Proceedings of the 11th International Conference of the Asian Association for Lexicography

 Lexicography in Asia: Challenges, Innovations and Prospects

Edited by Hai Xu
Center for Linguistics and Applied Linguistics
Guangdong University of Foreign Studies
Preface

All of us in China are proud to host the Asian Association for Lexicography (ASIALEX) Conference again after it has traveled around nine Asian countries and regions in the span of twenty years. The 11th Conference of ASIALEX (ASIALEX 2017, Guangzhou, June 10-12, 2017), organized by the National Key Research Center for Linguistics and Applied Linguistics at Guangdong University of Foreign Studies thus represents a happy opportunity for ASIALEX to celebrate the 20th anniversary of its founding.

Besides receiving felicitations from the Presidents of our global sister associations AFRILEX, AUSTRALEX, DSNA, and EURALEX, we have invited four world-renowned lexicographers as our keynote speakers:

- Prof. Jianhua Huang of Guangdong University of Foreign Studies, the First President of ASIALEX,
- Dr. Michael Rundell, Editor-in-Chief of Macmillan Dictionary,
- Prof. Andrea Abel of EURAC Research, President of EURALEX, and
- Dr. Julia Miller of Adelaide University, President of AUSTRALEX.

We have organized two advanced workshops, on Sketch Engine and DPS5, which will be run by Mr. Miloš Jakubíček, CEO of Lexical Computing, and by Mr. Holger Hvelplund, Vice President of Digital Solutions, IDM, respectively.

The theme of ASIALEX 2017 is Lexicography in Asia: Challenges, Innovations and Prospects. We think that it is timely to recognize our achievements in lexicographic research and practice in the past 20 years in Asia, and to look ahead to see how we can respond to the challenges of the revolutions in corpus linguistics and digital lexicography we are currently facing. In the four keynote speeches, Huang and Abel speak on the common theme of dictionary user orientation/participation in the digital age, and Rundell and Miller discuss extended units of meaning or phraseology, which lexicographers are increasingly aware of as representing the norm, rather than the exception, in language. All the issues the keynote speakers address are cutting-edge concerns, and most certainly deserve our special attention.

The enthusiasm of scholars and publishers from Asia and beyond that has greeted
this conference has been unexpectedly high. As one of the largest conferences in its series, ASIALEX 2017 hosts approximately 160 participants from 75 institutes over 24 countries and regions in Asia, Europe, Africa and North America. We received an astounding number of 130 abstract submissions. This volume of proceedings, which is 915 pages long, consists of 64 full papers and 49 abstracts, which are roughly divided into the sections digital lexicography, general-purpose lexicography, cognitive approaches to lexicography, bilingual lexicography, pedagogic lexicography, specialized lexicography, and historical lexicography. We are truly indebted to the contributors and the abstract reviewers for their hard work in bringing together such a remarkable collection.

While preparations for this grand event were under way, we sadly lost two great lexicographers who were highly influential in both China and the world, Professor Gusun Lu of Fudan University, who passed away on July 28, 2016, and Professor Boran Zhang of Nanjing University, who passed away on May 26, 2017. They both made enormous contributions to our field. To honour their great achievements, we have therefore set up a special session in their memory, and also dedicate this volume to these two great colleagues.

Finally, I would like to thank my PhD students, Yongfang Feng, Huilian Hu, Lingling Li, and Ziyue Chen, for assisting me in editing the proceedings. Ms. Yongfang Feng also painstakingly proofread the whole text. I am also grateful to my colleagues Prof. Martin Weisser and Dr. Vincent Ooi who helped revise some parts of the text.

Hai Xu

Chair, the 11th International Conference of the Asian Association for Lexicography (ASIALEX 2017)

June, 2017
CONTENTS

PREFACE i

I KEYNOTE SPEECHES

Searching for Extended Units of Meaning - and What to Do When You Find Them
Michael Rundell 1

User-oriented Compilation of the Grand Dictionnaire chinois-français contemporain
Jianhua Huang 17

Research in the Pipeline: Where Lexicography and Phraseology Meet
Julia Miller 18

Lexicography and User Participation in the Digital Age
Andrea Abel 19

II DIGITAL LEXICOGRAPHY

Africa’s Response to the Corpus Revolution
D.J. Prinsloo 20

Thai National Corpus (TNC) and a Corpus-based Monolingual Learners’ Dictionary of Thai
Jirapa Vitayapirak 32

Lexicological Module of Do It Yourself Corpora for Turkish
Bü lent ÖZKAN 40

From Ancient Manuscript Chinese Character Dictionary of TenreiBanshōMeigi to Deciphered
Electronic Text
Yuan Li 48

Building the KamusBesarBahasa Indonesia (KBBI) Database and Its Applications
David Moeljadi, Ian Kamajaya, Dora Amalia 64

The Development of Minority Dictionary via Digitalization: A Case Study of Uighur
Chuanming Sun, Yuting Zhu 81
Sketching the Difference between Word Senses
Meng-Hsien Shih 90

A Dictionary and Corpora Based Study of the Synonyms: “Contain” and “Include”
Congyi Joy Qu 102

The Lexicographical Treatment of Semantic Network Information in Online English Learner’s Dictionaries
Xiangming Zhang 113

A Sociocognitive Approach to English Online Dictionaries
Ying Lv, Congmei Chen 138

III GENERAL-PURPOSE LEXICOGRAPHY

Purism and Neologism: The Effect of Purism in the Development of the Northern Sotho Vocabulary
MV Mojela 168

Terminology, German Unity, the EU, and the Sorbian Minority
Gunter Schaarschmidt 175

Linguistic Purity? Profiling Lettered Words in Chinese Dictionaries
Lan Li, Qi Gong 185

The Quantitative Patterns of Lexical Neologisms: A Comparative-diachronic Corpus Approach
Tao Ma 197

Technical Words Boundary for General Purpose Dictionaries: A Grounded Theory
Ferdi Bozkurt 211

Gender Representation in the Contemporary Chinese Dictionary
Huilian Hu 221

A Brief Introduction to The American Heritage Dictionary of the English Language (5th)
Li Li 241

Yan Zhang 251
## IV COGNITIVE APPROACHES TO LEXICOGRAPHY

<table>
<thead>
<tr>
<th>Title</th>
<th>Author(s)</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defining Emotion Words: How to Bring Them down from Top to Bottom</td>
<td>Guiying Jiang, Xie Xi</td>
<td>259</td>
</tr>
<tr>
<td>Using Dictionaries in Metaphor Identification</td>
<td>Jihong Wu</td>
<td>277</td>
</tr>
<tr>
<td>The Presentation of Metaphor in the 7th Edition of <em>Modern Chinese Dictionary</em></td>
<td>Nan Geng</td>
<td>289</td>
</tr>
<tr>
<td>Method and Motivation of the Foregrounding Language Style of Paper English Learners’ Dictionaries——A Case Study of <em>COBUILD 8</em></td>
<td>Weifu Zhang, Congmei Chen</td>
<td>297</td>
</tr>
<tr>
<td>Heterosemy of Self-reference Lexemes in Modern Chinese from the Perspective of the Two-level Word Class Categorization Theory</td>
<td>Renqiang Wang, Changning Huang</td>
<td>309</td>
</tr>
<tr>
<td>On Multicategorization of Céngjing from the Perspective of the Two-level Word Class Categorization Theory</td>
<td>Yongqiang Ma</td>
<td>327</td>
</tr>
<tr>
<td>A Study of Multiple Class Membership in <em>The Chinese-English Dictionary (3rd Ed.)</em></td>
<td>Na Li, Renqiang Wang, Chao Wang, Jiangfan Yu, Huifang Tang</td>
<td>340</td>
</tr>
<tr>
<td>CSF-based Processing of Near-synonyms in the EFL Dictionaries for Chinese Learners</td>
<td>Huixiu Liu</td>
<td>376</td>
</tr>
</tbody>
</table>
V  BILINGUAL LEXICOGRAPHY
Anxing Li 384

A Discussion on New-Word Selection for Chinese-English Dictionaries
Yongwei Gao 393

Strategies Regarding the Translation of Neologisms in The Chinese-English Dictionary
Fushou Xiao 405

Translation Strategies of Culture-bound Words in The Chinese-English Dictionary (Unabridged) Volume I
Zi Wang, Lu Ye 418

On Bilingual Dictionary’s Strategy of Fixing Lexical Form in Dealing with Lexical Image Transformed between Languages
Yuanke Tao 430

The Translation of Keju Vocabulary in Chinese-English Dictionaries
Shan Ma 443

J. Huang’s Grand Dictionnaire chinois-français conteporain (2014) and His Lexicography
Bing Tian 456

Criteria for Selecting Trisyllabic Words as Headwords in the Chinese-French Dictionary
Shuyan Wang, Peng Zou 483

The Necessity of Compiling a Learners’ German-English-Chinese Etymological Dictionary
Changliang Qu 492

The Treatment of Polysemous Entries in Spanish-Chinese Bilingual Dictionaries
Yuqing Deng 500

Signposting Headwords for Increased User Friendliness in Bilingual Dictionaries
Mats-Peter Sundström 517

VI  PEDAGOGIC LEXICOGRAPHY
Connotations of Communicative Language Teaching for Developing Learners’ Dictionaries
Chuijun Qian 528
Treatment of Academic N-grams in English Pedagogical Dictionaries: A Corpus-based Study
Peng Chen

Which Words Should be Labelled Academic in an Indonesian Advanced Learner’s Dictionary?
Deny A. Kwary, Almira F. Artha

Characteristics of Lexical Items Appealing on the Websites of Japanese Corporations
Toshikazu Ezure

Illustrative Examples in Bilingual Learners’ Dictionaries
Naho Kawamoto

A Study on the Homonyms in the Advanced-level International Chinese
Wenting Wei, Shan Wang

A Chinese-English Dictionary for Production: With “yǒuxiào 有效” as an Example
Anmin Wang

Contrastive Study of Pragmatic Information between Learner’s Dictionaries
Yueting Rong

Weak-Form Representation in Online English Learner’s Dictionaries
Wenlong Sun

DICTIONARY-USE STUDIES

Dictionary User Research in the Digital Era: Value of User-generated Content
Yuyan He

Scaffolding Information of Learner’s Dictionaries for L2 Writing: A User Perspective
Na Yang, Yan Li, Jie Zheng

Dictionary Use and Peer Feedback in EFL Writing
Ying Cai

A Comparative Research on User’s Expectation of Learner’s Dictionary
Shan He

Lexicographical Delay
Erdogan BOZ
Analysis of the Dictionary Needs of China’s Intermediate Business English Learners
Yumei Cui, Julia Miller 737

Intermediate University Students’ Motivation to Receive Instruction for English-Japanese Dictionaries
Hiroe Kubota 757

Cultural Dictionary: An Instructional Enhancer to Better Reading Comprehension in Literature Classes
Rosario B. Dizon, Melba B. Ijan, Alia D. Ramber, Kim Minoza, Manna Hadassah B. Estrada, Gissele O. Dangdang 779

VIII SPECIALISED LEXICOGRAPHY

Complex Culture-Bound Notions in a Bilingual Dictionary of Idioms
Joanna Szerszunowicz 791

Creating Dictionary Entries in a Bilingual Dictionary of Proverbs: The Case of Tshivenđa
Munzhedzi James Mafela 803

How to Build a Specialized Chinese-English/English-Chinese Electronic Dictionary for Chinese Business English Learners
Zhiyi Zhou 814

Microstructure in Dictionaries of Turkish Grammar and Linguistics Terms: Examples of Mood and Modality Terms
Duygu Kamacı 823

IX HISTORICAL LEXICOGRAPHY

Some Aspects of Language in Samuel Johnson’s The Rambler
Masaaki Ogura 832

Two Lexicographers of English Collocation Dictionaries in Japan: Katsumata and Irie
Kazuo DOHI 840

A Turkish Historical Homonymy Dictionary “Eser-i Şevket” And Its Structure
Ezgi ASLAN 852

An Example of Interlinear (Subline) Dictionary Kitabû’l-müsellese
Tuğba POÇAN 860
ABSTRACTS

On the Presentation of Semantic Prosody in Advanced English-Chinese Learner’s Dictionaries
Lu Bai 869

The Effectiveness of a Bilingualized Dictionary in Helping Learners Determine Article Use
Alice Yin Wa Chan 870

A Discourse Approach to Critical Lexicography: Methods, Principles and Applications
Wenge Chen 871

Coding Transitivity Construction in English Monolingual Dictionaries: The Syntax-Semantics Perspective
Ziyue Chen 872

The Treatment of Figurative Collocations in Paper and Electronic Learner’s Dictionaries
Yuanjun Dai 873

Online Indonesian Comprehensive Dictionary: A Unifying Dictionary for Local Languages in Indonesia
Azhari Dasman Darnis, Adi Budiyanto 874

A Survey of the Use of Dictionaries by Medical Students
Jun Ding 875

Complex-Transitive Argument Structures in English Learner’s Dictionaries: A Usage-based Constructionist Approach
Yongfang Feng 876

Reflections on the Role of National Lexicographical Planning in China
Yundong Geng 877

FSPOST: A Part-of-Speech Tagger for Filipino
Matthew Phillip Go, Nicco Nocon, Nathaniel Oco, Rachel Edita Roxas 878

Topic Models of Philippine Disaster Preparedness E-Participation Responses
Joyce Emlyn Guiao, Jennifer Carreon, Brandie Nonnecke, Nathaniel Oco, Rachel Edita Roxas 879

A Review of Definition Research in Chinese Dictionaries for Foreign Learners (CDFFLs)
Qiping Guo 880

Computational Representation of a Model of Entries in a Business English Learners’ Dictionary
Jianing He, Mosi He 881
<table>
<thead>
<tr>
<th>Title</th>
<th>Author(s)</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cultural Aspects in Chinese Bilingual Dictionaries</td>
<td>Audrey Heijns</td>
<td>882</td>
</tr>
<tr>
<td>A Corpus-based Investigation on Definitions of Four-character Idioms in Chinese-French Dictionaries</td>
<td>Fang Huang</td>
<td>883</td>
</tr>
<tr>
<td>Issues in Bilingual Lexicography with Pashto and English as Pairs</td>
<td>Zafar Iqbal, Saida Jamal</td>
<td>884</td>
</tr>
<tr>
<td>A CiteSpace-based Knowledge Mapping of Lexicographic Research in <em>International Journal of Lexicography</em></td>
<td>Wenping Jiang</td>
<td>885</td>
</tr>
<tr>
<td>Subject Fields in EFL Dictionaries and ESL Textbooks</td>
<td>Takashi Kanazashi</td>
<td>886</td>
</tr>
<tr>
<td>A Study on the Method to Select Headwords for the Development of a L2 Learner of Korean Synonym Dictionary: Based on the Errors of the Learner's Corpus</td>
<td>Jeong-hyun Kim</td>
<td>887</td>
</tr>
<tr>
<td>Japanese Words in the Dictionarium Annamiticum Lusitanum et Latinum (1651)</td>
<td>Emi KISHIMOTO</td>
<td>888</td>
</tr>
<tr>
<td>Developing a Bi-Directional Ilocano-English Translator for the Travel Domain: Using Domain Adaptation Techniques on Religious Parallel Corpora</td>
<td>Ariel Neimond A. Lazaro, Nathaniel Oco, Rachel Edita Roxas</td>
<td>889</td>
</tr>
<tr>
<td>Features of <em>The English-Chinese Dictionary</em> Edited by Prof. Lu Gusun</td>
<td>Ang Li</td>
<td>890</td>
</tr>
<tr>
<td>The Presentation of Metaphorical Information in Chinese-English Dictionaries: An Empirical Study</td>
<td>Dehui Li</td>
<td>891</td>
</tr>
<tr>
<td>The Back-of-the-book Indexes in the Chinese Lexicographical Books</td>
<td>Huixian Li, Qin Dai</td>
<td>892</td>
</tr>
<tr>
<td>Coping with Medical Polysemy in Online Medical Dictionaries: A Case Study</td>
<td>Jinyan Li</td>
<td>893</td>
</tr>
<tr>
<td>Bilingual Glossaries in Chinese Almanac: Space for Informal Language Learning</td>
<td>Michelle Li</td>
<td>894</td>
</tr>
<tr>
<td>The Use of an Online Dictionary by EFL learners: A Case Study</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Xiqin Liu, Dongping Zheng, Yushuai Chen

Collocation Networks of “Community of Shared Future”
Tingting Luo

Structuring Polysemy in English-Chinese Learner’s Dictionaries: A Prototype Theory Perspective
Huaguo Lu

Counting and Expressing Numerals in Asian Languages: How People View Their World
Noa Nishimoto

Re-examining the Big Five English Learner Dictionaries for Semantic Prosody
Vincent B Y Ooi

The Effectiveness of Multimedia Lexical Knowledge Representation as a Learning Tool
Takeshi Sato, Tyler Burden

Exploring the Criteria of Dictionary Examples for Users at Different L2 Proficiency Levels
Yukio Tono

Zongzi, a Prospective New Word from Chinese Entering the OED?
Tai-yuan Hugo Tseng

On the Lexicographical Characteristics of The Chinese-English Dictionary
Jiangbo Wan

A Study on the Homonyms in the Advanced-level of International Chinese Teaching
Wenting Wei, Shan Wang

Parallel Corpus on Different Language Representations of the Same Chinese-English Concepts
Xiaofang Wu, Houwei Ge, Yu Wang, Li Wei

Middle Construction: Difficulties for L2 Acquisition and Adaptation to Dictionary Representation
Cheng Xiong

Using Think-Aloud Protocol to Investigate the E-Dictionary Use by Chinese Learners
Tingting Xu

EFL Dictionaries Should Treat Meaning More Properly
Shigeru Yamada
Study of Confucian Words in Chinese-English Dictionaries: Based upon Phraseology in Contrast
Rong Yang 911

Assisting Non-Native Chinese Learners: Robert Morrison’s Pioneering Efforts in Wuche Yunfu (1819)
Ying Ye, Xiangqing Wei 912

A Second Bite at the Ultimate Dictionary
Xuhua Zhang 913

Form-Concept Mapping and the Organization of Semantic and Contextual Information in The Chinese-English Dictionary (Unabridged)
Cuilian Zhao 914

The Influence of English-Japanese Lexicography Upon Early Monolingual English Learners’ Dictionaries
Lianzhen Zhao 915
Searching for Extended Units of Meaning - and What to Do When You Find Them

Michael Rundell
Editor-in-Chief, Macmillan Dictionary
michael.rundell@lexmasterclass.com

Abstract

Two of the key outcomes of corpus-linguistic research over the past 30 years have been the development of the idea that meanings are mostly constructed through context (undermining traditional notions of the individual word as an autonomous bearer of meaning); and the discovery that recurrence and regularity — our tendency to employ a limited number of conventionalized ways of expressing ideas — are essential features of the language system. Both findings have had a major impact on our understanding of how language works, and both have influenced the content of dictionary entries — contributing for example to improved word sense disambiguation (WSD) and to a greater emphasis on phraseology and collocation. But there is still much to do. Ever-larger corpora and more powerful corpus-query tools reveal areas where we can further improve our description of languages, and thus provide better resources for users. And the migration of dictionaries to digital forms (removing space constraints) opens up new opportunities for doing this.

In a characteristically far-sighted paper (Sinclair 1996), John Sinclair broadened the search for what he called “units of meaning” by investigating longer strings of words and identifying recurrent, and often quite extended, patterns of usage. Using this as a starting point, I will look at other examples in corpus data of the kinds of patterning Sinclair discussed, and we will see how current corpus-querying systems can help us identify these extended units of meaning. Finally, I will speculate about whether dictionaries should aim to describe these longer units, and if so, how this might work in practice.


Keywords: extended units of meaning, collocation, colligation, semantic prosody, multi-word sketch, longest-commonest-match
Part 1: Theoretical background

1.1 Introduction

I recently sent a text message (SMS) to a friend whose father was unwell. As I was composing the message, it felt as if the phone’s software already knew what I wanted to say. Each time I added a word, the system successfully predicted the next word, which was almost always one of the three options presented. My first keystroke was “S”, and this brought up three suggestions - one of which was *Sorry* (the word I was planning to write). Selecting *Sorry* then brought up options which included *to*, and this in turn led to *hear*, which led to *that*, and so on. Before long, I had written a complete sentence: “Sorry to hear that your father is in the hospital”. Or rather, the phone had more or less written it for me. The experience reminds us of a well-known quote from the scientist and science fiction writer Arthur C. Clarke, who said that “any sufficiently advanced technology is indistinguishable from magic.” This seemed like magic, but in reality of course it is science. The system on my phone used a combination of rule-based and statistical methods to predict the likely trajectory of my sentence. The science behind this rests in part on an essential fact about human language — a feature which has only become clear thanks to the the availability of large amounts of language data. That feature is recurrence: when we use words to encode meanings, we generally draw upon a fairly limited repertoire of conventionalized “ways of saying”. Consequently, a high percentage of what people say or write is predictable, because, as Hanks observes, “Although the number of possible combinations may in principle be limitless ...the number of probable combinations ...is rather limited” (Hanks 2013: 399).

However, although corpus analysis enables us to observe the inbuilt predictability of most language output, much of this is far from predictable to a learner or non-fluent user of a language. Even where a given word combination is semantically transparent, its status as a recurrent string, as a norm worth learning, is not necessarily obvious. This raises the question of what dictionaries can or should do in order to identify and describe recurrent patterns of usage, and this is one of the themes of this paper. I will start by looking at a paper written over 20 years ago by John Sinclair (Sinclair 1996), in which he explores the idea of “extended units of meaning”. In many earlier works, Sinclair had already developed (and demonstrated the workings of) what he called “the idiom principle” — the idea that language users regularly resort to an inventory of “semi-preconstructed phrases that constitute single choices” (Sinclair 1991: 110). At this point, relatively small units of meaning, such as two-word collocations and other so-called “lexical bundles” (Biber et al. 1999: 990ff), had been extensively discussed (and were beginning to be accounted for in some dictionaries). In his 1996 paper, however, Sinclair was interested in longer patterns, sometimes of considerable complexity, which the data shows to be remarkably frequent.

From a lexicographic point of view, this interest in multiword units of meaning is relatively recent. Traditional lexicographic practice has rested on the assumption that individual words are autonomous bearers of meaning — a view reflected in the names
we give to dictionaries (called “word books” in many Germanic languages) and even
in definitions of the word dictionary itself, for example:

*a book that gives a list of words in alphabetical order and explains what they mean*

(Macmillan Dictionary, first edition, 2002)

This definition (now superseded) exactly describes Cawdrey’s *Table Alphabetical* of
1604 (generally thought of as the first monolingual English dictionary), where the
“definitions” typically consist of one or two almost-synonymous words (see Figure 1),
while contextual information (showing the conditions in which word X is equivalent to
word Y) is entirely absent.

![List of words](image)

**Figure 1** extract from Cawdrey’s *Table Alphabetical*

A focus on the word as the primary unit of meaning is not so surprising given that “in
the majority of writing and printing conventions, words are separated by spaces…A
text is therefore seen as a succession of discrete items, those items being words”
(Sinclair 1996: 75). In this model, the meaning of an utterance is a concatenation of the
meanings of the individual words which it comprises. And when dictionaries have to
deal with longer units of meaning such as idioms and phrasal verbs, these have
traditionally been relegated to the bottom of a main dictionary entry, so that *break out*
and *break the bank* are “nested” at the end of the entry for *break*. These longer units, as
Sinclair notes, “are considered as marginal phenomena, almost aberrations” (Sinclair
1996: 76).

1.2 Understanding how meanings are created

All of this would change under the impact of corpus study, which provided the
empirical basis for a radically different understanding of how meanings are created —
a model of meaning summed up in Sinclair’s well-known observation that "Many if
not most meanings depend for their normal realization on the presence of more than
one word" (Sinclair 1998). A number of pre-corpus scholars had already begun to
develop the idea that meaning is at least partly dependent on context and co-text, rather
than being an inherent property of individual words. As far back as 1755, Samuel
Johnson had recognised that “It is not sufficient that a word is found, unless it be so
combined as that its meaning is apparently determined by the tract and tenour of the
sentence” (Johnson 1755). Much more recently, J. R. Firth demonstrated that features
such as collocation, caggiation and phraseology had a central (rather than marginal)
function in the language system, and that the meaning of a word could not be fully
understood without knowing “the company it keeps”.

Similar ideas surfaced among people involved in the teaching of English as a
second language. Several years of experience as a language teacher brought Harold
Palmer — working in Japan during the 1920s and 1930s — to the realization that “it is
not so much the words of English nor the grammar of English that makes English
difficult … The vague and undefined obstacle to progress … consists for the most part
in the existence of so many odd comings-together-of-words.” (Palmer 1933, quoted in
Cowie 1999: 52-53). The first major learner’s dictionary of English, compiled in Japan
by Palmer’s protégé A.S. Hornby, moved away from the standard model by including
information about syntactic behaviour and (to a lesser extent) collocation and
phraseology. Since Hornby, pedagogical dictionaries have paid increasing attention to
longer units of meaning. Salient two-word collocations are extensively covered in
today’s dictionaries, while in Longman’s Activator (1993), the lists of near-synonyms
which lexicalize a given concept make no distinction between single words and
phrases. For example, the concept “Usually” is instantiated not only by single words
(like generally and routinely) but also by longer units (such as nine times out of ten and
as a rule). And in a significant recent development (facilitated by the migration of
dictionary text from print to digital platforms), some pedagogical dictionaries now treat
phrasal verbs and idiomatic phrases as headwords in their own right, only loosely
connected to the entries under which they were formerly nested. Gradually, then,
dictionaries’ exclusive focus on single words has given way to a more mixed picture,
where the role of longer units of meaning is recognised at both microstructural and
macrostructural levels.

These developments in lexicographic practice reflect theoretical insights, gained
through the study of corpus data, into how meanings are created and understood.
Hanks proposes (e.g. Hanks 2013: 73f.) that words on their own do not have meanings;
rather, they have “meaning potentials”. These meaning potentials are activated by
specific contextual features, and many of the resulting patterns (of word plus syntactic
and/or lexical context) recur frequently enough in corpus data to be regarded as normal
units of meaning. And since the primary job of a dictionary is to account for linguistic
norms, Sinclair’s interest in longer units of meaning (beyond those already described in
dictionaries) is a logical next development.

1.3 Sinclair’s 1996 paper: The search for units of meaning

In this paper, Sinclair looks in detail at corpus data for the expression naked eye, a
frequent though more or less opaque combination. He discovers a complex network of
recurrent patterns. To summarize the main features of this patterning: the two positions
to the left of the node, which we will refer to as N-1 and N-2, are typically filled by to
the [naked eye] or with the [naked eye]. As we move further to the left, things get more interesting. At position N-3 there is a strong preference for words relating to visibility: this slot tends to be filled by verbs like see, spot, and perceive, or by adjectives like visible, evident, and detectable. For the verbs in N-3, there is a colligational preference for use with a modal at N-4 (especially can or could): …these could be seen with the naked eye from a helicopter....On top of all this, Sinclair finds that a semantic prosody of “difficulty” is evident in 85% of the instances in his sample: we see this in expressions like too faint to be seen by the naked eye or barely visible to the naked eye.

What Sinclair presents here is a model of a lexical item consisting of several words which — unlike items that we categorize as idioms or fixed phrases — can tolerate “a great deal of internal variation”. Distinct linguistic features such as collocation, semantic preference, colligation and semantic prosody all combine to create one of the “semi-preconstructed” phrases which Sinclair referred to in his earlier work and which he sees as effectively a single lexical choice. And despite all the internal variation found in his set of naked eye phrases, “there is always a clearly preferred selection right down to the actual words”.

There are of course many exceptions to the patterns which Sinclair focuses on: in position N-2, for example, prepositions other than to and with can sometimes occur, and there are some sentences which don’t exhibit any of the semantic preferences Sinclair identifies as being typical, such as:

To the naked eye, he is easily one of their the fittest...

Anyone who has spent much time looking at corpus data knows that it is not difficult to find exceptions to whatever generalizations emerge from one’s analysis. But for dictionary-makers, the exceptions are of less interest than the norms.

From the point of view of practical lexicography, two key messages emerge from Sinclair’s investigations. Firstly, the need to broaden our notions of what constitutes a lexical unit to be accounted for in a dictionary: “So strong are the co-occurrence tendencies of words, word classes, meanings and attitudes that we must widen our horizons and expect the units of meaning to be much more extensive and varied than is seen in a single word.” The second and related point is that, in the theory of meaning Sinclair proposes in this paper, “the idea of a word carrying meaning on its own would be relegated to the margins of linguistic interest, in the enumeration of flora and fauna for example”. From a lexicographic point of view, this is a significant reversal of traditional practice, where the word is central and longer units are seen as anomalies.

1.4 Some new examples

Before we look further at the implications of all this, it would be useful to get a clearer idea of how pervasive this sort of patterning is, so we will attempt a similar analysis on a few other items. The corpus used here is the LexMCI corpus, which was the main evidence base for the creation of the DANTE lexical database (Convery et al. 2010). LexMCI is a collection of about 1.7 billion words of contemporary English, and is
described in more detail on the DANTE website: http://www.webdante.com/the_corpus.html.

1.4.1 wreak

Number of instances in the corpus: 2334

A word sketch for wreak (Figure 2 shows an extract) vividly illustrates the stand-out feature of this verb.

![Figure 2 extract from a Word Sketch for wreak](image)

This verb’s typical objects fall into just two semantic classes: revenge or some form of chaos and destruction. There is some variation in the choice of collocates, but havoc is so dominant that the collocation wreak havoc is almost a fixed phrase. Regardless of the object type, a prepositional phrase follows the noun object in almost 60% of cases, with on being by far the most frequent preposition. The usual pattern is wreak havoc/revenge on, but in about 14% of cases the object is modified by an adjective such as untold, terrible, or enormous.

1.4.2 untoward

Number of instances in the corpus: 992

This is an unusual adjective. A noun complement follows in about 45% of cases, and two semantic types dominate: words meaning (roughly) “event” or “consequence”. The most frequent of these is incident: the collocation untoward incident makes up over 13% of all instances of untoward. Typically occurring in the context of discussions about health and safety, it could be seen almost as a technical term in its own right:

Staff also need some avenue for whistleblowing, to voice concerns about untoward incidents or bad practice.
The striking colligational feature of *untoward* is that it functions, in a majority of
cases, as a postpositive adjective, usually preceded by the words *nothing* (201
instances), *anything* (172), or *something* (50):

*He was keen to reassure parents that nothing untoward was going on.*
*After initially denying that anything untoward had happened... he later confessed to
raping the boy.*

In these and many similar cases, there is a broadly negative semantic prosody. And
when that does not apply, a conditional appears in around 10% of the postpositive
examples:

*If you spot anything untoward, let me know.*

*Wreak* and *untoward* are relatively rare words in English, so it would not be surprising
if the range of patterns in which they appear is limited. But even here, there is clear
evidence of of recurrent “units of meaning”, some of which are quite extended.

1.4.3 *sink in*

Our focus here is the phrasal verb. A simple search for the lemma *sink*-verb followed
immediately by *in* (*sink in* is a non-separable phrasal verb) generates a concordance of
3819 lines. But more than half are instances of the verb followed by a prepositional
phrase, and therefore not relevant here. For example:

*Their fishing boat sank in the Bristol Channel between Penarth Pier and Cardiff Bay.*
*He was so sunk in his despair, he scarce observed the change.*

To filter out the noise, a search string was created using CQL (Corpus Query
Language), a syntax used in Sketch Engine for specifying complex searches. The CQL
query reduced the output from 3819 lines to 1592, and a random sample of 600 was
taken from this “candidate” set. Finally, manual methods reduced the 600 candidates to
470 *bona fide* instances of the phrasal verb *sink in*, and this dataset forms the basis for
what follows.

One of the most striking features of the verb’s behaviour is its colligations.
Colligation refers to a word’s observable preference for occurring in — or for avoiding
— a particular form, a particular position in the sentence, or a particular grammatical
function (see Hoey 2005: 43ff. for a fuller description). Three facts stand out. First, the
phrasal verb has a strong preference for the infinitive form, with 144 of 470 instances
being infinitives. Secondly, in well over 70% of cases, *sink in* appears at — or very
close to — the end of a sentence or clause. 265 instances are immediately followed by
punctuation; in a further 50 or so cases, we find the pattern *sink in* + adverb +
punctuation (e.g. *It hasn’t sunk in yet*); and in 20 or so other cases, *sink in* is followed
(without intervening punctuation) by a conjunction such as *and* or *but* introducing a
new clause. In those cases where sink in is not clause- or sentence-final, it is often followed by a that-clause (47 instances) or occasionally a wh-clause (6). And thirdly, in almost a quarter of cases sink in occurs in a broadly negative environment, such as:

*It hadn’t really sunk in until I spoke to mum.*  
*I don't think the shock of it all has sunk in yet.*  
The realization that they are ‘just like us’ has yet to sink in.

Looking now at the characteristic co-text of sink in, there are three common types of subject (referring to what it is that sinks in): information (instantiated by words like message, words, and news), impact (implications, scale, realisation, gravity, impact, extent), and —most frequently — the pronoun it:

*Read that again. Let it sink in.*  
*I miss him already and it hasn't really sunk in.*  
"*It still hasn't sunk in," says McGoldrick.*

A key fact about the meaning of sink in is reflected in another recurrent contextual feature: in a high proportion of cases, there is some indication that “sinking in” — the full absorption of new information — is a process, and it takes time. This feature is realised in a number of ways:

- co-occurring with start or begin (40 instances)  
  *Now that she has done that, the shock begins to sink in.*  
  *It is just starting to sink in now but when my name was announced I was just dumbstruck.*

- in the pattern take + time marker + to sink in (51 instances):  
  *...the reality took a little while to sink in*  
  *This came as a complete shock to me and has taken a few days to sink in.*  
  *(As variation on this, we also find patterns like let/allow/give something (time) to sink in.)*  

- co-occurring with adverbs such as gradually, finally, slowly, eventually, or in questions or negatives with yet (has it sunk in yet?)

- the use of the progressive form (typically with an adverb like still or only just):  
  *The awful truth was slowly sinking in*  
  *(... the shock of what happened is still sinking in)*

Finally to the role (if any) of semantic prosody in the way sink in behaves. In his analysis of naked eye Sinclair found a semantic prosody of “difficulty” in a high proportion of corpus instances. Prosodies like this differ from collocation or...
characteristic co-text in that they are not instantiated by specific lexical items. In the case of *naked eye*, the sense of something being difficult is generally present, but the ways in which this is conveyed lexically can be quite diverse. In the case of *sink in*, the picture is less clear. “Bad” situations outnumber “good” ones by about three to one, so we are more likely to encounter instances like this:

*As the scale of the catastrophe sank in, he began to fear for his family...*

…than like this:

*As the initial euphoria sinks in you say to yourself, 'What do I do now?’*

But both good and bad types are outnumbered by cases which are “neutral” (or cases where it is impossible to tell one way or the other), and typical subjects like *implications, message, scale, or realisation* are not inherently positive or negative.

Much of what Sinclair found in his analysis of *naked eye* holds true for words like *wreak, untoward*, and especially *sink in*. In every case, corpus analysis reveals — beneath the surface variation — patterns which appear repeatedly in the language data, including frequently co-occurring words and clear colligational preferences. As well as further undermining the idea of words as independent bearers of meaning, the examples discussed here support what Sinclair calls “the case for compound lexical items” which may be of considerable length.

The evidence of usage leaves little doubt that such extended units of meaning are a pervasive feature of everyday language, and that “the independence of the choice of words is compromised, because other patterns cut across them and constrain them” (Sinclair 1996). In order to understand why our language output should be “compromised” in this way, and why it relies so much on recurrent patterns (of whatever length), it is helpful to invoke Michael Hoey’s notion of “lexical priming”. Hoey proposes that “every word is primed for use in discourse as a result of the cumulative effects of an individual’s encounters with the word” (Hoey 2005: 13). As the data for *sink in* suggests, we are likely to encounter this verb in one of a limited set of contexts, showing one of a limited number of selectional and colligational preferences and (sometimes) semantic prosodies. These are its “primings”, and they influence us when we use the word ourselves. The process is circular and self-reinforcing, and for Hoey, such primings are “the driving force behind language use, language structure and language change” (Hoey 2005: 12).

**Part 2: Implications for practical lexicography**

One of the lessons of corpus linguistics is that words tend to occur frequently with certain other words in predictable patterns and contexts. This insight has informed many of the innovations in pedagogical dictionaries over the last twenty years, notably a sharper focus on collocation. In his 1996 paper, Sinclair showed that these networks
of co-occurrence could be considerably more extensive than had been envisaged in earlier corpus studies. So if we accept that extended units of meaning are a significant feature of the language system, it follows that a dictionary which aims to describe normal usage should find ways of incorporating such information. This raises two questions: how can lexicographers identify recurrent extended units in an efficient and non-time-consuming way; and, once found, how should they be accounted for in dictionaries?

2.1 Finding extended units in a corpus

Since the beginnings of corpus-based lexicography, efforts have been made to identify recurrent patterns in the language and to account for them in dictionaries. When the main (or only) analysis tool was the concordance, finding patterns could be a laborious process, and one whose outcomes were not necessarily complete or systematic. But lexical profiling software, of which the best-known example is the Word Sketch (Kilgarriff et al. 2004), has transformed this operation. A Word Sketch presents the lexicographer with lists of a word’s most significant collocates. Lists are grouped according to the grammatical relations they instantiate (such as word+NP, ADJECTIVE+word), and can be ranked by frequency or salience. The Sketches also list prepositions which typically follow a word (and in some cases those that precede it), and sometimes also show “constructions” such as that-clauses or infinitive clauses. In all cases, a further click takes the user to a concordance of the selected pattern. The automatic detection of patterns like these is a well-researched topic in natural language processing, and methods for extracting this information are well established and widely used. Finding shorter units of meaning (typically, two words which regularly co-occur) is now a relatively straightforward process.

If we broaden our search horizons to take in Sinclair’s extended units of meaning, what kind of software tools will we need? In Sketch Engine, some progress has already been made in this direction, with two features which have been added fairly recently: multi-word sketches and “longest-commonest match” (Kilgarriff et al. 2015). For a multi-word sketch, the starting point is any two-word collocation, and the observation that in many cases a third collocate is found in the corpus data. For example, when we look at instances of the common collocation seek+advice we find that many include an adjective referring to the type of advice being sought: professional, legal, medical, financial, and so on. These are effectively three-word collocations, and in the large corpora available now it is easy to find numerous examples of this type. The second feature, longest-commonest-match, is based on the concordance for a search word, and identifies any multiword string which accounts for a high proportion of the corpus instances. (Technical aspects of this feature — how the algorithm works, and how it could be improved — are discussed in Kilgarriff et al. 2015.)

To give an idea of how these functions work, Figure 3 shows an extract from a Word Sketch for the adjective vocal. The corpus used is the very large (>20-billion-word) EnTenTen13 web corpus of English (freely available in Sketch Engine), and
collocates are ranked according to frequency, not salience. The screenshot shows the top four collocates for two grammatical relations:

![Word Sketch for vocal](image)

**Figure 3** extract from a Word Sketch for *vocal*

Each of the eight collocates shown in this extract has a multiword string, in grey, below the collocate, and this is the longest-commonest-match. So, at the collocate *quite*, the algorithm has detected that the sequence “quite vocal about” is especially frequent. All the collocates in this extract (*very, quite, increasingly, etc*) are shown in bold and followed by a + sign, and this gives access to a multi-word sketch. Clicking this + symbol brings up a new Word Sketch for the composite item *increasingly vocal,*

![Word Sketch for increasingly vocal](image)

which occurs 1000 times in this corpus (Figure 4):

**Figure 4** extract from a Word Sketch for *increasingly vocal*

This multi-word sketch shows that combinations such as *increasingly vocal critic* are fairly common, and that (see top right-hand column) prepositional phrases with *in* or
about often follow increasingly vocal. But the word that occurs most frequently with increasingly vocal is the verb become (in the bottom right-hand column), accounting for 269 instances out of the 1000 examples. Because of the frequency of this combination, the word become has its own + symbol, and this brings up an even more granular multi-word sketch for the combination “become increasingly vocal” (Figure 5):

![Fig 5](image)

**Figure 5** extract from a Word Sketch for become increasingly vocal

We have now drilled down almost as far as we can, but our final observation is that this three-word string is often followed by a PP with either about or in. If we click on one of the numbers (64 or 62) we bring up a concordance like this (Figure 6):

![Fig 6](image)

**Figure 6** extract from a concordance for become increasingly vocal in

There are 62 instances of the multiword unit “become increasingly vocal in”. Even here the highlighted words to the right of the node draw our attention to nouns which are common collocates of vocal. At this level it is feasible to scan all the concordances in the “old-fashioned” way, and this reveals that the seven word string “become increasingly vocal in [one’s] criticism of” (with 13 of the 64 concordance lines) is the commonest pattern — and a clear example of what Sinclair called an “extended unit of meaning”.

We have seen that even currently-available corpus querying tools — provided
they are working with sufficiently large corpora — can help us to detect recurrent multiword units of considerable length. For purposes of linguistic research, the software already works well, but it does not yet meet the needs of time-poor lexicographers. Their task is to analyse corpus data and identify all of the lexicographically-relevant facts about a word or phrase — effectively, all frequently-occurring patterns of whatever type — and to complete this operation as rapidly as possible. This requires a high level of automation. Automation works optimally when lexicographers are not required to make too many subjective choices, and can feel confident that the software has provided them with a complete set of each linguistic feature they are looking for (see e.g. Rundell and Kilgarriff 2011). The “classic” Word Sketch meets both these needs. But the recently-added functionality which facilitated our analysis of *vocal* is not yet ideal for lexicographic purposes. It raises doubts about the level of “recall” (have all relevant extended units been found?) and it adds a degree of complexity to the task which will result in a significant (and probably unacceptable) overhead in terms of time.

None of this is unsurmountable. The search mechanisms do their job and the information is all there if you have the time and skill to find it. So the issue is largely one of optimising the presentation, so that all relevant information is made available to the working lexicographer in an easy-to-digest form.

Possible improvements include a function which extracts what the software sees as the most significant multiword strings in which a search word participates, and presents them in a single list ranked by frequency or salience. A further refinement would be to classify such n-grams and present separate lists for different types of expression. These could include lexical collocations of more than two words (such as *seek professional advice*), or prepositional phrases that frequently follow the search word. Many of these n-grams will be semi-fixed patterns with slots for words belonging to a particular semantic set, and the system could show which actual words, or which types of word, typically fill the slot. For example: *vocal in [one’s] [criticism, opposition, condemnation, concerns]* or *X takes [time marker] to sink in*. The goal is to maximize the usefulness of the information for lexicographers, and this will call for some further processing, some design tweaks, and perhaps for the use of data visualization techniques.

### 2.2 Implementation: what goes in the dictionary, and why

Let us assume that improvements in corpus-querying software can be made which will provide lexicographers with a user-friendly overview of the significant extended units typical of a given search word. We then need to consider whether this is the kind of information dictionaries should include — and if so, how this might be done.

Over the last 30 years or so, corpus-based research has led to major changes in our understanding of how meanings are created and interpreted. The idea of words as semantically autonomous is no longer sustainable, at least for mainstream non-specialist discourse, and this insight is increasingly applied to the content of dictionary entries. In older monolingual dictionaries, the description of meaning was often limited
to short definitions, and one-word translation equivalents performed a similar function in bilingual dictionaries. But meaning is now understood to be distributed across longer sequences of words, and the scope of dictionary entries has broadened significantly to reflect this view, taking in a range of features such as selectional restrictions, constructions, collocation, and phraseological conventions.

Things have now taken another step forward. With larger corpora and more powerful search tools at our disposal, we are learning even more about the pervasiveness of what Sinclair called the “idiom principle”, and of the extent to which meanings are conveyed through multiword strings which may be longer than previously suspected. The goal of corpus lexicography is to describe linguistic norms — to produce what Hanks calls “an inventory of normal uses of each word in a language” (Hanks 2013: 92) — so it follows that dictionary entries should expand further to incorporate the kind of information discussed in this paper.

There are clear benefits here for computational applications, such as word sense disambiguation. Indeed, as Sinclair predicted in his paper on extended units, the approach he proposes should mean that “some of the problems of conventional description are much reduced — for example there will be little word-based ambiguity left when this model has been applied thoroughly” (Sinclair 1996). But what about human users of dictionaries? Many of the units described both by Sinclair and in this paper are compositional in nature. Expressions such as barely detectable to the naked eye or increasingly vocal in one’s (criticism/ opposition etc) are certainly frequent, but they do not pose particular problems of comprehension. However, the same is true of many two-word collocations, but these are nevertheless regarded as worth describing, especially in pedagogical dictionaries. Even if their meaning is transparent, their form is often unpredictable — and the same rationale applies to longer units. As examples of normal, frequent usage, they can help dictionary-users to understand the conventions of mainstream discourse, or sometimes of specific types of discourse. (Some extended units are especially typical, for example, of journalistic, academic, or technical registers.) Non-fluent speakers are thus provided with models for production which will help them to avoid unnatural, non-idiomatic language.

In his last paper, written in 2007 and published posthumously, John Sinclair returned to this theme. Observing once again “the tendency of words to occur together more often than their frequency would predict”, he saw far-reaching implications for lexicography, because “the definiendum…is no longer a simple entity, a headword” (Sinclair 2007/2010: 37). In practical terms, this probably does not imply that extended units should themselves be full and separate headwords, in the way that fixed idiomatic phrases like as let the cat out of the bag are often now presented in online dictionaries. The information belongs more usefully at the main headword (a word such as vocal or sink in), on the assumption that the user wants to understand the word more fully and to use it naturally. In many dictionaries, the entries already tell the user about the headword’s syntactic behaviour and its salient collocational preferences. If we now aim to describe the common extended units associated with it, there will be some tough design issues to contend with. When adding new categories of data to dictionary entries which are already information-rich, we face the challenge of giving users
access to the facts which they need at a given moment, while also making it easy for
them to ignore data-types which do not interest them (see Rundell 2015: 308-309 for a
fuller discussion). But, without the space constraints of paper-based dictionaries, there
is no reason why we cannot solve these problems, and the resulting dictionary entries
would provide a new and deeper level of information on how words naturally and
typically behave and combine in text.

Acknowledgments

I am grateful to Vojtěch Kovář of the Sketch Engine team for his helpful comments on
the functions discussed in section 2.1.

References

Grammar of Spoken and Written English. London: Pearson Education.
Convery, C., Ó Mianáin, P., Ó Raghallaigh, M., Atkins, S., Kilgarriff, A., and Rundell,
M. 2010. The DANTE Database (Database of ANalysed Texts of English). In
Anne Dykstra & Tanneke Schoonheim (Eds.). Proceedings of the XIV EURALEX
Congress. Leeuwarden: Fryske Akademy.
Oxford University Press.
Press
Routledge.
Johnson, S. 1755. Preface to A Dictionary of the English Language. Edited by Jack
Lynch, and available at http://andromeda.rutgers.edu/~jlynch/Texts/preface.html
Williams and Vessier (Eds.) Proceedings of the Eleventh Euralex Congress, UBS
Lorient, France. 105–116.
In Kosem, I., Jakubiček, M., Kallas, J., Krek, S. (eds.) Electronic lexicography in
the 21st century: linking lexical data in the digital age. Proceedings of the eLex
Rundell, M. 2015. From Print to Digital: Implications for Dictionary Policy and
will it all end? In Meunier F., De Cock S., Gilquin, G. and Paquot M. (Eds), A
Taste for Corpora. A tribute to Professor Sylviane Granger. Amsterdam:


User-oriented Compilation of the *Grand Dictionnaire chinois-français contemporain*

**Jianhua Huang**
Guangdong University of Foreign Studies

This paper specifically discusses how one bilingual dictionary—the *Grand dictionnaire chinois-français contemporain* (GDCF), of which I am the editor-in-chief (Huang, 2014)—was designed to respond to the needs of the source language and target language speakers. To cater for the needs of Chinese-speaking learners of French, GDCF takes into account the wide coverage of wordlist, and enriches the information on the frequency, grammar, and register of French equivalents. To be geared to the needs of French-speaking learners of Chinese, GDCF incorporates various Chinese dialect words, and provides useful information on the formulaicity and flexibility of Chinese morphology (classifiers in particular). Some other design features such as the indication of POS in examples of usage, establishment of a semantic network of the nomenclature, and provision of dynamic equivalents are expected to benefit both groups of users.

**Keywords:** user-orientation, coverage, frequency, grammar, register, dialect words, formulaicity, flexibility, classifier, POS, semantic network, dynamic equivalent

**References**

Research in the Pipeline: Where Lexicography and Phraseology Meet

Julia Miller
The University of Adelaide
julia.miller@adelaide.edu.au

In the pipeline is an English idiom referring to something abstract, such as a project, that is in progress but is not yet complete. The inclusion of idioms such as this in a dictionary raises many questions for lexicographers. Which headword would a user choose when looking for this idiom? Which dictionaries should include the idiom? How could it be translated into another language? How frequently is it used? Who uses it? Are idioms purely informal or can they be used in formal contexts, such as journal articles?

There is a widespread belief that idioms are little used in academic speaking and writing. Research indicates, however, that this is not the case. In the pipeline, for example, occurs in 3916 articles and books in the Springer Exemplar corpus. Idioms like this give the writer greater entry into the field of academic discourse, both spoken and written. Indeed, failure to understand an idiom can lead to misunderstanding of key parts of a text. Conversely, an ability to use idioms in their own speech and writing gives the researcher greater inclusion in an academic community.

This presentation will explore preliminary findings based on an analysis of the British Academic Spoken English (BASE) corpus and the Michigan Corpus of Academic Spoken English (MICASE), cross-checked with articles in Springer Exemplar, to identify which idioms (multi-word figurative expressions with limited lexical and syntactic variability) are used most often in English academic speaking and writing. The most frequently used idioms can then be used more confidently by researchers with English as an additional language and taught more confidently to students. These idioms can also be included in monolingual, bilingual and academic dictionaries.

The presentation will then explore ways in which phraseology contributes to lexicography in diverse fields, with reference to each of the conference themes. It is hoped that this will inspire joint research projects between phraseologists and lexicographers so that future conference presentations can report on more research in the pipeline.
Lexicography and User Participation in the Digital Age

Andrea Abel
The European Academy of Bozen/Bolzano, EURAC
Andrea.Abel@eurac.edu

Online dictionaries rely increasingly on their users and leverage methods for facilitating user participation at basically any step of the lexicographic process. This ranges from giving feedback or correcting errors to creating new articles and discussing language- or subject-related issues beyond the explicitly encoded knowledge. The ease of communication and collaboration between publishers and users has enormous potential, not only for keeping a reference work up to date and at a high level of quality, but also for developing improved, user-adapted views of and access to the contents of a dictionary.

This talk will discuss user participation as a new field of metalexicographic research, bringing together, as a first step, multiple isolated previous works using a common framework. To this end, the different types of user participation are presented, backed up by multiple practical examples found in existing online reference works:

1. direct user participation comprises dictionary articles entirely or partly written by users in a collaborative effort,
2. indirect user participation occurs in different forms of explicit and implicit feedback,
3. accessory user participation goes beyond the dictionary content by initiating a discourse either between the dictionary makers and their users or among the users themselves.

The talk will be concluded by discussing strengths and drawbacks of different types of user participation. It will be argued that user participation research is an emerging field of lexicography in order to properly benefit from and provide better products for users.
Africa’s Response to the Corpus Revolution

D.J. Prinsloo
University of Pretoria, South Africa
danie.prinsloo@up.ac.za

Abstract

The aim of this paper is to reflect on Africa’s response to the corpus revolution in the late 1980’s which was pioneered by Collins COBUILD on a mere 20 million-words corpus of English of the 1980’s. The first corpora built for African languages also ranged from a modest one million to 10 million words. This paper reflects on how the use of corpora improve the quality of dictionaries for African Languages ranging from basic enhancement of the macrostructure as well as the microstructure to more sophisticated applications and lexicographic tools built from corpora. On the level of the macrostructure reference will be made to compilation of the lemmalist from frequency lists culled from the corpus, frequency indication in the dictionary, combatting inconsistencies and more advanced applications e.g. lexicographic tools built from corpora such as rulers and block systems. On the level of the microstructure keywords in context culled from the corpus assist the lexicographer with especially sense distinction and the selection of examples. More advanced applications on microstructural level include dictionary use support systems such as domain specific guidance, linking processed and unprocessed corpus data to e-dictionaries and corpus verification for text production in a dictionary-use situation.

Keywords: corpora, dictionaries, African languages, corpus-based dictionaries, dictionary use support systems

1. Introduction

African language lexicography does not exist in isolation. It is subjected to the same user demands and trends, changes and developments as in international lexicography. However, African languages have their own unique challenges mostly revolving around being lesser-resourced languages. Lesser-resourced languages generally lack dictionaries of a high lexicographic standard resulting from a combination of criteria such as a limited number of speakers, spoken by under-developed communities, subjected to political oppression and lacking a standardized orthography, lack of financial resources, text and oral corpora, a dictionary culture, human language technologies (HLT) tools, etc. It is not possible to give a comprehensive overview of corpus-based lexicographic activities for the entire continent of Africa, the focus in this
paper is on highlighting a number of corpus enabled strategies and tools developed for African language lexicography.

2. African languages

According to Wikipedia the number of African languages is estimated at between 1,250 to 2,100 and some estimates suggest more than 3,000. They are divided into six major language families as indicated in Figure 1 (Languages of Africa: Wikipedia).

![Figure 1 Major African language families: Wikipedia](image)

The focus of this presentation is on Bantu languages (Niger-Congo B in Figure 1). Most African languages, especially the Bantu languages, have complex grammatical systems, classification of nouns into different classes, a complex concordial and pronominal system, problematic lemmatisation traditions and orthographic systems. These characteristics pose a challenge to the lexicographer to produce user-friendly dictionaries for especially learners of African languages. The lexicographer has to design and employ a number of dictionary support systems to enable access on demand especially for highly problematic constructions such as the copulative system of Sepedi as explained below. The computer era which enabled corpus based lexicography surely represent one of the major changes in lexicography and offers significant opportunities but also challenges to African language lexicography. Generally speaking it could be argued that African languages are slow in responding to the computer era.

3. Maximizing limited corpora

In most cases African language lexicographers have only small and raw corpora at their disposal. Prinsloo (2015) tried to ascertain the value of such limited corpora for
dictionary compilation. He came to the conclusion that raw corpora, even if they are compiled only from written sources, can substantially assist the lexicographer in the compilation of especially small bilingual dictionaries. On the macrostructural level a corpus of one million words is useful to pinpoint the most commonly used words in the language. As far as microstructural elements are concerned, he concluded that a one million corpus is useful in determining the basic senses of a word as well as typical examples of usage of these basic senses.

Due to the scarcity of resources, it is not possible to design corpora in terms of balance and representativeness, e.g. as for the pioneering Brown corpus (Francis and Kucera 1979) or to follow the corpus design criteria of Atkins et al. (1991). The answer to corpus building for African languages is in terms of Atkins (Gouws and Prinsloo 2005:25) to build what she calls organic corpora. She says: “A corpus may be thought of as organic, and must be allowed to grow and live if it is to reflect a growing living language”. Corpus sizes generally range between 1-10 million and many corpus compilations are still work in progress.

4. The significance of high frequency

The frequency of occurrence of words in written and spoken corpora play an important role in modern corpus-based lexicography. The impact of top frequencies in everyday written and oral communication is extremely high. In terms of Collins Cobuild English Dictionary (Sinclair 1995:xiii) the top 1,900 English words account for approximately 75% of all English usage and the top 14,700 for 95% of all spoken and written English. Calculations for Sepedi indicate that coverage exceeding 90% is reached by an even smaller number of words in the top frequency band.

5. Macro structure enrichment

The use of corpora enhanced the quality of African language dictionaries on macrostructural level in a number of ways which could be described as ranging from basic strategies, e.g. compilation of lemmalists, putting frequencies in the dictionary and combatting inconsistencies, to more sophisticated ones such as dictionary support systems.

5.1 Compilation of lemmalists and solving lemmatisation problems

The compilation of corpus-based lemmalists are inspired by frequency of use – frequency lists (types) culled from a corpus are used as input to the lemmatisation process. A major advantage of such an approach is that frequently used words are unlikely to be accidentally overlooked and pinpoints the most common words of the languages in cases where the lexicographer is restricted by the publisher for
commercial reasons to a limited number of lemmas – typically 5,000 in many dictionaries. (Prinsloo 2009).

5.2 Solving lemmatisation problems

Due to a conjunctive orthography for many African languages a stem lemmatisation approach is followed which requires the user to identify the stem in order to look up the word in the dictionary. To date lexicographers were unsuccessful to solve this problem for the hundreds if not thousands of derivations of e.g. an isiZulu word that the user is likely to find when reading isiZulu books. So, for example, the user has to know that the meaning of the following words angakhulumi, esakhulumi, basikhulumisa, ngikhume, ngokukhulumi, ukukhulumi, bezikhulumela, sikhulumile, okhulumayo, should all be looked up under the stem -khumla ‘talk’. An isiZulu corpus can generate the entire paradigm for -khumla in seconds and all derivations can consequently be linked to the appropriate dictionary articles. The user can simply type the entire word or a section thereof and the software automatically takes him/her to the relevant dictionary articles – thus the prerequisite to know the stem is no longer applicable. Lemmatisation problems in African languages are described in great detail in the literature and the interested reader is referred to sources such as Van Wyk (1995), Prinsloo (2009, 2011), Gouws and Prinsloo (2005), Prinsloo and De Schryver (1999).

5.3 Putting frequencies in the dictionary

Corpora enabled lexicographers to give users an indication of frequency. This is for instance done in Collins COBUILD English Dictionary by means of one to five filled/unfilled diamonds, Macmillan English dictionary for advanced learners (Rundell 2007) by one to three stars and in the Longman Dictionary of Contemporary English (Summers 1953) by means of three levels S1,S2, S3 for spoken and W1, W2 and W3 for written language.

Collins COBUILD English Dictionary: need 限限限限, latest 限限限, flag 限限限限限, ordeal 限限限限, flaming 限限
Macmillan English dictionary: answer 限限限限限限限, dictionary 限限限限限, diesel 
Longman Dictionary of Contemporary English: leg S1W1, legislation W2, legal S3W1, prosecution W3

On these scales the topmost frequently used words will be indicated by five filled diamonds, three stars, S1 and W1.

Indicating frequencies in dictionaries for African languages is uncommon but Oxford Bilingual School Dictionary De Schryver (2007) can be quoted as an example. The one to three star approach was followed.

5.4 Combatting inconsistencies

De Schryver and Prinsloo (2000) discuss a number of typical inconsistencies in African language dictionaries compiled without a corpus. The first type of inconsistency highlighted is obvious gaps in the lemmalists of dictionaries compiled on introspection detected when the lemmalists of such dictionaries are compared to a frequency list culled from a corpus between two fixed points.

The second type of inconsistency pertains to unequal treatment of verbal derivations, i.e. that lexicographers fail in terms of their editorial policy to treat a number of frequently used verbal derivations. Consider Table 1 (De Schryver and Prinsloo 2000:296) as a typical example.

<table>
<thead>
<tr>
<th>root --&gt;</th>
<th>bolela</th>
<th>dira</th>
<th>hwetša</th>
<th>rata</th>
<th>reka</th>
<th>Tseba</th>
</tr>
</thead>
<tbody>
<tr>
<td>derivation</td>
<td>(5,735)</td>
<td>(5,475)</td>
<td>(3,371)</td>
<td>(2,786)</td>
<td>(551)</td>
<td>(5,851)</td>
</tr>
<tr>
<td>↓ + passive</td>
<td>BOLELWA</td>
<td>dirwa</td>
<td>hwetšwa</td>
<td>ratiwa</td>
<td>rekwa</td>
<td>tsebja</td>
</tr>
<tr>
<td></td>
<td>(408)</td>
<td>(636)</td>
<td>(260)</td>
<td>(5), ratwa</td>
<td>(122)</td>
<td>(441)</td>
</tr>
<tr>
<td>+ applicative &amp; passive</td>
<td>BOLELELW</td>
<td>DIRELW</td>
<td>A</td>
<td>A</td>
<td>Rekelw</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>(6)</td>
<td>(40)</td>
<td></td>
<td></td>
<td></td>
<td>(19)</td>
</tr>
<tr>
<td>+ perfectum &amp; passive</td>
<td>boletšwe</td>
<td>dirilwe</td>
<td>HWEDITŠ</td>
<td>RATILW</td>
<td>rekilwe</td>
<td>TSEBILW</td>
</tr>
<tr>
<td></td>
<td>(44)</td>
<td>(137)</td>
<td>E</td>
<td>E</td>
<td>(17)</td>
<td>E</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(10)</td>
</tr>
</tbody>
</table>

Words given in lowercase were either entered as lemma signs or treated in the microstructure in The New Sesotho – English Dictionary while those in uppercase were not, representing inconsistent treatment.

5.5 Lexicographic rulers and block systems

Alphabetical stretches in dictionaries are not equal in size. The alphabetical stretches for A, B, D, M, R and especially C and S, in English dictionaries contain large numbers of lemmas, occupying almost 50% of the dictionary, while categories such as J, K, Q, U, V, X, Y and Z are relatively small, and consequently fill only a few pages. Alphabetical word lists culled from corpora for African languages enabled the compilation of rulers
by simply calculating the number of words in each alphabetical stretch as a percentage of the total number of words in the list. Consider Figure 2 representing a ruler for Tshivenda.

![Figure 2 A lexicographic ruler for Tshivenda](image)

Alphabetical rulers can also be calculated on a percentage basis i.e. dividing the ruler into 100 blocks, a so-called block system as in Figure 3, to guide and monitor the dictionary compilation progress.

<table>
<thead>
<tr>
<th>% Marker</th>
<th>% Marker</th>
<th>% Marker</th>
<th>% Marker</th>
<th>% Marker</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ALAF</td>
<td>21 FAHL</td>
<td>41 KUKE</td>
<td>61 MONO</td>
<td>81 SEET</td>
</tr>
<tr>
<td>2 AROG</td>
<td>22 FETL</td>
<td>42 LAMO</td>
<td>62 MOŠA</td>
<td>82 SEJA</td>
</tr>
<tr>
<td>3 BAFE</td>
<td>23 FOŠW</td>
<td>43 LEDI</td>
<td>63 MOTO</td>
<td>83 SEMA</td>
</tr>
<tr>
<td>4 BANK</td>
<td>24 GALA</td>
<td>44 LEKA</td>
<td>64 MPHO</td>
<td>84 SERE</td>
</tr>
<tr>
<td>5 BEAB</td>
<td>25 GAYA</td>
<td>45 LEPO</td>
<td>65 NASW</td>
<td>85 SETS</td>
</tr>
<tr>
<td>6 BITL</td>
<td>26 GOLE</td>
<td>46 LETŠ</td>
<td>66 NGWE</td>
<td>86 SITE</td>
</tr>
<tr>
<td>7 BOGE</td>
<td>27 HAHA</td>
<td>47 LOGW</td>
<td>67 NKOK</td>
<td>87 STEF</td>
</tr>
<tr>
<td>8 BOKO</td>
<td>28 HLALH</td>
<td>48 MABE</td>
<td>68 NTEB</td>
<td>88 SWEL</td>
</tr>
<tr>
<td>9 BOMM</td>
<td>29 HLOGL</td>
<td>49 MAGA</td>
<td>69 NTSE</td>
<td>89 TEKE</td>
</tr>
<tr>
<td>10 BOPU</td>
<td>30 HOSE</td>
<td>50 MAKA</td>
<td>70 NYAK</td>
<td>90 THAT</td>
</tr>
<tr>
<td>11 BOLT</td>
<td>31 IHLO</td>
<td>51 MAMO</td>
<td>71 OLEL</td>
<td>91 THOM</td>
</tr>
<tr>
<td>12 BUWA</td>
<td>32 ILWA</td>
<td>52 MARA</td>
<td>72 PANK</td>
<td>92 TIKR</td>
</tr>
<tr>
<td>13 DEAP</td>
<td>33 IPIT</td>
<td>53 MATH</td>
<td>73 PHAK</td>
<td>93 TLAH</td>
</tr>
<tr>
<td>14 DIIP</td>
<td>34 ITIH</td>
<td>54 MEAG</td>
<td>74 PHET</td>
<td>94 TONA</td>
</tr>
<tr>
<td>15 DIKI</td>
<td>35 JESU</td>
<td>55 MELO</td>
<td>75 PIPA</td>
<td>95 TSEN</td>
</tr>
<tr>
<td>16 DIPE</td>
<td>36 KATO</td>
<td>56 MIDI</td>
<td>76 PŠHA</td>
<td>96 TŠHI</td>
</tr>
<tr>
<td>17 DIJE</td>
<td>37 KGAN</td>
<td>57 MMAS</td>
<td>77 RANG</td>
<td>97 TSOL</td>
</tr>
<tr>
<td>18 DIJO</td>
<td>38 KGHOH</td>
<td>58 MOBO</td>
<td>78 RETA</td>
<td>98 TUME</td>
</tr>
<tr>
<td>19 DUDI</td>
<td>39 KGWA</td>
<td>59 MOHIL</td>
<td>79 RRAG</td>
<td>99 WABO</td>
</tr>
<tr>
<td>20 EMAE</td>
<td>40 KLAS</td>
<td>60 MOKO</td>
<td>80 SATH</td>
<td>100 ZOUN</td>
</tr>
</tbody>
</table>

![Figure 3 A Block System for Sepedi](image)

6. Micro structure enrichment

As in the case of the macro structure the use of corpora enhanced the quality of African language dictionaries on microstructural level in a number of ways. Studying concordance lines generated from corpora assist the lexicographer in respect of sense distinction and examples of usage and quite a number of African language lexicographers build and employ corpora for this purpose. However more sophisticated
dictionary support systems are also developed from corpora to assist the user in terms of his/her text production and text reception needs.

6.1 Sense distinction and choice of examples

Studying even a limited number of concordance lines for a specific word is normally sufficient to determine the most important senses and sub-senses. Consider a few concordance lines for the Sepedi word ntšha ‘take out’ in Table 2

<table>
<thead>
<tr>
<th>#</th>
<th>Left co-text</th>
<th>Node</th>
<th>Right co-text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>bo bonala bo sa le gona. Ba mo tšea ntšha ba mo</td>
<td>moleteng. Ge ba mo lebelediša ka seetša</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ka gana ge e le yona ka re ke noko. O ntšha ile a</td>
<td>mphaka wa gagwe wa bogale. Ka wona a nko. Ba gorogile ka mogobo wa nngalaba, tawana e tee ya e iša go mmagoyona, ya e</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>seatla sa yona. Ge ba goroga ke ge ntšha letšatsi le</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Ba tliše!” Phukuje ya tsena ka ntšha mphomeng ya</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>a ba Jabes wa ga Gileada. Bjale ntšha phuthego ya</td>
<td>banna ba dikete tše lesome le metšo e</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Ka lebaka la bodiidi le la bohumi ntšha kgadi yeo ya</td>
<td>lentšu ya re: “Re tlo bona ge o ka tla wa</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>ke utswitšego thekethe ya Seila ya ntšha tšhelele ka yona. O be a romilwe ke dipere ka yo</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These lines reveal the basic sense ‘to take out something’, for instance a knife in Line 2 and a person or animal from a hole Lines 1 and 4, ‘to pay / earn money’ (literally ‘to take out money’) in Line 7, or idiomatic uses in the sense of ‘the sun taking out its nose’ (that is ‘rise’) in Line 3, or ‘the congregation taking out men’ (that is ‘nominate / identify’) in Line 5.

More advanced microstructural enrichment initiatives for the African languages include corpus assisted homonym detection and separation of concordance lines for homonyms, corpus verification of text production, domain specific information, term extraction and lexicographic guidance for typical errors in student writing.

6.2 Dictionary use support systems

Due to the complex grammatical system of African languages, as mentioned above, the dictionary user needs more guidance beyond the standard treatment provided by the dictionary article. So, for example, an English speaking learner of Sepedi who wants to look up a basic top frequency copulative such as is, am or are in a Sepedi-English dictionary does not expect that correct text production with is, am and are is so
complicated that it is described in up to 30 pages of fine print in a Sepedi grammar. The lexicographer cannot serve the text production needs in such cases within the boundaries of a dictionary article, nor can (s)he expect the user to study the entire copulative system of Sepedi as a prerequisite to the use of copulatives. (S)he has to find quick and effective ways of linking guiding support systems to the standard dictionary article, e.g. by means of pop-up boxes, guidance paths and selection tree structures. The aim is to give access on demand, guiding the user through a complex system directly to the answer (s)he is looking for.

6.2.1 Linking processed and unprocessed corpus data to e-dictionaries

The aim is to extend the traditional dictionary article with additional information that can be digested at a glance, e.g. filling a single computer screen as in Figure 4.

![Figure 4 Linked information for Afrikaans aandag ‘attention’](image)

In this example basic concordance lines with aandag as the keyword in context are given as well as typical collocation statistics, spreading across sources, etc.

6.2.2 Frequency trajectories

Frequency trajectories reflecting the use of words over a long period of time was first introduced for Afrikaans in Prinsloo and Gouws (2006) cf. Figure 5a. Such trajectories are of use to the lexicographer and informative to the user indicating whether words
retain a stable pattern of usage or whether it is increasingly used or antiquating, e.g. in labelling words as neologisms, obsolete, etc. Provision of a frequency trajectory is currently given as a standard information category in google-lookups, cf. Figure 5b.

**Figure 5a** Frequency trajectories for *rowwe* versus *ruwe* ‘rough’ in Afrikaans

**Figure 5b** Frequency trajectory for *rough* in Google

**Figure 5** Frequency trajectories for *rowwe*, *ruwe* and *rough*

Figure 5a indicates that the use of *ruwe* is on the decline showing an upward trend for *rowwe*.

### 6.2.3 Verification of text production

Given the complex nominal and verbal systems of many African languages, text production in itself is challenging to the learner and there is a strong need for verification of constructions and sentences produced by users in a dictionary use situation. Such verification is attempted in terms of exact matching, i.e. finding an identical construction in the corpus or at least a part-of-speech match confirming that the correct morphemes, concords and pronouns for specific noun classes and verbal moods have been selected and that they are presented in the correct syntactic order. This is work in progress for Sepedi which will be linked to the *Sepedihelper* (Prinsloo et al. 2015), a writing tool for text production.

### 6.2.4 Domain specific guidance

Corpora can assist the lexicographer to provide the user with domain-specific indication on the usage of words. Consider Table 3 for the Afrikaans word *diens* ‘service’ in religion-related texts versus newspaper texts.
Table 3 Diens in religious texts versus newspaper texts (Prinsloo et al. 2013)

<table>
<thead>
<tr>
<th>Collocation</th>
<th>Relig.</th>
<th>News</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>in diens</td>
<td>206</td>
<td>410</td>
<td>in service</td>
</tr>
<tr>
<td>jaar(are) diens</td>
<td>18</td>
<td>51</td>
<td>year(s) of s.</td>
</tr>
<tr>
<td>beter/beste diens</td>
<td>0</td>
<td>65</td>
<td>better service</td>
</tr>
<tr>
<td>swak diens</td>
<td>0</td>
<td>55</td>
<td>bad service</td>
</tr>
<tr>
<td>goeie diens</td>
<td>2</td>
<td>75</td>
<td>good service</td>
</tr>
<tr>
<td>aan diens</td>
<td>8</td>
<td>131</td>
<td>on duty</td>
</tr>
<tr>
<td>onbaatsugtige diens</td>
<td>15</td>
<td>1</td>
<td>unselfish s.</td>
</tr>
<tr>
<td>besondere diens</td>
<td>180</td>
<td>0</td>
<td>special service</td>
</tr>
<tr>
<td>diakonale diens</td>
<td>15</td>
<td>0</td>
<td>deacon service</td>
</tr>
</tbody>
</table>

From Table 3 it is clear that collocations such as *beste diens* and *swak diens* are frequently used in newspaper texts but not in religious texts while *besondere diens* and *diakonale diens* are exclusively used in religious texts.

6.2.5 Homonym separation

Corpora are not only invaluable in providing concordance lines for homonyms but also to facilitate the process to distinguish and group them for each homonym using high frequency collocators and clusters typical for each homonym as determiners. A tool for homonym detection is currently compiled to separate and group together concordance lines for homonyms in Afrikaans and Sepedi.

7. Conclusion

The corpus era has also dawned upon African language lexicography. Dictionaries for these languages are subjected to the same user demands and trends, changes and developments as in international lexicography. The computer era and the availability of corpora brought new challenges but also new opportunities to African language lexicography. In this paper a brief overview of such challenges and opportunities was attempted. Corpus based lexicography needs to be expanded in African languages. More and larger corpora for an increased number of African languages should be compiled. The quality of dictionaries for African languages should be enhanced through basic and advanced utilization of corpora.
Acknowledgment

This research is supported by the National Research Foundation of South Africa (grant-specific unique reference number (UID) 8576). The Grantholder acknowledges that the opinions, findings and conclusions or recommendations expressed in any publication generated by the NRF-supported research are those of the author, and that the NRF accepts no liability whatsoever in this regard.

References


Prinsloo, D.J. & De Schryver, G.-M. 2004. Crafting a Multidimensional Ruler for the


Thai National Corpus (TNC) and a Corpus-based Monolingual
Learners’ Dictionary of Thai

Jirapa Vitayapirak
Faculty of Liberal Arts
King Mongkut's Institute of Technology Ladkrabang, Thailand
kvjirapa@gmail.com

Abstract

Corpus technology has become central to dictionary development. This paper presents a research project on developing a new dictionary for learners of Thai basing on Thai National Corpus for print and digital media following similar principles of modern pedagogical lexicography. It was targeted for intermediate-level users. It was compiled by a linguist team at Faculty of Liberal Arts, King Mongkut's Institute of Technology Ladkrabang. It follows a simple principle of making it easy for non-native Thai users to find what they want. Based on Thai National Corpus (TNC) for headword selection using word frequency, it initially consisted of approximately 12,000 entries. The authentic examples of usage in the entries were also selected from the Thai National Corpus. The entries included phonetic transcription for pronunciation, part of speech, irregular forms, definition, and example of usage for each sense, and other information such as synonym and antonym, register, subject domain, qualifier, hyponym and various compositional phrases.

Keywords: Thai National Corpus, dictionary, learner, monolingual

1. Introduction

Dictionary compilation has undergone a phase of innovation with the help of computer corpora. At present major monolingual dictionaries for EFL students were published by using general reference corpora, for instance, the Collins COBUILD, Oxford, Cambridge, Longman and so on. The corpora contribute positive features in almost newly revised learners’ dictionaries (Vitayapirak 2001).

For Thai dictionaries, it should be noted that little previous Thai-Thai general dictionaries has been compiled on the use of corpora to develop dictionaries in terms of checking word frequencies, searching for word patterns of words combination, using examples of the uses of particular words combination. It seems that the information recorded in Thai dictionaries may not be authoritative and back up by hard evidence from
the authentic information from the text corpus. This research is an attempt to use Thai National Corpus (TNC) for Thai lexicography. This paper argues that Thai dictionary compilation should be prior in corpus design. It has three dimensions. One is the background of Thai dictionaries, in which Thai dictionaries in the past were reviewed. The second dimension is the details about the Thai National Corpus (TNC). The third dimension is the creation of a corpus-based monolingual learners’ dictionary of Thai. The integration of these three dimensions is intended to demonstrate the good principles of dictionary design to improve the quality of dictionary for learners of Thai.

2. Background of Thai Dictionaries

Thai dictionaries compilation has extended over a period exceeding two centuries by the work of Christian Missionaries in Thailand or Siam (Vitayapirak, 2016). When the missionaries were in Siam, they realized the great need to learn Siamese language to communicate with Thais. The need for reference source or dictionary is apparent to all. The following is the list of Thai dictionaries during 1846-1964:

<table>
<thead>
<tr>
<th>No.</th>
<th>Dictionary Title</th>
<th>Language Combinations</th>
<th>Publication Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Dictionary of the Siamese Language</td>
<td>Thai-English</td>
<td>1846</td>
</tr>
<tr>
<td>1.2</td>
<td>Dictionarium Lingue Thai (Pallegoix)</td>
<td>Thai-Latin-French-English</td>
<td>1854</td>
</tr>
<tr>
<td>1.3</td>
<td>Dictionary of the Siamese Language</td>
<td>Thai-Thai</td>
<td>1873</td>
</tr>
<tr>
<td>1.4</td>
<td>Siamese-English Dictionary (Michell)</td>
<td>Thai-English</td>
<td>1892</td>
</tr>
<tr>
<td>1.5</td>
<td>The Royal Institute Dictionary (Bradley)</td>
<td>Thai-Thai</td>
<td>1927</td>
</tr>
<tr>
<td>1.6</td>
<td>Thai-English Dictionary (McFarland)</td>
<td>Thai-English</td>
<td>1944</td>
</tr>
<tr>
<td>1.7</td>
<td>Thai-English Students Dictionary (Haas)</td>
<td>Thai-English Learners’</td>
<td>1964</td>
</tr>
</tbody>
</table>

2.1 The first monolingual Thai dictionary published

During King Mongkut’s period, the first monolingual Thai-Thai dictionary published was Dictionary of the Siamese Language or ด้วยภาษาไทยเพิ่ม (Bradley: 1873) which the headwords had been defined wholly by good Thai scholars, mainly Nai Muang, under the supervision of Dr. D.B. Bradley. He edited the outstanding work of publishing the first monolingual Thai dictionary. In his diary in 1854, Bradley commented about the unreliable and poorly written Siamese reference materials at that time. He thus started his own project
on the Dictionary of Siamese Language. The objectives were first for the benefits of missionaries in acquiring the language, and second for a standard work for the Siamese themselves.

2.2 Thai-English Student’s Dictionary (for non-native Thais)

During the 1930s, there was the end of Siamese absolute monarchy in 1932 and at the same time the world had conflict as the Second World War (WWII) had started in 1939 involving most of the world’s nations. Due to the need for communication of oriental languages such as Thai during the WWII, Mary Haas started fieldwork in the phonology and syntax of Thai (Siamese) in 1941 at the University of Michigan because of the need for speakers of Asiatic languages as war developed. In spite of expanding interest in Thai studies during that decade, there were the needs of Thai-English Students’ Dictionary for non-native speaker of Thai. The first Thai-English Student’s Dictionary was compiled by Haas in 1964. The main aim was to assist students or those who have engaged in work involving the Thai language.

Haas (1964) noticed about most of Thai dictionaries that: (1) they fail to provide a scientifically accurate and up-to-date guide to the pronunciation of Thai words. (2) they neglect to distinguish important levels of usage e.g., vulgar, common, elegant, colloquial, literacy, etc. (3) they lack many important idioms, compounds, and other specialized combinations of words. (4) Examples, when given, are more likely to be chosen to clarify the peculiarities of English than those of Thai. Haas’s dictionary thus contains the traditional Thai spelling as shown in the Thai monolingual dictionary, phonemic transcription using Roman letters, an abbreviation designing the word class, the level of usage, both social and technical, the English definition or equivalent, synonyms and antonyms and examples of usage when the definition alone is insufficient. An important feature is the indication of stress for every entry and the tones of the individual syllables.

2.3 Thai-Thai Dictionaries for native Thais

The Royal Institute Dictionary (RID) or พจนานุกรมมัธยมศึกษาปีที่พิมพ์ is the most well-known well-known work of Thai-Thai dictionary for native speakers of Thai. It was first appeared in 1927, first printing in 1950. The Institute has published four fully revised editions of the dictionary, and many intermittent reprints with minor revisions, i.e. 1950, 1982, 1999 and 2011. Since the Royal Institute is the most respectful lexicographical organisation in Thailand whose main function is the general development of the Thai language, both in terms of codification of new concepts (terminology) and establishing norms for written Thai (standardization), the RID is still officially remained useful reference for Thais (Vitayapirak, 2001).
3. Thai National Corpus (TNC)

The Thai National Corpus (TNC) is a general corpus of Thai language. It is conducted by the linguistics department, Chulalongkorn University under the patronage of H.R.H Princess Maha Chakri Sirindhorn with collaboration from many researchers and publishers (Aroonmanakun, 2007). It is designed to be comparable to the British National Corpus in terms of its domain and medium proportions. The corpus size is targeted at eighty million words. Only written texts are collected and the words are segmented and tagged following the Text Encoding Initiative (TEI) guidelines on text encoding. Word segmentation and transcription program proposed in Aroonmanakun and Rivepiboon (2004) is used as a tagger. TNC Header is used for inputting contextual information and generating header tag for each text. Output from TNC Tagger is combined with the header tag as an XML document. In general, the TNC was designed as a resource for general applications, such as lexicography, language teaching, and linguistic research. Therefore, the TNC is fruitful for this research especially on the compilation of corpus-based monolingual learners’ dictionary of Thai.

4. Creation of a Corpus-Based Monolingual Learners’ Dictionary of Thai

Before deciding on a corpus-based monolingual learners’ dictionary of Thai, it was of interest to review the existing Thai dictionaries. It was found that most of them are no more than glossaries of words, and the selection of their headwords is not decided by linguistic evidence such as the frequency of occurrence or concordance from a corpus. In general, Hartmann and James (1988) suggested the dictionaries should be decided in terms of two main aspects, i.e. macrostructure and microstructure. Macrostructure includes internal structure of the wordlist and other components such as the front, middle, and back matters. Microstructure includes the elements of entry, i.e. the format and information provided in an entry. For the compilation of corpus-based monolingual learners’ dictionary of Thai in this study, the first task is the selection of headwords using frequency data from the TNC and then the construction and internal arrangement of the entries. The following sections explains about them.

4.1 Frequency data as criteria for headword selection

Frequency data from TNC corpus is used explicitly within dictionary entries to indicate how likely the word is to turn up in the students reading. The potential headword list for this dictionary thus starts with those words with middle to high frequency in the TNC corpus. An example of the word ‘ชิน’ and the frequency was shown below:
4.2 Usage in context

Words in context are an important feature of language instruction as a means of providing detailed information on collocations and grammatical information in authentic language. This dictionary thus provides a number of examples in the dictionary. It should be noted that some of the examples are adapted in order to make then self-standing and shorten. This is exemplified in the following sample concordance below:

Concordance

"นางเกลี้ยง คือผู้ที่สร้างงานเขียน อย่างไรก็ตามคำนี้มักใช้เฉพาะกับผู้ที่เขียนงานสร้างสรรค์หรือเป็นอาชีพ หรือผู้ที่ได้สร้างงานเขียนในลักษณะอื่น ๆ ...
ที่กุมมีนางเกลี้ยง นั่นก็คือในทุกวัน ชุดนักเขียนนั้นรู้จัก มีผลงานให้อ่านเพียบ ส่วนนักเขียนนักเขียนพนักงานศิลป์ นักเขียนพนักงาน, นักเขียน กามาที่สร้างสรรค์อยู่ที่นักเขียน ...
อาชีพนักเขียน เป็นอาชีพหนึ่งที่ให้ทั่วไป สำหรับผู้ที่มีอาชีพที่สร้างสรรค์ในยุคปัจจุบัน สำหรับผู้ที่มีอาชีพที่สร้างสรรค์ในยุคปัจจุบัน สำหรับผู้ที่มีอาชีพที่สร้างสรรค์ในยุคปัจจุบัน สำหรับผู้ที่มีอาชีพที่สร้างสรรค์ในยุคปัจจุบัน สำหรับผู้ที่มีอาชีพที่สร้างสรรค์ในยุคปัจจุบัน สำหรับผู้ที่มีอาชีพที่สร้างสรรค์ในยุคปัจจุบัน สำหรับผู้ที่มีอาชีพที่สร้างสรรค์ในยุคปัจจุบัน..."

Figure 1 Concordance

4.3 Non-corpus based Aspects of the Dictionary entries

4.3.1 Structure of an entry

This section shows a sample entry from the dictionary. It is intended to illustrate the key features of the dictionary. All entries show the use of Thai synonyms and translations for each sense of the headword as well as pronunciation, grammatical information or part of speech, and Thai example of usage. Moreover, other special features are included, i.e.
idioms (Thai collocational phrases), sense indicators, register, subject field, hypernym, synonym, and antonym (Vitayapirak and Kernerman. 2011).

<table>
<thead>
<tr>
<th>Headword</th>
<th>ข้อบัญญัติ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pronunciation</td>
<td>kʰɔ̂n ʔɔ́n ját</td>
</tr>
<tr>
<td>Part of speech</td>
<td>Noun</td>
</tr>
<tr>
<td>Definition</td>
<td>(กฏ) กฎหมายที่ห้องการบริหารส่วนท้องถิ่นตราขึ้น เพื่อใช้บังคับในเขตขององค์การบริหารส่วนท้องถิ่นนั้นๆ</td>
</tr>
<tr>
<td>Example</td>
<td>ข้อบัญญัติกำหนดขั้นเวลาเพื่อความสงบของบ้านเมือง</td>
</tr>
</tbody>
</table>

**Figure 2 Structure of an entry**

### 4.3.2 Sense Indicators

Wherever possible the first sense is a common one – usually the sense that most people would expect.

<table>
<thead>
<tr>
<th>ก่ก /kok/</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Noun 1</strong> (ไวยากรณ์) แม่ก หรือ คำที่เหล่าด้วย ก ข ค ฅ</td>
</tr>
<tr>
<td>◊ คำว่า 'ก่ก' เป็นคำที่อยู่ในแม่ก</td>
</tr>
<tr>
<td>2 โตกลับ, ลำกลับ</td>
</tr>
<tr>
<td>◊ ก่กสา</td>
</tr>
<tr>
<td>3 (หั้น) ซึ่งไม่กลมถุนของหนัง</td>
</tr>
<tr>
<td>◊ แม่สามเสือตัวยืนก่ก</td>
</tr>
<tr>
<td><strong>Verb</strong> แนวใช้กับก่ก</td>
</tr>
<tr>
<td>◊ ก่กอล, ก่กใส, ก่กลูก</td>
</tr>
</tbody>
</table>

**Figure 3 Sense Indicators**

### 4.3.3 Subject field

This dictionary disambiguates the sense by providing the general subject category of the word entry. Example:

(กฏ) กฎหมาย legal: รัฐธรรมนูญ |
(คณิต) คณิตศาสตร์ mathematical: คณิตศาสตร์ |
(วิทยาศาสตร์) วิทยาศาสตร์ scientific: ไวยากรณ์ |
(วรรณ) วรรณคดี literature: ฮัตถุน

**Figure 4 Subject field**
4.3.4 Register

Comments related to register, such as formal, impolite, informal, spoken, etc. can be provided.

**Abbreviations:**
- (จาษ) ราชสีห์ royal word: เสด็จ
- (พะษ) ไปสุภาพ impolite word: แตก
- (พุธ) ภาษาพูด colloquial language: กระแสที่ชี้ทิศทาง
- (ทางการ) เป็นทางการ formal language: ชาวพันธุ์
- (ทิม) ภาษาทิม dialect: กังจ่อง
- (เก่า) ทิมเก่า obsolete or old-fashioned: ตะวัน
- (อุปมา) อุปมาอุปมัย figurative: กลืนไม่ลง
- (หญิง) ผู้หญิง female: ค่ะ
- (ชาย) ผู้ชาย male: คุณ

**Figure 5** Register

4.3.5 Sense qualifier

Phrases such as แซ่ด้า (black sheep) indicate figurative language (อุปมาอุปมัย). (อุปมา) indicates the translation which is the literal meaning.

For example:
แซ่ด้า [kae dam] Noun (อุปมา) คนที่ทำผิดจากเพื่อนฝูง (ใช้ในทางไม่ดี) แซ่ด้าของครอบครัว

**Figure 6** Sense qualifier

4.3.6 Hypernym, Synonym, Antonym

The hypernym is a generic term, to whose field of application the headword belongs. It helps to disambiguate its sense. For example:

**Hypernym**
สัตว์ Noun (สัตว์) พันธุสัตว์ระหว่างลำดับม้า ต้องเป็นสัตว์ตัดขนาด ■ Verb ใช้จุบปลาน้ำ ต้องแยกลักษณะจากตัว

**Synonym (=) and antonym (#)**
แห้งกิ [log rák ] Verb คลื่นไอน้ำบนหนึ่ง = เฉยหน้า, บุช สำหรับข้อความ
ขาว [kaaw ] Adjective สีชัดสีต่าง; # ต่าง ≠ หน้าขาวมาก

**Figure 7** Hypernym, Synonym, Antonym
5. Conclusion

In conclusion, this research charts an approach to Thai dictionary compilation, using corpus-based approach to develop a model of good linguistic practice from the corpus. It integrates the findings to produce more informative kind of dictionary entries, and this demonstrates some effective ways of monolingual learners’ dictionary in Thai. It argues that Thai dictionaries should be based on corpus data of Thai literature. Frequency data and concordance can play an important role in the design, evaluation, and revision of Thai dictionaries in order to sure that the information recorded in the dictionaries is authoritative and backed up by empirical evidence.

References

Lexicological Module of Do It Yourself Corpora for Turkish

Bülent ÖZKAN
Mersin University
ozkanbulent@gmail.com

Abstract

The aim of this paper is to introduce the lexicological module of a corpus platform, which is flexible according to the research questions of scholars, and which is specific to the scholar; is user friendly, and corpus-database. Considering this perspective, it is expected to provide a corpus platform in which the results of the research can be obtained in a functional way. In addition to these, in lexicographic studies, the system can present a corpus database for users as a corpus output.

The study, which is named as To Built Do It Yourself Corpus for Turkish (DIYCT), supported by TÜBİTAK* 1005 New National Ideas and Products Research Support Group. In this study the Lexicological Module of Do It Yourself Corpora for Turkish will be introduced. Firstly the outline of compiling a corpus by using DIYCT will be introduced and using Lexicological Module will be present to the researchers by explaining adding application like etymological knowledge, collocations, run-ons, spelling and if necessary voice and picture file adding and also other steps of compelling a dictionary.

Keywords: lexicography, corpus linguistic, DIY corpora

1. Introduction

It has been proved that corpora are important resources for linguistics studies. Almost in all linguistics disciplines, corpora have achieved to open new areas of research or to bring new insights to many traditional research questions (Meyer, 2004).

Currently, applications of corpus linguistics are used mainly in lexicography and lexical studies in parallel with applied linguistics. Additionally, these applications are used in other fields such as grammatical studies, register variation and genre analysis, historical studies, translation studies, diachronic studies, language change, language learning and teaching, semantics, pragmatics, sociolinguistics, discourse analysis, stylistics and literary studies, forensic linguistics etc. (McEnery, 2006: 80-122).

* The Scientific and Technological Research Council of Turkey
From the earlier text collections to large linguistics database, named as corpus, the studies of lexicography evolved to the empirical perspective. “Today, advances in computer technology have given several advantages to the corpus-based lexicographic research over earlier work. First of all, computers have made possible to the collection and storage of big chunks of texts and so that analyses are not limited to sentence-length excerpts.” (Biber, 2002: 22). With the ability of designing large corpora is provided to describe the language for lexicographer. In addition to this, “… using corpora allows dictionary-makers to extract all authentic, typical examples of the usage of a lexical item from a large body of text in a few seconds.” (McEnery, 2006: 80).

The structure of corpus design provides particular knowledge about lexical item that are examined. According to metadata of the corpus or structured datum, researcher can get this knowledge automatically such as frequency, co-occurrence, and collocations, key words in context (KWIC); register, genre and domain and part-of-speech… etc. Researcher also can build semantic patterns of lexical units while querying lexical items from a corpus.

On the other hand, in corpus linguistics literature, many researchers (Kennedy, 1998; McEnery, T. et. al. 2006; Sinclair, 1991; McEnery et. al. 1996; Barnbrook, 1996) have underlined the importance of determination of the motivation of corpus design and according to this, they have emphasized the importance of research questions. Moreover, the corpus have evaluated as data sets, which are used in linguistics researches in corpus linguistics literature (Nesselhauf, 2005; Gries, 2006; Kawaguchi, 2004; Dale, 2000; Scott, 2006; Sterkenburg, 2003).

Experts, suggest to use specific softwares or web applications in lexicological corpus linguistics studies for standart-simple outputs such as frequency, KWIC etc. (Sinclair, 1991; Stevens, 1995; Todd, 2001); but these outputs cannot be adequate for lexicological studies. There are same specific patterns to form a headword in a dictionary. Some of these are spelling, pronunciation, inflections, word class, senses, definition, examples, usage, run-ons, etymology etc. (Jackson, 2002: 26-27; Hanks, 2003: 56-57). At this point, the important issue is to decide which of these patterns will be take place in a headword. The determinative approach is the aim of lexicological studies -in other word, research questions- and according to this, decision of dictionary-makers.

Even in this case, a dictionary-maker needs to use a corpus tool, which is adequate for completing her/his studies. Unlikely, lack of a database and/or a storage support is the biggest disadvantages of corpus tools. Another difficulty that dictionary-makers face to face is the learning period and adaptation period of these tools.

2. DIY Corpora Project for Lexicological Issues

The aim of the project DIYCT is provide a database-supported corpus, which can be shaped according to the research questions, and this corpus is specific and user friendly
for the scholar. Considering this perspective, this corpus platform can generate flexible reports according to tagged corpus units for the linguistics researcher. By using DIYCT, linguistics researcher can built a Turkish corpus and also can tag this corpus via the Lexicological Module, Semantic Module, Syntactic Module, Morphological Module, Discourse Analysis Module, and Learner Corpora Modules (http://kld.mersin.edu.tr/index.php?dil=en).

After building a corpus in Lexicological Module, researchers can form the headword patterns, according to their research questions and can tag the headwords pattern in the Lexicological Module. Researchers can tag the patterns of headwords to the units that are determined via the Lexicological Module.

As soon as the corpus is built via DIYCT, the platform automatically process the texts as lemma, deduplication, and makes morphological analysis, frequency analysis, parse the sentences, shows the n-gram and collocation computing in a few seconds.

After these steps, the data that will be used for lexicological research becomes available for headwords tagging. These processes can be summarised as below:

I. Building a corpus
   - Definition of layers and metadata of the corpus
   - Uploading the texts to corpus

II. Standard corpus processes
   - Lemmatization and stemming
   - Deduplication
   - Morphological analysis
   - Frequency analysis
   - Parsing the sentences
   - N-gram and collocation computing

III. Lexicological Module
   - Lemmatization and stemming
   - Definition of flexible tags of the headwords (spelling, pronunciation, inflections, word class, senses, definition, examples, usage, run-ons, etymology etc.)
   - Data processing
   - Dictionary-makers can report the headwords as output from the module via flexible tagging.

3. Lexicological Module Processes

As it is mentioned above (III. Lexicological Module) this process consists of three steps. The first step is lemmatization and stemming, second one is definition of flexible tags headwords, and the last one is data processing. These stages are shown in Figure 1.
Figure 1 Processes of lexicological module.

- **Lemmatization and stemming process:**

This process is the main source of headwords of dictionary via lemmatization and stemming interface. Because of agglutinative structure of Turkish, stemming is important to determine the headwords of dictionary.

In conclusion, respectively lemmatization, stemming, and headwords steps can be structured for a dictionary via the interface of DIYCT software (Figure 2). The stemming list is shown in Figure 3.

Figure 2 Interface of lemmatization and stemming
Figure 3 Stemming list of headwords.

- **Definition of flexible tags of the headwords (spelling, pronunciation, inflections, word class, senses, definition, examples, usage, run-ons, etymology etc.):**

The flexible tagging interface of DIYCTL (Figure 4) can determine the dictionary structure tags such as spelling, pronunciation, inflections, word class, senses, definition, examples, usage, run-ons, etymology etc. after the lemmatization and stemming processes in the basis of research question and/or dictionary researcher’s (dictionary-maker) aim. On the other hand, other tagging such as voice and picture files can be added through this interface.

Figure 4 Interface of flexible tagging.

- **Data processing:**

In this stage, the stemmed headwords are listed for data processing. The data processing sample of “baba” (father) headwords is shown in Figure 5. The tags that have been determined at the **Definition of flexible tags** stage, like meanings of headwords and choosing examples (see also Figure 6.); part of speech tagging, pronunciation/spelling of headwords, compound word structure, etymology, picture
file (see also Figure 7.) and collocational structures (see also Figure 8.) can be tagged via these interfaces.

**Figure 5** Data processing of “baba” (father) headwords.

**Figure 6** Interface of meanings and examples tagging.

**Figure 7** Interface of etymology, part of speech, pronunciation/spelling, etc.
Flexible report:

Consequently, the data processing of headwords, researcher can take flexible report from the Lexicological Module of DIYCT as an output (see Figure 9). System allows flexible tagging for dictionary-makers to built headwords and according to this flexible tagging, the determined headwords are always available as an output (doc. docx etc.) for the researcher.

Figure 9 Flexible headwords report of lexicological module of DIYCT.

Acknowledgments

This study, which is presented and named as To Built Do It Yourself Corpora for Turkish, supported by TÜBİTAK 1005 New National Ideas and Products Research Support Group. Many thanks for this contribution to TÜBİTAK.
References


Özkan, B. et. al. (2016). Result report of Do It Yourself Corpora for Turkish Language. Project number: 114E791. Mersin/TURKEY


From Ancient Manuscript Chinese Character Dictionary of

TenreiBanshōMeigi to Deciphered Electronic Text

Yuan Li
Hokkaido University
liyuansapporo@yahoo.co.jp

Abstract

This paper provides a detailed report of the transliteration, the full-text deciphered electronic text, and publication systems, based on the ancient manuscript Chinese character dictionary of KōsanjibonTenreiBanshōMeigi 高山寺本篆隷万象名義.

The TenreibanshōMeigi(after 830) is an abridgement of the YuanbenYupian 原本玉篇, which was compiled in 543 by Guyewang 顧野王 in China. Because the Yupian was lost in China, and only 7 volumes have preserved in Japan, this is a very valuable resource for attaining a glimpse of the Yupian’s original form. With the spread of Unicode, chief Chinese dictionaries like ShuowenJiezi 説文解字 and Guang Yun 廣韻 have been provided open access. Meanwhile, though many old manuscripts are well preserved in Japan, owing to the problems such as variant characters and erratum, the data origination of old manuscripts is falling behind. In the research group to which author is affiliated, a general database named Integrated Database of Hanzi Dictionary in Early Japan (abbreviated HDIC) is being constructed. TenreiBanshōMeigi is a part of HDIC, and the full-text was released (http://hdic.jp/) in 2016. This paper focuses on the three points related to TenreiBanshōMeigi below:1.Ancient Manuscript, 2.Compilation Style, 3.Unicode based full-text.

Keywords: KōsanjibonTenreiBanshōMeigi, Yupian, Chinese character dictionary, Open data

1. A Digital Model of Ancient Manuscript Chinese Character Dictionaries

Ancient manuscript Chinese Character dictionaries are different from the modern ones in that they usually lack clear explanatory notes or a standardized compilation style. However, it may be shown that the contents have a hidden systematic design,
for example, in how the volumes are separated, the setting of the radicals, and also the arrangement of the characters.

This paper provides a detailed report of the model used to decipher an ancient manuscript Chinese character dictionary, *TenreiBanshôMeigi* (or referred to as *Banshô Meigi* if necessary), and convert it to an electronic text.

Therefore, there are three steps necessary to realize the decipherment, transliteration, and digitization.

① Confirm the contents of the dictionary by the reproduced text, sometimes inspection of the original documents is necessary.
② After the confirmation in ①, determining and describing the systematic design elements of the dictionary to reveal the hidden compilation style visible.
③ After ① and ②, finishing the digitization of the text using Unicode.

---

**Figure 1** Model of the Digitization of Ancient Manuscript Chinese Character Dictionaries

In previous research, the list of headwords (Miyazawa 1997), and transliteration of the entire contents (Lv 2007) were published. Figure 2 shows the stages in deciphering an ancient manuscript Chinese character dictionary. The handwriting decipherment and printing decipherment are pre-stages of the electronic text. The base of the printing decipherment must be included in the database in some way; however, it must be processed in such a way as to lead to its release as open data.

**Figure 2** Stages in Deciphering Ancient Manuscript Chinese Character Dictionaries

Making the e-text public makes it possible to research and revise the text,
confirming and saving of revisions thereby becomes more practicable. By using information processing methods, the epoch-making progress in the study of the ancient dictionaries may be realized.

The issues shown in Figure 1 above will be discussed in this paper. In Section 1, the whole model and the first issue are discussed, while Section 2, 3 and 4, the second and third issues will be discussed.

2. The Compilation Format of the KōsanjibonTenreiBanshōMeigi

2.1 The Radical System

The KōsanjibonBanshōMeigi is composed of six books, each of which is divided into volumes. Each volume covers several radicals. (Some Radicals with many headwords extend over two or more volumes, such as 草 ‘grass radical’ and 木 ‘wood radical’.) On each page, the items are written in two vertical columns, with six lines from right to left. In the frame, the headwords appear first in large seal and clerical script, followed by annotation. The SongbenYupian compiled around 11C, continued the system and content of the YuanbenYupian. While there are many parts common to the two dictionaries, there are many differences in format and content.

The components of the BanshōMeigi can be separated into two parts. The first part, called the original part, was compiled by Kūkai in 9C and includes Books 1-4. The second part, the sequel, was compiled by someone else and includes Book 5-6. The dictionary was organized following different principles from the YuanbenYupian, and it would appear the whole works was to be divided into 100 volumes. However, the SongbenYupian is organized into 30 volumes in all, following the same principles of organization as YuanbenYupian.

Comparing the three dictionaries (the YuanbenYupian, BanshōMeigi, and SongbenYupian), their structures can be described as follows:

YuanbenYupian: Volume-Radical-Headword-Annotation
BanshōMeigi: Book-Volume-Radical-Headword-Annotation
SongbenYupian: Book-Volume-Radical-Headword-Annotation

The YuanbenYupian is a scroll-style book, while BanshōMeigi and SongbenYupian are booklets. The BanshōMeigi and SongbenYupian are both the digest of the YuanbenYupian. The correspondences between the three dictionaries are indicated in Table 1. The original part and sequel are organized differently in that the original part is divided more specifically and is differently organized than the YuanbenYupian; however, the sequel is organized the same as YuanbenYupian.
Table 1 Comparative table of BanshōMeigi and SongbenYupian

<table>
<thead>
<tr>
<th>BanshōMeigi</th>
<th>SongbenYupian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Books</td>
<td>Radical</td>
</tr>
<tr>
<td>1</td>
<td>一〜目</td>
</tr>
<tr>
<td>2</td>
<td>目〜心</td>
</tr>
<tr>
<td>3</td>
<td>心〜中</td>
</tr>
<tr>
<td>4</td>
<td>木〜禾</td>
</tr>
<tr>
<td>5</td>
<td>禾〜氐</td>
</tr>
<tr>
<td>6</td>
<td>山〜亥</td>
</tr>
</tbody>
</table>

2.2 Item Structure

The Item Structure of BanshōMeigi is relatively simple.

Item

Headword A  Annotation

Fanqie B  Zhiyin b  Meaning C  Variant Form a

ItemContent: glyph A Headword a Variant Form pronunciation B Fanqie b Zhiyin\(^1\) meaning C Meaning

A 譄  B 子恒反。C 加也。a 增字。(BanshōMeigi Book3, f.15r)

Headword A 譄, indicates the character, fanqie 反切 B 子恒反, indicates the pronunciation. C 加也, the meaning, and annotation a 增字 provides variant forms.

\(^{1}\)A homophonous character.
3. Headwords of $KōsanjibonTenreiBanshōMeigi$

3.1 Classification of Headwords

A Chinese character dictionary contains Chinese characters of that era, to which explanations concerning pronunciation, definition, and orthography are added, and is a helpful book for Chinese character learning and as a standard of Chinese characters in social life. The Chinese characters contained are the objects to be explained in the dictionary and the framework of the dictionary, but different terms are used, depending on the country and the researcher. For example, a headword is referred to mainly as Zitōu 字頭 in China, and as hyōji 標字, hyōshutsuji 標出字, oyaji 親字, keishutsuji 掲出字, and midashiji 見出字 in Japan. In this paper, the term, “headword (掲出字)” will be used in referring to “a Chinese character to be explained in a Chinese character dictionary”.

This paper attempts to comprehensively examine the headwords that may have existed in BanshōMeigi and the YuanbenYupian, assuming that BanshōMeigi inherits the YuanbenYupian faithfully, but before that, I will define the term, “headword.” It is believed that the headwords in BanshōMeigi are defined either in a narrow sense or in a broad sense, considering their style and the fact that they are inherited from ShuowenJiezi and the YuanbenYupian, which are preceding dictionaries. Headwords in a narrow sense refer to the headwords in large clerical script in BanshōMeigi, and headwords in a broad sense include the headwords in other scripts, including embedded headwords and omitted headwords.

In $KōsanjibonBanshōMeigi$, the items are arranged into two-tiered columns with six lines per page described in Section 2. For each item, the headword is provided in seal script (only approximately 6% of the total are attached with headwords in seal script) and in large clerical script, and followed by explanations. Judging by the style of $Kōsanjibon$, the characters written in large clerical script are referred to as headwords in a narrow sense (headwords in clerical script).

On the other hand, there are some characters that are headwords in large print in the Yupian fragmentary volumes but are included in annotations without being headwords in clerical script in BanshōMeig. Miyazawa (1977) recognizes this type of character as “a headword included in an annotation in BanshōMeigi,” and it is inserted at a lower position by one character than the original position in the list of headwords. In this paper, this is defined as embedded headwords among headwords in a broad sense. This embedded headword has two types.

One type is provided with phonetic annotation, or with phonetic and semantic annotations, and is believed to have been originally a headword. This is referred to as embedded headwords A in this paper.

The other type is not provided with phonetic annotation or semantic annotation, and only shows an allographic relationship. This type is referred to as
embedded headwords B in this paper. Those that correspond to embedded headwords B exist as variant headwords in the *YuanbenYupian*. Ueda (1970) refers to those as “allographs,” and the them is defined as: “it is used as a collective term for characters that are regarded as headwords in *Yupian* but that are the same characters as the ones above those without *fanjie*, such as ancient script, large seal script, and variant characters.” In this paper, variant headwords in the *YuanbenYupian*, which are included in annotations in *BanshōMeigi*, are treated as embedded headwords B.

In contrast, a headword dropped out in *BanshōMeigi* is referred to as omitted headwords. There are also two types of omitted headwords. One is a headword that is omitted but has an annotation, and this is referred to as omitted headwords A. The other is a character that is believed to have had a text in *BanshōMeigi* because there exist texts in *ShuowenJiezi*, the *Yupian* fragments, citation from *Yupian* in *ShinsenJikyō*, and *SongbenYupian* (excluding the augmented parts), and this is referred to as omitted headwords B. In Miyazawa (1977), this type of characters are recognized as the characters dropped out in *BanshōMeigi*, the characters that exist in *SongbenYupian* (excluding the augmented parts), the characters seen in *ShinsenJikyō* or *ShuowenJiezi*, which probably existed in *Yupian* originally, and reflected in “the list of headwords.”

In this paper, the headwords in *KōsanjibonTenreiBanshōMeigi* are classified and recognized as above. First, the headwords written in large clerical script are counted. Then, after embedded headwords B included in annotations are classified and recognized as embedded headwords A and embedded headwords B by the difference of their characteristics, characters that seem to be omitted in *BanshōMeigi* are classified into omitted headwords A and omitted headwords B based on the evidence. The criteria of classification will be mentioned in the following section.

### 3.2 Analysis of embedded and omitted headwords

#### 3.2.1 The issues of embedded and omitted headwords related the *YuanbenYupian*

In this paper, I will define a concept of embedded headwords and omitted headwords and organize them in order to determine the number of headwords, because I would like to reconstruct the *YuanbenYupian*, which *BanshōMeigi* was based on, as much as possible, keeping the problems that occurred in the process of transcription up to *KōsanjibonTenreiBanshōMeigi* in mind. It is especially necessary to organize embedded headwords B, in order to grasp the situation of variant characters in the *YuanbenYupian* through the examination of the headwords and their orthography in *BanshōMeigi*.

---

2 Embedded headwords included in the annotation of a certain listed item exists not only in *BanshōMeigi* but also in other old dictionaries and definitions. The concept of embedded headwords is convenient to sort out this type of problems.
There have been few studies that focus on the problems of embedded headwords and omitted headwords, and especially for embedded headwords, not enough attention has been paid. Embedded headwords A have often been counted without distinguishing from headwords, and embedded headwords B have often been not recognized as headwords and been excluded from the counting.

I will examine each of embedded headwords A, embedded headwords B, omitted headwords A and omitted headwords B below.

### 3.2.2 Embedded headword A

First, we must confirm the part of the annotation that unrelated to the main headword, and next recognize those provided with phonetic, or with phonetic and semantic annotations as embedded headwords. There are 150 items in all, 105 items in Book 1, 1 in Book 2, 1 in Book 3, 2 in Book 4, 40 in Book 5, and 1 in Book 6, of which we will see several examples below. They are given in the order of embedded headword, main headword, and annotation of the main headword (including the embedded headword). The embedded headword is emphasized with underlining, punctuation marks have been provided by the author.

| (1) | 柴 祡 仕佳反。 | 祭天曰焚、祟也。[柴;仕佳反。燔。] 篾、柴文。(Book 1, f. 17r) |
| (2) | 祼 瑣 湯典反。 | 古文。祼也。祼、古文。[祼;去理反。] (Book 1, f. 21v) |
| (3) | 泛 汲 或駃也。 | 泛;孚劔反。舟流皃。] (Book 5, f. 93v) |
| (4) | 潃 滕 上同。 | 潃;徒見反。萍也、澤也。] (Book 5, f. 97v) |

The items of the Yuanben Yupian (Volume 19 Radical 氷’water’) corresponding to (3) and (4) are shown below, as per the order in the Yuanben Yupian)

(5) (淦) □欠部。又音古暗□…□□□曰淦、水所出西入湖漢。

(6) 潃説文或淦字也。廣雅淦取也。

(7) 泛孚劔反。國語泛舟于河、賈逵曰泛浮也。毛詩泛彼栢舟、傳曰泛RetVal流兒也。

又曰泛々其景、傳曰泛RetVal駃疾而不疑也。説文從之聲也、此亦汽字、相似而不不同。漢書或以爲亹字、亹覆也。方音鰣反。亹駕之馬是也。在西部。
「汵」 and 「泛」 are continued in the *YuanbenYupian*, probably because they have similar forms, although they have no connections in meanings. 「汵」 is a variant script of the previous headword 「淦」. However, in the *BanshōMeigi* 「泛」 is an embedded headword A in the annotation of 「汵」, probably because 「泛」 had been omitted when the dictionary was compiled or written and then was supplied later. In Book 5, the number of items of this type can be confirmed as 27. (Those not present in the fragments of *YuanbenYupian*, have been checked in the *SongbenYupian*.)

### 3.2.3 Embedded headword B

An embedded headword B is the one that included in the annotations of another item, without phonetic annotation or semantic annotation, and only serves to show an allographic relationship. There are 561 such items in all: 183 items in Book1, 11 in Book2, 29 in Book3, 58 in Book4, 269 in Book5, 11 in Book6. We will see several examples below. An embedded headword is highlighted with underlining, and punctuation marks have been provided by the author.

8) 庙靡召反。兒。〔庿：古廟。〕 (*BanshōMeigi* Book 6, f. 9v)

9) 庙靡召反。尚書七世之庿可以観德、孔安國曰天子七庿、有德之主則爲宗其庿

干不毁。尔雅室有東西廂曰廟。韓詩鬼神所居曰廟神。礼記夫子七庿、三昭

三穆与大祖之庿而五、大夫三庿、一昭一穆与大祖之庿而三、士一庿。鄭玄

曰此同制也、殷即六庿之也。白虎通曰先祖之尊兒所在也。

(*YuanbenYupian* Volume 22 Radical 广 ‘wide’)

庿 説文古文廟字也。（*YuanbenYupian* Volume 22 Radical 广 ‘wide’）

10) 庙靡召切。宗廟也。（*SongbenYupian*, Book 3, f. 15v）
In the above examples, the part “庿:古廟” indicates anallographic relationship between “庿” and main headword “廟”, which can be confirmed in the YuanbenYupian. This has been processed as an embedded headword B in BanshōMeigi, as well as in the ShuowenJiezi.

To recognize an embedded headword B, it is necessary to compare the item with the YuanbenYupian; however, as the greater part of YuanbenYupian has been lost, it must be compared with such other documents as the SongbenYupian.

### 3.2.4 Omitted headword A

This refers to a headword that is omitted but for which an annotation is present, including 15 items in all (垚・巖・壇・里・囲・嘆・右・뙇・讃・邇・桍・様・箇・栽). In details, there are 4 such items in Boo1, 5 in Boo2, 2 in Boo3, 3 in Boo4, and 1 in Boo5. An example is shown below in the order of omitted headword A, annotation, location.

(12) （垚）呼玦反。深也、空也。(BanshōMeigi Book 1, f. 36v)

垚呼玦反。深也、空也。□穴高字。(ShinsenJikyō Volume5, f. 22v)

垚呼決切。深也、空也。亦作粫。(SongbenYupian, Book 1, f. 15v)

In the above example, in the BanshōMeigi, the headword has been omitted, and the first character of the fanqie is given as the headword. The content is confirmed in In ShinsenJikyō and SongbenYupian as above.

### 3.2.5 Omitted headword B

Based on Miyazawa (1977) and Ikeda (2014), the term “omitted headword B” refers to a headword that is believed to have had a text in BanshōMeigi because there exist texts in the ShuowenJiezi, the YuanbenYupian fragments, and citations from Yupian in ShinsenJikyō and SongbenYupian(excluding the augmented parts),
and this category includes 476 items in all as follows: 34 in Book1, 92 in Book2, 110 in Book3, 132 in Book4, 35 in Book5, and 73 in Book6. Many examples of omitted headwords B are allographic headwords in *YuanbenYupian*.

### 3.2.6 Entry Counts in Tenrei*BanshôMeigi*

In this paper, the headwords that can be confirmed in *BanshôMeigi* and recognized as ordinary headwords, embedded headwords, and omitted headwords A, constitute 16,523 items in all. If we add omitted headwords B added, the total is 16,999 items. It cannot be confirmed, however, whether all the omitted headwords B are present *BanshôMeigi*. Currently, the most that can be said is that the headword count of *BanshôMeigi* ranges from 16,523 to 16,999.

### 4. Unicode-Based Full-text

#### 4.1 Classification in Full-text

The headwords can be grouped into three basic classes: the Regular Headwords (about 88%), Embedded Headwords (about 4%), and Omitted headwords (about 3%). We may then add the information of seal or clerical script to these three groups in a subordinate position (Group a, b, c, mentioned later).

The remaining group is the supplementary classification of Variant scripts (about 5%) supplied in the *SongbenYupian*. This is a measure to take into consideration the *YuanbenYupian* and *SongbenYupian* as well, which also belong to the Yupian-system dictionaries into consideration as well.

Full details of the four groups are given below:

**Entry_type**

**Group a** -- Regular headword  
1 Regular (= Regular_clerical)  
2 Regular_seal  

**Group b** -- Embedded headword  
3 Embedded_clerical  
4 Embedded_omitted  
5 Embedded_seal  

**Group c** -- Omitted headword  
6 Omitted  
7 Omitted-regular  

**Group d** -- Variant of the SongbenYupian  
8 Songben-Yupian

Group a, basically contains Regular Headwords written in larg clerical script. They are classified into two classes according to the presence of seal-script forms.
1 Regular_clerical means those with only clerical script headwords.
2 Regular_seal means those with both clerical script and seal script headwords.

Regular _clerical is abbreviated Regular because it is the most typical class of headwords.

Group b, basically contains headwords embedded in annotations. Moreover, font information is added.

3 Embedded_clerical means the font is clerical.
4 Embedded_omitted means it is omitted.
5 Embedded_seal means the font is seal

Group c, basically contains omitted headwords. In case of only the headword is omitted while the annotations are present, it will be noted as regular.

6 Omitted means the headwords exist in other documents, and thus should be considered as having also been present in the Banshō Meigi.
7 Omitted_regular

Group d, basically contains headwords for which Variant Forms (8 Songben-Yupian) are provided in the SongbenYupian.

4.2 Processing Conditions Based on Unicode

Under the current computer environment, implementing Unicode allows multiple Chinese character documents to be processed. However, the results of previous studies were published by handwritten form. For example, the published identifications of the headwords of Banshō Meigi in Shirafuji 1977, Miyazawa 1977, and Ueda 1986, are all by handwritten.

Over 70,000 Chinese characters could be processed after the release of Unicode 3.1 in 2001. Lv (2007) was published against this background, in which the chief Chinese dictionaries like the Shuowen Jiezi and the Guangyun were released under open access.

The Konsanjibon Tenrei Banshō Meigi database was constructed within the Unihan range offered by Unicode. In transliteration it follows the standards of the Kangxi Zidian 康熙字典 scripts, and some of the Characters were processed following the original script. Also, some Characters are indicated with Ideographic Description Sequence (IDS), which cannot be processed by Unihan. IDS is a method that uses IDC (Ideographic Description Character, U+2FF0〜U+2FFB) and parts of characters to describe the whole character.

In fact the percentage of characters in these dictionaries that can be processed
with Unihan is as below: BanshōMeigi 99.2%, SongbenYupian 99.8%, and ShinsenJikyō 89.2%. The numbers of headwords that cannot be processed are 128 in the BanshōMeigi, 46 in the SongbenYupian, and 2,592 in the ShinsenJikyō. The details are shown in Tables 2 and 3.

Table 2 The Processing Situation of the BanshōMeigi (KTB) SongbenYupian (SYP) and ShinsenJikyō (TSJ)

<table>
<thead>
<tr>
<th>DB</th>
<th>Unicode Han</th>
<th>IDS</th>
<th>Others</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KTB</td>
<td>15,872</td>
<td>80</td>
<td>48</td>
<td>16,000</td>
</tr>
<tr>
<td></td>
<td>(99.2%)</td>
<td>(0.5%)</td>
<td>(0.3%)</td>
<td>(100%)</td>
</tr>
<tr>
<td>SYP</td>
<td>22,954</td>
<td>23</td>
<td>23</td>
<td>23,000</td>
</tr>
<tr>
<td></td>
<td>(99.8%)</td>
<td>(0.1%)</td>
<td>(0.1%)</td>
<td>(100%)</td>
</tr>
<tr>
<td>TSJ</td>
<td>21,408</td>
<td>1,512</td>
<td>1,080</td>
<td>24,000</td>
</tr>
<tr>
<td></td>
<td>(89.2%)</td>
<td>(6.3%)</td>
<td>(4.5%)</td>
<td>(100%)</td>
</tr>
</tbody>
</table>

Table 3 Details of the UnicodeHan Shown in Table 2

<table>
<thead>
<tr>
<th>DB</th>
<th>CJK</th>
<th>Ext.A</th>
<th>Ext.B</th>
<th>Ext.C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KTB</td>
<td>10,160</td>
<td>2,000</td>
<td>3,712</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(63.5%)</td>
<td>(12.5%)</td>
<td>(23.2%)</td>
<td>(0.0%)</td>
</tr>
<tr>
<td>SYP</td>
<td>13,386</td>
<td>3,243</td>
<td>6,325</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(58.2%)</td>
<td>(14.1%)</td>
<td>(27.5%)</td>
<td>(0.0%)</td>
</tr>
<tr>
<td>TSJ</td>
<td>15,048</td>
<td>2,400</td>
<td>3,960</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>(62.7%)</td>
<td>(10.0%)</td>
<td>(16.5%)</td>
<td>(0.0%)</td>
</tr>
</tbody>
</table>

4.3 TSV Information

We adopted the TSV data format. A tab-separated values (TSV) file is a simple text format for storing data in a tabular structure. Each record in the table is one line of the text file, and each field value of a record is separated from the next by a tab character. The TSV format is thus one type of the more general delimiter-separated values format. TSV is a simple file format widely supported in both lexicography and information processing that is often used in data exchange to move tabular data between different computer programs that support the format.

For the reference of researchers in other countries and areas, we put the TSV Format Information at the head of the open data. The numbers 01~10 are for convenience of explanation as follows: The information encoded 01~10, can be divided into three groups by their contents.

I. Basic Information: Location, system, structure information [01-03],
transliteration of headwords and annotations in the *BanshôMeigi* [04, 07].

II. Relevant Information: Information on the location of the corresponding entries in the *Yupian* systematic dictionaries (*YuanbenYupian* and *SongbenYupian*) [08, 09].

III. Revision Information: Classification of headwords by the author, differences with the identification in previous researches, revision comments [05, 06, 10].

I is the original information in the *BanshôMeigi*.

II is the corresponding information in the relevant dictionaries.

III is the revision information with references to previous research, authors, and relevant documents, respectively.

**TSV Format Information:**

01  TBID (v_www_xyz) : Book(v), leaf(www), recto-verso(x), line(y) and number(z)

02  TB_vol_radical (xx#yyy): Volume(xx) and radical number(yyy)

03  TB_radical : Radical of Chinese character

04  Entry: Headword

05  Entry_type: For details, refer to the following section

06  Entry_diff: Differences of transliteration with other scholars

07  TB_def: Definition of pronunciation, meaning and variant(s)

08  SYID (vwwwxyyyzz): Book(v), leaf(www), recto-verso(x), line(yy) and number(z)

09  YYID (Ywwxxxxyy-z): Volume(ww), leaf(xxx), line(yy) and number(z)

10  TB_remarks: Editor’s notes

### 4.4 Data Sample

Here are the details of the TSV Data for the headword 哥 (Table 4).

**Table 4** Details of TSV Data for the Headword 哥

<table>
<thead>
<tr>
<th></th>
<th>TBID</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>TBID</td>
<td>3_024_A31</td>
</tr>
<tr>
<td>02</td>
<td>TB_vol_radical</td>
<td>v9#96</td>
</tr>
<tr>
<td>03</td>
<td>TB_radical</td>
<td>可</td>
</tr>
<tr>
<td>04</td>
<td>Entry</td>
<td>哥</td>
</tr>
<tr>
<td>05</td>
<td>Entry_type</td>
<td>Regular</td>
</tr>
<tr>
<td>06</td>
<td>Entry_diff</td>
<td></td>
</tr>
<tr>
<td>07</td>
<td>TB_def</td>
<td>古何反。詠言也。</td>
</tr>
<tr>
<td>08</td>
<td>SYID</td>
<td>a006b043</td>
</tr>
</tbody>
</table>
5. Conclusion

The issues ①②③ shown in the model below (Figure 3) have been discussed in case of *TenreiBanshōMeigiin* this paper.

*Figure 3* The Processes of the Model of the Digitization of Ancient Manuscript Chinese Character Dictionaries
These three steps mutually influence each other and indeed form a recycling-oriented process. Hereafter follow the details for each sub-process of these steps from different directions.

A ①→②→③
[Standard transliteration process of ancient manuscript Chinese character Dictionaries]
B ③→①
[Search, specify the items, or characters (navigation)]
C ①→③
[Revise the text as occasion arises, improve the precision of the deciphered text]
D ①→②
[Make the hidden compilation style visible]
E ③→②
[Verify the compilation style by patterning information of the text]

By repeating the processes above, we may realize continuing improvements in the study of this dictionary. We hope that the data, which has been made public, would be of practical use to researchers, in both pre-modern Chinese lexicography and information processing.

Acknowledgments

Massive thanks to Kōsanji Temple authorities Ogawa Chie (the previous chief priest) and Ishizuka Harumichi (the representative director) for permission to publish the decipherment text of Kōsanjibon TenreiBanshōMeigi (KTB). Sincere thanks to the chair of HDIC Shoju Ikeda.

References

Li, Yuan [李媛]. 2015. Embedded and Omitted Headwords: Issues Regarding Entry Counts in
Tenreibanshomeigi. Diacritical Language and Diacritical Materials 135: 37-56
Abstract

The official dictionary of the Indonesian language, KamusBesarBahasa Indonesia (KBBI), is published by BadanPengembangandanPembinaanBahasa (The Language Development and Cultivation Agency) or BadanBahasa, under the Ministry of Education and Culture of the Republic of Indonesia. The fourth edition of KBBI (Sugono 2008) has more than 92,000 entries and 100,000 senses and contains a wealth of linguistic information and cultural diversity of Indonesia. However, the data was available only in Microsoft Excel and Word files in exactly the same format as the one in the printed dictionary. Its online edition was only meant for basic word search by entry words. Thus, in order to create an online dictionary application which has advanced search capabilities, building a database is very vital: the data structure needs to be identified and the data itself needs to be cleaned so that it can be broken down based on its components. Atkins and Rundell (2008: 114) state that a database is one of the three main components of Dictionary Writing System (DWS). This paper describes our efforts in building the KBBI database in SQLite (www.sqlite.org) using Python programming language (www.python.org) and presents some applications for lexicographic and linguistic research and analysis. The KBBI database is employed for the online DWS application called KBBI DalamJaringan or KBBI Daring (https://kbbi.kemdikbud.go.id) (Kamajaya et al. 2017), the offline KBBI mobile applications in Android and iOS, and the printing of the latest, fifth edition of KBBI (Amalia 2016).

Keywords: KBBI, database, Indonesian language dictionary, machine-tractable dictionary

1. Introduction

KamusBesarBahasa Indonesia (KBBI) is the official dictionary for Indonesian,1

---

1 Indonesian (ISO 639-3: ind), called bahasa Indonesia (lit. “the language of Indonesia”) by its speakers, is a Western Malayo-Polynesian language of the Austronesian language family. Within this subgroup, it belongs to the Malayic branch with Standard Malay in Malaysia and other Malay varieties (Lewis 2009). It is spoken mainly in the Republic of Indonesia as the
published by BadanPengembangandanPembinaanBahasa (The Language Development and Cultivation Agency) or BadanBahasa under the Ministry of Education and Culture of the Republic of Indonesia. Up until present, KBBI is the most comprehensive and the most authoritative reference for the Indonesian language. The first edition of KBBI, published in 1988, has 62,000 entries. The number of entries increased to 72,000 or about 10,000 entries over three years in the second edition (1991). The third edition of KBBI, published in 2001, contains 78,000 entries and seven years later, the fourth edition of KBBI’s number of entries increased to more than 92,000. The latest, fifth edition of KBBI was released for the first time in 2016 in three formats: printed, online, and offline versions. These three versions are launched to meet the needs of all users. Figure 1 shows the printed version of KBBI from the first edition to the fifth edition. This paper describes our work in 2016 on making a database for the fourth edition of KBBI which is then employed for the printed, online, and offline versions of the fifth edition of KBBI.

Regarding the online KBBI before 28 October 2016, it used the data from the third edition of KBBI and allowed searches only by headwords. The search results were presented exactly in the same format as the one in the printed version, i.e. using bold or italic typefaces and different punctuations, such as colons and semicolons (see Section 2 for the details of the formatting effects). These formatting effects serve only as stylistic presentations and do not distinguish the fields or their structure explicitly. For example, to look up mengacang, a user must first look up the root word (kata dasar) kacang, as shown in Figure 2. This may present some difficulties if the user is not familiar with Indonesian morphological rules. The users cannot perform more targeted searches and computer applications cannot utilize the data fully. This can be overcome by identifying the data structure, cleaning the data, and breaking it down based on its components or structure of dictionary entries.

We identify the data structure and break it down using regular expressions in Python programming language (www.python.org). The results are converted into a SQLite database (www.sqlite.org) to facilitate more specific and targeted word lookup and analysis. Lim et al. (2016) mention the categorization of lexical resources in terms of their digital readiness for natural language processing (NLP) work, from paper dictionaries, machine-readable dictionaries, machine-tractable dictionaries, to semantic rich resources. Paper dictionaries are traditional dictionaries printed on paper. They are only for human consumption. The contents are presented with text formatting effects and organized by headwords. Machine-readable dictionaries (MRDs) are digitized versions of the original paper-printed versions and are the most common form of electronic dictionaries, which retain the text formatting styles. The previous online KBBI, as shown in Figure 2, was an MRD. Machine-tractable
dictionaries are MRDs with machine-tractable structures, i.e. all fields and hierarchy of the entries are specifically marked and delineated, such that different information can be identified and extracted. Our work was to bring the KBBI to the level of this digital-readiness. Semantic rich resources are machine-tractable dictionaries with semantic information for each sense entry. They are very useful for NLP tasks, such as text categorization, sentiment analysis, and information extraction. However, this is outside the scope of our work.

Figure 1 KamusBesarBahasa Indonesia (KBBI), from the first to the fifth edition

Figure 2 Screenshot of the online KBBI before 28 October 2016

2. The KBBI dictionary format

KBBI is a general dictionary whose macrostructure has a hierarchical order. The schematic is arranged by placing the basic form (the root word or kata dasar) as the headword or the lemma. The information fields in an entry structure include a headword or lemma; variant forms; pronunciations; labels: parts-of-speech, styles, languages, domains, idioms, abbreviations; sublemmas/subentries: derived words, multiword expressions (MWEs) including compounds, idioms, and proverbs; definitions; cross-references; examples; scientific names; and chemical formulas.

Figure 3 shows us that the headword or the lemma is in bold type with periods for syllabification, followed by the pronunciation, surrounded by slashes. The part-of-speech label is written in italic type, following the
pronunciation. If there is more than one definition phrase in one sense, the
definition phrases are separated by semicolons. If there is an example, a colon
is put after the last definition and followed by a space; the example is in italic
type. The lemma is represented by two hyphens in the example. If a lemma is
an abbreviation, a label for abbreviation is placed before the definitions. If
there is more than one sense, a polysemy number is written in bold type before
each sense and senses are separated by semicolons. If the definition is in a
foreign language, it is written in italics. Figure 4 illustrates these formatting
effects. If a lemma has a chemical formula or a scientific name, it is written
after the definition and preceded by a semicolon, as shown in Figure 5. The
scientific name is written in italic type. Some numbers in the chemical formula
are subscripted.

If a lemma is homonymous, a homonymy number is placed before the
lemma in superscript bold type. If a particular label is appropriate for every
sense, it is written before the first polysemy number. If a label is appropriate
only for a particular sense, it is written after the respected polysemy number for
that sense. Subentries, such as compounds and derived words, are in bold type.
Subentries are separated by semicolons. If a derived word has an example, it is
represented by a tilde in the example. Figure 6 shows these formatting effects.
There is a special feature which distinguishes KBBI from other monolingual
dictionaries, i.e. the order of the derived words is not arranged alphabetically,
but in accordance with the paradigm of word formation. For example,
tinju “boxing”, as a lemma, has meninju “to box” as a transitive verb followed
by peninju “boxer”, peninjuan “act/process of boxing”, and tinjuan “the result
of boxing”. This sequence of verbs, actors, acts/processes, and results is called the
paradigm of word formation.

If a lemma or a sublemma appears in proverbs or idioms, the lemma is
represented by two hyphens, while the sublemma is represented by a tilde, same
as in the example field. Both the proverbs and the idioms are in italics,
followed by a comma and pb for proverbs (pb stands for peribahasa) or ki for
idioms (ki stands for kiasan), as shown in Figure 7. For cross-references, if a
lemma is non-standard, a right arrow is placed after it, followed by the standard
lemma in bold type. If a lemma is a part of an idiomatic compound, it is
followed by lihat “see” and the cross-referenced lemma printed in bold type
(see Figure 8). Up until the fourth edition of KBBI, the dictionary data with the
formatting effects mentioned above was available only in Word and Excel files.
The following section describes our work in breaking down the components
based on the formatting effects.

<table>
<thead>
<tr>
<th>lemma</th>
<th>pronunciation</th>
<th>POS label</th>
<th>definition sentence(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mereka</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3 Example entry mereka “they”
**Figure 4** Example entry AD

**Figure 5** Example entries digitalin and digitalis

**Figure 6** Example entries kaul “vow”

**Figure 7** Example entry karam “shipwrecked”
3. Cleaning-Up, Conversion, and Database Creation

The cleaning-up and conversion processes of data in Word and Excel files are quite tricky. This is because the data available in Word and Excel are formatted text (i.e. some of the data are in bold type, some others are in italic type, some others are superscripted, subscripted, or any combination of those, as described in Section 2) and that we want to keep the format as we transfer the data into the database. Figure 9 shows a part of the Word and Excel files. The format, not just the text, is a part of the lexicographic information and must not be removed during the conversion process. Hence, simple data extraction from Excel cells and Word paragraphs to the database entries cannot be done as it would not retain the text format.

In order to retain the text format in the conversion, the formatting effects in the Word and Excel files must be read, too. Therefore, a Windows Form application named KBBICleaner (see Figure 10) is created using .Net Framework to help us complete the task. The program uses Word- and Excel-compatible Microsoft-created dynamic link libraries (.dll), namely Microsoft.Office.Interop.Word and Microsoft.Office.Interop.Excel, to extract the data from Word and Excel in the Rich Text Format (RTF). Figure 11 shows a part of the RTF file. Furthermore, to ease the cleaning-up process, the program is designed with three additional main functionalities: (a) File and string manipulation, (b) List of text filter and conversion, and (c) List of regex filter and conversion, explained in the following subsections.
3.1 File and String Manipulation

This function helps us determine which portions of the Excel file to be cleaned using the program. This is used primarily for cleaning up the Excel file as it contains multiple sheets with different (inconsistent) “Range” to be cleaned up (for example, in one sheet, there might be three columns of data while in another sheet there might be four columns). It also helps us determine which starting and ending strings (in RTF format) can be used as division of cells, lines, or paragraphs.

3.2 List of Text Filter and Conversion

This function helps us process the file with a list of text filter and conversion. Some data in the RTF format are not needed (for example, the header of the file and the unused format code) and some need to be changed (for example, indicators of bold type, italic type, superscript, and subscript formats). Thus, a list of text filter and conversion will greatly help us process such data. All entries in the list will be applied to the original text in a sequential fashion, i.e. from the top entry to the bottom entry.

If a user needs to remove a certain consistently unused string, he or she needs to specify it in the filter with no conversion value. In addition, if a user needs
to replace a certain consistently appearing string, he or she needs to specify it and the desired conversion value in the list. Moreover, if a user needs to change two or more different formats into a single final format, he or she can exploit the sequential behavior of the filter to convert the earlier format(s) to the uniformed format in a sequential fashion and convert the uniformed format to the desired (single) final format.

This list of filter can be applied with both (real-time) user input values and predefined, loaded .txt file containing the filter information to further help a user save his or her filter midway whenever he or she finds the list of filter non-final and wants to continue to do it conveniently next time.

3.3 List of Regular Expression Filter and Conversion

Similar to the list of text filter and conversion above, this function helps a user with a list of regular expression (regex) -instead of text- filter and conversion. Regular expression or regex is a language for specifying text search strings which requires a pattern that we want to search for and a corpus of texts to search through (Jurafsky and Martin 2009). The regex filter behaves the same way as the text filter: it obtains and converts the filtered text according to the given list in a sequential fashion. However, it filters and converts the filtered text using regex patterns instead of doing direct conversion. Thus, this function can simply be perceived as a more powerful version of its text filter counterpart.

Naturally, however, being made of a set of regexes, this regex filter and conversion is significantly slower than the text filter and conversion. For cleaning-up process of a text data as large-sized as dictionary data, the time difference can be significant. Thus, this function is meant to help us process unused or to-be-converted data which form certain patterns. For statically written data, although they can be processed by this filter, they should be efficiently processed using the text filter instead of this regex filter. Table 1 shows some examples of the conversion from Excel to RTF and Hypertext Markup Language (HTML) and Figure 12 shows a part of the HTML file, as a result of the filter and conversion process using KBBICleaner.

3.4 Cleaning-up

After we converted the RTF file to a HTML file using KBBICleaner, we found some inconsistencies in the formatting effects and we did some cleaning-up for the data. We observe that these inconsistencies in formatting are sporadic and are due to the manual formatting work by hand. Table 2 shows some of the inconsistencies we found. In addition, we modified some definitions in order to make the formatting more consistent and to extract more information, such as chemical formulas, scientific names, and examples. Some examples are shown in Table 3.

3.5 Breaking down the components and creating a database

We wrote a Python script to break down the components or fields for each dictionary entry based on the patterns and formatting effects described in
Section 2, using regex. Figure 13 illustrates the algorithm we used to extract a number of fields in an entry. To facilitate easier manipulation of the data, all broken-down components such as lemmas, definitions, and examples were exported to a SQLite database.

### 3.6 Dictionary data structure

The data structure of KBBI consists of four types of data: entry, sense, example, and category. The relationship between entry and sense, as well as the one between sense and example are one-to-many. The category is a list of descriptions or metadata for entry, sense, and example. Figure 14 illustrates the KBBI data structure. An entry can be a fixed expression (ungkapan) or a root word (kata dasar). A fixed expression should have at least one sense and one example. In this case, one fixed expression may have one to multiple senses and one sense may have one to multiple examples. A root word should have at least one cross-reference, one sense, one compound, or one derived word. In this case, one root word may have zero to multiple senses and one sense may have zero to multiple examples. A root word may also have variant(s), proverb(s), and idiom(s). A proverb or an idiom should have at least one sense. A compound should have at least one cross-reference or one sense. One sense may have zero to multiple examples. Similar to the root word, a derived word should have at least one cross-reference, one sense, or one compound. It may also have variant(s), proverb(s), and idiom(s). The root word can be in the form of compound if it can be a fixed and have derived word(s).

<table>
<thead>
<tr>
<th>Field</th>
<th>Excel</th>
<th>RTF</th>
<th>HTML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lemma</td>
<td>A, a</td>
<td>\b A f5 f22 b0 , f6 f822 b a f7 f5 f22 b0</td>
<td>&lt;b&gt;A, a&lt;/b&gt;</td>
</tr>
<tr>
<td>Label</td>
<td>n</td>
<td>\i n f5 f22 u0</td>
<td>\i n&lt;/i&gt;</td>
</tr>
<tr>
<td>Homonymy number and lemma</td>
<td>^ab</td>
<td>\b super f5 f22 nosupersub f5 f22 b0</td>
<td>&lt;b&gt;ab (1)&lt;/b&gt;</td>
</tr>
<tr>
<td>Chemical formula</td>
<td>Cu_3</td>
<td>Cu f9 f22 sub 3 f5 f22 Cu sub 3 /nosupersub</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1** Some examples of the conversion

<table>
<thead>
<tr>
<th>Type</th>
<th>Example Before cleaning-up</th>
<th>After cleaning-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>incomplete syllabification</td>
<td>&lt;b&gt;(ke)ro.boh.an&lt;/b&gt;</td>
<td>&lt;b&gt;(ke)ro.boh.an&lt;/b&gt;</td>
</tr>
<tr>
<td>a semicolon should be a colon before an example</td>
<td>...pangkatdsb)_i_j~nyasbgduta...</td>
<td>...pangkatdsb)_i~nyasbgduta...</td>
</tr>
<tr>
<td>a comma should be a semicolon separating examples</td>
<td>...spt air mengalir: _i_j--udara_i_j --lalulintas&lt;/i&gt;:...</td>
<td>...spt air mengalir: _i_j--udara&lt;/i&gt;:... --lalulintas&lt;/i&gt;:...</td>
</tr>
</tbody>
</table>

**Table 2** Some inconsistencies in the KBBI format and the cleaning-up process
Pronunciations should precede labels:

- /gadéh/belanak
- /gadéh/n

Ki should be written in an idiom:

- /tidak
- /tidakbukan
- /yg

A comma should precede pb in a proverb:

- ...taklapukolehhuanpb
- ...taklapukolehhuan,pb

Scientific names should not be put inside brackets:

- /Aquilariamalaccensis

---

**Table 3** Some modifications in the KBBI definitions

1. Move chemical formulas to the end of the definitions, preceded by a semicolon

   **Before**
   ```
   <b>n</b>it.rat</b><i>n</i>Kim</i>/garamasamintratHNO</sub><i>3</i>
   dipakai dl campuranpupuk
   ```

   **After**
   ```
   <b>n</b>it.rat</b><i>n</i>Kim</i>/garamasaminstrat_HNO<sub>3</sub>,
   dipakai dl campuranpupuk
   ```

2. Change rumuskimia “chemical formula” to a semicolon

   **Before**
   ```
   <b>Kam</b>.fa.na</b><i>n</i>Kim</i>kristal...pdshu 158–159°C
   danrumuskimia
   ```

   **After**
   ```
   <b>Kam</b>.fa.na</b><i>n</i>Kim</i>kristal...pdshu 158–159°C,
   C<sub>10</sub>H<sub>18</sub>
   ```

3. Move scientific names to the end of the definitions, preceded by a semicolon

   **Before**
   ```
   <b>T</b>al</b><i>n</i>1</i>tumbuhanpalem</i>Borassusfiabellifer</i>da
   unnya...; batanglontar;
   ```

   **After**
   ```
   <b>T</b>al</b><i>n</i>1</i>tumbuhanpalem,daunnya...;
   batanglontar; Borassusfiabellifer</i>;
   ```

4. Change msl “e.g.” before examples to a colon

   **Before**
   ```
   (dl bentuk kata kerjaber--an); msl</i>bersikutat</i>, berkutat-
   kutatan; i</i>bersipandang</i>, berpandang-pandangan
   ```

   **After**
   ```
   (dl bentuk kata kerjaber--an): <i>bersikutat</i>, berkutat-
   kutatan; i</i>bersipandang</i>, berpandang-pandangan
   ```

---

**Figure 12** A part of the HTML file of KBBI
For each entry in each line in the HTML file,
if <b> is at the beginning of that line,
    extract the lemma between <b> and </b>
    if there is an opening bracket and a closing bracket in the lemma,
or if there is a comma in the lemma,
    extract the variant form(s)
if there is a slash after </b>,
    extract the pronunciation(s) between slashes
if <i> appears after the second slash,
    extract the label(s) for POS, language, domain, etc.
if there is a number surrounded by <b>…</b>,
    split and extract the senses
for each sense,
if <i> appears after </b>,
    extract the label(s)
    extract the definition
if <i> appears after a colon and </i> is at the end,
    extract the example(s)
if <i> appears after a semicolon and </i> is at the end,
    extract the scientific name(s)
if there is some chemical elements,
    extract the chemical formula(s)

Figure 13 A part of the algorithm used to extract a number of fields in an entry

Figure 14 The KBBI data structure
4. The current state of the KBBI database and its applications

After the data was broken down into its components, we can check how many each component is in the database. As of 15 May 2017, the KBBI database contains:

- 48,142 root words (kata dasar)
- 26,198 derived words (kata turunan)
- 30,375 compounds (gabungan kata)
- 2,040 proverbs (peribahasa)
- 267 idioms (kiasan)
- 126,635 definitions (makna)
- 29,255 examples (contoh)

There are many applications can be made possible using the KBBI database. This section will provide some examples of those applications, especially for lexicography and linguistics field.

4.1 Targeted lookups

A user can search for all definitions for a word which may originate from two different headwords, e.g. mereka, using the following search procedure. The results are shown in Table 4.

```
SELECT entri, jenis, induk, lafal, kelas, makna FROM baseview WHERE entri="mereka";
```

The task of looking up phrases and MWEs such as idioms and proverbs is also made simpler, as a user would no longer need to find out which headword to look up first, e.g. the following search procedure can be used to lookup a proverb sedia payung sebelum hujan (the headword is payung ‘umbrella’). Table 5 shows the result.

```
SELECT entri, jenis, makna FROM baseview WHERE entri="sediapayungsebelumhujan";
```

Linguists and etymologists can also search specific entries by their labels. For example, a user can search archaic (ark) lemmas originating from Javanese (Jw) using the following search procedure. Table 6 shows the search results.

```
SELECT entri, ragam, bahasa, makna FROM baseview WHERE ragam="ark" and bahasa="Jw";
```

<table>
<thead>
<tr>
<th>Entri (entry)</th>
<th>Jenis (type)</th>
<th>Induk (root)</th>
<th>Lafal (pronunciation)</th>
<th>Kelas (wordclass)</th>
<th>Makna (definition)</th>
</tr>
</thead>
</table>
4.2 Lexicography analysis

The definitions and examples in KBBI can be regarded as a corpus which can be employed for various analyses and give further insights to the Indonesian language. We extracted the twenty most frequent words in definitions using the Python NLTK library (http://www.nltk.org) (see Table 7). These frequent words can be used as a part of a lexical set for the Indonesian learner’s dictionary we are making now which uses limited words in the definitions and examples. We can also look for the genus words whose result is shown in Table 8. Lim et al. (2016) present the fifty most frequent words and genus words used in definitions in KamusDewan, the authoritative dictionary for Standard Malay. With these data, we can make a comparison of the vocabularies of Indonesian and Standard Malay.
Table 7 Twenty most frequent words in KBBI definitions

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>yang</td>
<td>43,61</td>
<td>untuk</td>
<td>10,312</td>
<td>pada</td>
<td>6,793</td>
<td>dapat</td>
<td>3,020</td>
</tr>
<tr>
<td>dan</td>
<td>26,22</td>
<td>dalam</td>
<td>8,638</td>
<td>orang</td>
<td>6,110</td>
<td>tempat</td>
<td>2,970</td>
</tr>
<tr>
<td>atau</td>
<td>14,41</td>
<td>di</td>
<td>8,537</td>
<td>tentang</td>
<td>4,746</td>
<td>sebagai</td>
<td>2,917</td>
</tr>
<tr>
<td>sebagainya</td>
<td>12,41</td>
<td>tidak</td>
<td>7,756</td>
<td>seperti</td>
<td>3,422</td>
<td>oleh</td>
<td>2,910</td>
</tr>
<tr>
<td>dengan</td>
<td>12,01</td>
<td>dari</td>
<td>7,280</td>
<td>ke</td>
<td>3,247</td>
<td>sesuatu</td>
<td>2,851</td>
</tr>
</tbody>
</table>

Table 8 Twenty most frequent genus words in KBBI definitions

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>orang</td>
<td>2,703</td>
<td>tempat</td>
<td>806</td>
<td>keadaan</td>
<td>526</td>
<td>ilmu</td>
<td>401</td>
</tr>
<tr>
<td>proses</td>
<td>1,858</td>
<td>hasil</td>
<td>656</td>
<td>ikan</td>
<td>521</td>
<td>fobia</td>
<td>350</td>
</tr>
<tr>
<td>alat</td>
<td>1,595</td>
<td>sesuatu</td>
<td>573</td>
<td>hal</td>
<td>512</td>
<td>nama</td>
<td>337</td>
</tr>
<tr>
<td>bagian</td>
<td>835</td>
<td>kata</td>
<td>557</td>
<td>tumbuhan</td>
<td>443</td>
<td>zat</td>
<td>300</td>
</tr>
<tr>
<td>perihal</td>
<td>823</td>
<td>pohon</td>
<td>547</td>
<td>tiruan</td>
<td>413</td>
<td>penyakit</td>
<td>297</td>
</tr>
</tbody>
</table>

4.3 Linguistic analysis

The Indonesian language has a very rich morphology for word derivation process. It has a rich affixation system, including a variety of prefixes, suffixes, circumfixes, non-productive infixes; and a variety of reduplications. Most of the affixes are derivational (Sneddon et al. 2010). Using regular expressions in Python, we made a table of more than 100 patterns of word formation based on affixes and various types of reduplication in Indonesian. Table 9 shows a part of it. It has been used in a linguistics research for analyzing the difference between meN-...-i and meN-...-kan (NurAmirahKhairulAnuar et al. 2017). There are many possibilities we can do with the data, such as analyzing other affixes and reduplications.

Table 9 Some derived words in KBBI, grouped by affixes

<table>
<thead>
<tr>
<th>Affix/Redup.</th>
<th>Example</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>meN-</td>
<td>mengabadi</td>
<td>5,185</td>
<td>21.1%</td>
</tr>
<tr>
<td>meN-...-kan</td>
<td>mengabadikan</td>
<td>2,884</td>
<td>11.7%</td>
</tr>
<tr>
<td>ber-</td>
<td>berabang</td>
<td>2,704</td>
<td>11.0%</td>
</tr>
<tr>
<td>-an</td>
<td>abaian</td>
<td>1,873</td>
<td>7.6%</td>
</tr>
<tr>
<td>peN-...-an</td>
<td>pengabadian</td>
<td>1,780</td>
<td>7.2%</td>
</tr>
<tr>
<td>peN-</td>
<td>pengabai</td>
<td>1,552</td>
<td>6.3%</td>
</tr>
</tbody>
</table>
4.4 Linking to other lexical resources

KBBI contains a number of scientific names for flora and fauna. Using them as a pivot, we aligned more than 600 entries in KBBI to the entries in other lexical resources, such as WordnetBahasa (Bond et al. 2014). Table 10 shows some examples of aligned entries via scientific names.

Table 10 Some examples of aligned KBBI entries and the Wordnetsynsets

<table>
<thead>
<tr>
<th>KBBI entry</th>
<th>Scientific name</th>
<th>Wordnet lemma</th>
<th>Wordnetsynset</th>
</tr>
</thead>
<tbody>
<tr>
<td>abaka</td>
<td>musatextilis</td>
<td>abaca</td>
<td>12353431-n</td>
</tr>
<tr>
<td>abalone</td>
<td>haliotis</td>
<td>Haliotis</td>
<td>01942724-n</td>
</tr>
<tr>
<td>abrikos</td>
<td>prunusarmeniaca</td>
<td>common apricot</td>
<td>12641007-n</td>
</tr>
<tr>
<td>acerang</td>
<td>coleus amboinicus</td>
<td>country borage</td>
<td>12845187-n</td>
</tr>
<tr>
<td>adas</td>
<td>foeniculumvulgar</td>
<td>common fennel</td>
<td>12939282-n</td>
</tr>
<tr>
<td>adasmanis</td>
<td>pimpinellaanium</td>
<td>anise, anise plant</td>
<td>12943049-n</td>
</tr>
</tbody>
</table>

4.5 Online and offline applications

KBBI database serves as the vital part in building the online DWS (https://kbbi.kemdikbud.go.id), called ‘KBBI DalamJaringan’ or KBBI Daring, launched on 28 October 2016 (Kamajaya et al. 2017) and offline mobile applications, both for Android (https://play.google.com/store/apps/details?id=yuku.kbbi5) and iOS (https://itunes.apple.com/us/app/kamus-besar-bahasa-indonesia/id1173573777), launched on 17 November 2016. Figure 15 shows the homepage of the online KBBI and figure 16 shows both the screenshots of the Android and iOS applications. In order to facilitate the workflow of the editorial staff for the online application and the online public participation to add, edit, and deactivate lemmas, definitions, and examples, the KBBI database is equipped with tables for proposals.

![Figure 15 Screenshot of the online KBBI homepage](image)
6. Conclusion and future work

We have described our work in creating a database for KBBI from Microsoft Excel and Word files by converting them to a RTF file and a HTML file, identifying its structure, cleaning up the data, and breaking it down based on the structure. The broken down components were then exported to SQLite database. The database allows lexicographers, linguists, and researchers in NLP field to access the rich lexicographic and linguistic contents in the Indonesian language in more flexible ways, opening up possibilities in discovering new insights into the language, as well as helping the KBBI editorial staff work on the dictionary more effectively.

In the near future, the database will be expanded with etymological information. Our work on compiling and editing the etymological information has been done since 2015 and is still in progress. We have finished working on lemmas from Sanskrit and are working on lemmas originating from Old Javanese and Dutch. In addition, the database will be connected to a corpus. The source of the corpus we are building is from scientific publications. We have finished the first stage and are now adding about five million words per year.

Acknowledgments

Thanks to Francis Bond and Luís Morgado da Costa for the precious advice on the database structure. Thanks to Ivan Lanin for improving the database and making it more efficient. Thanks to Lim Lian Tze who inspired us to write this paper. Thanks to NTU HSS library support staff: Rashidah Ismail, Raihana Abdul Wahid, and Tan Chuan Ko for allowing the first author to borrow the fourth edition of KBBI for months and to Wong Oi May who helped order the dictionary.
References


The Development of Minority Dictionary via Digitalization: A Case Study of Uighur

Chuanming Sun  
National Research Center of Cultural Industries,  
Central China Normal University  
cms@mail.ccnu.edu.cn

Yuting Zhu  
National Research Center of Cultural Industries,  
Central China Normal University  
1448152952@qq.com

Abstract

With the prevalent of Mobile Internet, many challenges and opportunities have been brought to the development of minority dictionary. When people encounter language barrier, they can quickly use dictionary APP to solve these problems, and learn the knowledge of minority culture by using mobile phones. This is one of the important factor that promotes people to use modern digital dictionary to learn. At the same time, it is also an opportunity to promote the transmission of minority language and the local culture. China is a multinational country, and there are some barriers in the development and inheritance of minority languages because of the language and cultural assimilation. With the rapid development of Internet technology, the corpus of the dictionaries and the compilation system are becoming more and more mature, so the application of the dictionary should focus on the innovation of technology and service. This paper aim to reveal the development tendency of minority dictionary through researching and analyzing the development of Uighur digital dictionary. There are three parts in this paper. Firstly, through using new media technology, how to enhance the interaction of dictionaries to attract more users, and how to provide personalized customization of dictionaries to help them learn easily are discussed. Secondly, how to facilitate the joint development of minority dictionary and culture through digitization technology will be answered. Finally, how to exploit the advertising market of minority dictionary by digital innovation, and make the advertising profit become the support of the dictionary innovation is going to be analyzed.

Keywords: minority dictionary, digitization, Uighur

1. Introduction

With the development of society, all kinds of modern civilization and modern
technology have a great influence on minority culture. Thus in the process of intercultural communication, there have been cultural vacancies and cultural conflicts, which affect the cultural transmission of ethnic minorities and the communication between different people. The text of Chinese minority language is a special cultural symbol, which is an important symbol of ethnic minorities and also a bridge between ethnic groups. A dictionary is a reference book that provides phonology, meaning explanation, example sentence, usage, and so on. It not only helps the language learning of the minority region, but also promotes the protection and inheritance of the minority culture.

As one member of the Chinese nation, Uighur has a long history and splendid culture. However, with the development of modernization and the continuous influence of various cultural style, the protection and inheritance of the language and culture that the Uighurs have accumulated over a long period are facing many difficulties. As we all know, rote learning is unable to master a language, and most of the traditional Uighur learning tools are printed dictionary. Therefore, there are few smart tools for learning on the computer and the Internet. According to some education data statistics, there is a positive correlation between the degree of cultural identity and the contact frequency of cultural and recreational media for non-professional people. Thus the text of Chinese minority language can help to enhance the local people's cultural identity. With the “The Belt and Road” policy as well as the expand influence of the new media and digital technology which can be put on Uighur, it has attracted more and more people to learn and use the Uighur language, which has a great role in promoting the development of Uighur dictionary and its culture.

With the rapid development and popularization of smart mobile terminals, Uighur APP has been developed as a new style of Uighur Dictionary. Therefore, based on the background of mobile learning in this research, some suggestions will be proposed for the development of minority dictionary from the following aspects: analyzing the features and demands of the dictionary users, synthesizing the transmission needs of promoting minority language and culture, combining the advantages of the current media integration in resource and platform.

2. The service of Uighur electronic dictionary in PC and mobile terminals

2.1 Digital dictionary based on PC terminal

In September 1987, CANET in Beijing Institute of Computer Application Technology officially built China's first Internet e-mail node. And in September 14th, they issued China's first e-mail: “Across the Great Wall we can reach every corner in the world” which opened the prelude of the use of Chinese Internet. Since then, the network dictionary industry has also been attracted a wide range of attention. The network
dictionary covers a wide range from the initial simple machine-readable dictionary to the current digital dictionary and dictionary APP. This section focuses on the PC terminal of Uighur-Chinese online dictionary.

According to “Statistics Report of the 39th China Internet Development”, the number of Internet users in Xinjiang is about 13 million by the December 2016 and the popularizing rate is 54.9%. According to the data, people engaged in academic research and civilian work are the main consumer groups in the dictionary market. And those people who have the higher level background of culture and education are also the mainstream consumer groups in the dictionary industry. In addition, the Uyghur college students are also the main consumer groups of the Uighur digital dictionary in PC terminal. Internet provides more opportunities for Uyghur cultural dissemination, and in the meanwhile, the network has promoted the development of Uighur dictionary in PC terminal. In recent years, the developmental trend of new media is diversification, and personalization. And the related portal website continues to be revised to strive to close to the life of the users, such as the Xinjiang news website “Tianshan Web” network which utilize the timeliness and interactivity of the new media to integrate videos, audios, pictures and other services into their platform. In the process of cultural communication, digital dictionary can provide rich and detailed explanation to users who do not understand the characteristics of the Uyghur culture conveniently, and then push the relevant language and cultural content and other personalized services to these users.

Different from the printed dictionary, digital dictionary is always not so authoritative, normative, scientific and practical. But some popular Uighur digital dictionaries, like “izda”, “Atila”, are basically based on “Uighur-Chinese Dictionary” as the reference. And these have been able to meet the needs of the vast majority of users, especially some personalized services are warmly welcomed by online translation customers. The digital dictionary of minority is not only to promote their own applications, but also to promote minority culture and the cultural integration between the various national cultures. The compilation and development of dictionaries is the standard and norm of the language cognition, and it is also the record of human development and cultural inheritance. But when Uighur digital dictionary appears, it not only becomes the tool to help people to learn, but also becomes the carrier of cultural information.

2.2 Digital dictionary based on Mobile Internet

2.2.1 Opportunities of digital dictionary based on Mobile Internet

According to the data of “Statistics Report of the 39th China Internet Development”, the scale of China's Internet users has reached 731 million until December 2016, include 695 million people who use the Mobile Internet. And the proportion of Mobile
Internet users increase from 90.1% in 2015 to 95.1%. The development of Mobile Internet makes people more convenient in communication, making friends, entertainment and learning, which makes it possible for people to build up their social circle in various fragments time, thus making life more exciting and creative by Internet.

Mobile Internet is becoming more and more important for cultural communication. And online video has become an important tool for the transmission of minority culture with the popularization of smart mobile terminals. China's online video users have reached to 545 million by the end of December 2016, and online video user usage was 74.5% according to the data from “Statistics Report of the 39th China Internet Development”. Among it, the scale of the mobile video users is close to 500 million. Mobile Internet video usage was 71.9%, which is increasing by 6.5% compared to the end of 2015. Minority culture and language communication on the Internet can increase the usage requirements of minority digital dictionaries, which become an important turning point in the development of the minority digital dictionary in mobile terminals.

2.2.2 The Current Situation and Problem of Uighur dictionary APP

According to statistics, there are about 400 Uighur mobile APP in iOS system and Android system up to March 2017. In these APPs, the amount of Uighur dictionary APPs (not including input method APPs) is nearly 10. There are two main types of Uighur APPs: (1)The traditional translation type, such as Uighur-Chinese dictionary, Uighur-Chinese bilingual dictionary and Uighur dictionary, which can complete the basic functions of Uighur learning, including Uighur translation, pronunciation, words explanation and so on. (2) Search type, such as “Bilkan Dictionary” and “izdax search engine”, which increases the search and translate functions of Uighur website in the mobile phone.

Although the related APPs of minority culture and education continue to emerge, the update frequency of some Uighur APPs, like the Uighur Dictionary APP, is less than the Chinese dictionary, English-Chinese dictionary, and Oxford Dictionary. The less update frequency makes Uighur Dictionary APPs have many problems to solve.

(1) The theme and the content is monotonous

Until now, the learning content of Uighur Dictionary APP is monotonous without any culture pushed or introduced.

(2) Lack of enough information

Although the Uighur dictionary APP can translate the Uyghur website easily, there are only words introduction without pictures, animations, videos and so on. There are few
learning resources and materials on the Uighur dictionary APP, and the update speed is also very slow.

(3) Poor interaction between users

APPs can make the interaction between dictionary and the users, and also between the users and the users, but there are few users who write and share comments and reviews in the interaction function set up by Uighur- Chinese dictionary.

3. The Countermeasures for the development of digital innovation in Uighur dictionary

3.1 The Countermeasures for digital dictionary and new media innovation service

Since the implementation of “The Belt and Road” strategy, Uygur in Xinjiang has become the core area of the Silk Road Economic Zone; the digital innovation service of Uighur dictionary APPs can effectively solve the communication of Uighur people with the outside world, and exploit economy market of “The Belt and Road” and promote the cultural transmission and inheritance. The innovation development of digitalization and new media on Uighur dictionary can be implemented from the following aspects.

(1) Enhance sociability

It is difficult to find like-minded people who like to learn minority languages or who need to learn to do economic trade and other things. Therefore, the social service should be strengthened in the minority dictionary. At present the social function in Uighur APP hasn’t performed its role effectively because of few user who often use interaction function. So, Uighur dictionary should strengthen its social interaction to make users quickly get to know each other through the APPs and communicate with each other on the Internet or in real life. At the same time, users can carry out the corresponding social activities based on the theme of language learning, which enhance the mutual learning experience and enrich the exchange experience between these learners.

(2) Gamification

Traditional dictionary can meet the user's need of translation and interpretation, but the APP can provide learners with a relaxed and pleasant learning environment through playing to learn. Using early childhood education APP for reference, some small games can be added in the Uighur APP. For example, users can fill in the correct words
according to Uygur characteristic pictures; it not only add the interesting of the Uighur language to learn, but also can show and spread the Uygur culture through these pictures.

(3) The integration of digital dictionary and the media

The dictionary is a tool for language learning through providing phonology, meaning interpretation, and example sentences; Bilingual dictionary is not only as a tool for learning language words, but also carries the communication and spread of two national languages and cultural. But learner can hardly understand their culture when dictionary just provide a translation, especially one as Uighur whose living environment, living habits, national culture, religion and so on are different from other nationality. In recent years, with the rapid development of media integration in China, some curriculums of various colleges and universities in the open classes and quality classrooms have a very in-depth introduction of the language and culture. If the user can use the dictionary APP to watch these courses, their learning effect may be greatly improved. For example, the professor live-transmission classroom of Zhejiang University has attracted a lot of attention in these years. Although the professor taught the curriculum “calculus” online, it can be seen that the teaching quality of classroom can bring better learning experience to learners. Therefore, it may be bring better learning experience and learning effect if the Uighur digital dictionary can strengthen the functions of live web casting or online excellent course.

3.2 The Countermeasures for personalized learning in dictionaries

Personalization is very important for dictionaries because each user has their own learning style, cognitive structure and thinking pattern. The best way to use the dictionary APP is to present the content to different users and provide personalized information flexibly. And then, users can get a comprehensive and personalized language information. In a word, digital dictionary can provide users with a wide range of personalized needs.

The procedure of personalization is to collect the individual's interest, and then to generate the personalized information. Different from the other APP which find user’s needs and interests from a variety of advertising, promotions, news and other information, the personalization of minority dictionary APP is mainly reflected in the learning mode and strategy, and the cultural interest and so on. Some countermeasures will be proposed as follows.

(1) Independent choice

Independent choice is reflected in two points: First, selecting words and cultural content to learn. Consumer can freely search and translate text, or select voice
broadcast news through the function provided by the Uighur APP, which is more simple, convenient and time-saving for user. Second, selecting proper time to use APP. At present, most types of APP have a reminder function, for example, “BaiCiZhan APP” can set a fixed time for language learning. It will automatically remind the users every day when the time is coming. As mentioned in the above section, the integration of the dictionary APP and media can also be used for reference. The minority dictionary can be combined with classroom teaching. In addition, remind function can be combined with Internet broadcast and remind users to watch. In terms of Uighur dictionary, when the user chooses to watch the program of Uighur song, dance and costumes and so on, it is also a good opportunity to spread local culture.

(2) Interested information push

On the one hand, personalization can give users a more comfortable experience and lead to the concentration of information. On the other hand, it is difficult for users to find other topics other than interest, which is unfavorable to the study of language and culture. The user of the minority dictionary is a person who is interested in the language and culture of the nation or a student need to learn, so they may have little understanding of minority culture, or only a certain understanding. Take Uighur as an example, most people are only interested in exotic clothing, dances, and delicious food, but do not know the others. So, in the course of the use of the minority dictionary, not only to delivery what the user used to find, but also to push other information which the user may not select but important. Only the comprehensive study of language can promote the communication of language.

(3) Cultural pushing in dictionary application

Some digital dictionaries have expanded the amount of cultural information, strengthened the pushing functions. Especially in recent years, the combination and integration of dictionaries and media can effectively break the barriers between different ethnic groups. For example, when looking for wedding in the Uighur dictionary APP, it can push the Uighur’s wedding video, which not only allows users to understand the related words in the dictionary more easily, but also promotes the transmission of national culture through showing the Uighur culture to the users. At the same time, completed personalization may cause the knowledge of the narrow-minded on the minority language and culture learning. When users search for words and other information, dictionary can push Uygur national culture in an all-round way for a comprehensive collection of words. It is not only to enhance the user's understanding of national culture, but also play a role in the protection and promotion of national culture heritage. Therefore, it can be seen that the minority digital dictionary can promote the exchange and development of culture and economy. At the same time, the progress of science and technology also has stimulated the continuous development of
minority dictionaries.

3.3 The Countermeasures of advertising marketing for dictionary

The development of dictionaries depends on the talents of compilation and technology. So, cultivating excellent talents needs the support of funds. As mentioned above, these functions, for example social, media integration and cultural pushing in the minority dictionaries, can be used to get the benefits from advertisements. The advertisements income can also promote the development of minority dictionary.

Taking Uighur as an example, the Uygur nationality has its special characteristics, including Uygur diet, clothing, dance, music, agriculture and so on, which can be used as a good resource for advertising marketing. When the users use the dictionary to check the Uighur words, the interpretation with text, sound, image, and video can help their learning in the dictionary APP. In addition, the image and video can add advertising as marketing. For example, when look up “grape” in Uighur APP, it can display some pictures or videos which can push some grape products of the local companies to make a profit.

Whether the digital dictionary is in the PC terminal or mobile terminal, it is an effective method to improve the added value to make profit. Minority dictionaries can promote the marketing products by means of their own cultural and creative industry elements. With the implementation of “The Belt and Road” policy, the demand of Uighur language learning is increasing because of the cultural exchange and trade exchange. Therefore, Uygur region, as the economic core area, is feasible to increase the advertisements in the dictionary to get the benefits.

4. Conclusion

Internet is the first real sense media in the history of the world. It makes the study and exchange more conveniently without the boundaries of time and space. The development of information technology has promoted the progress of digital dictionaries which can provide a convenient platform for language learning and cultural communication, and can make people learn and communicate better.

Along with the further implementation of “The Belt and Road” policy, Uygur region in Xijiang is to become the core area. The development of Uighur digital dictionary is very necessary for the economic and cultural communication. If Uighur APP can strengthen some functions include group social interaction, personalized learning, culture and advertising pushing, it can not only help the development of the minority dictionary, but also promote the cultural and economic development of minority areas.
Acknowledgement

This research was supported by Projects of the National Social Science Foundation of China: “Research on the Digital Protection Situation and Countermeasures of Intangible Cultural Heritage in Minority Areas” (15CMZ022).

References


MaJiayue. 2016. What is the customization information for us? - Take the “headlines” APP as an example. Journal of News Research, 2016(06), 131.


Sketching the Difference between Word Senses

Meng-Hsien Shih
National Taiwan University
d00142002@ntu.edu.tw

Abstract

Grammatical collocation is useful information in language learning, and its extraction is important for natural language processing (Pearce 2001). In particular, a collocation resource with sense annotation could be valuable in disambiguating word senses. While most collocation resources focus on the pattern of parts-of-speech (e.g., Oxford Collocations Dictionary 2002), or grammatical collocations (Kilgarriff et al. 2004) for a word form, this paper further considers the collocational behavior of a specific word sense. The project begins with a small Chinese dataset with sense annotation from SemEval-2007 Task 5 (Jin et al. 2007) to extract collocations by a dependency parser (Chang et al. 2009). The parser produces dependency relations between words in terms of triples such as object(eat, fish). The sense-annotated sentences from SemEval-2007 are parsed to obtain dependency relations between words, which are treated as grammatical collocations in this paper. These grammatical collocations with sense annotation can be queried online from the proposed resource.1 This resource can not only sketch the collocation behavior of a specific sense (e.g., “fish” as the object for the sense of “eat”), but also compare the collocation difference between word sense (e.g., “pain” as the object for the other sense “suffer”). The preliminary results are evaluated by applying to the task of word sense disambiguation, with a simple bag-of-words model by counting the frequency of the corresponding dependency collocations for a specific sense. The performance based on only these collocations (55%) over the SemEval-2017 baseline (40%) implies that the collocation behavior of word senses is important in language understanding.

Keywords: corpus lexicography, collocations dictionary, word sense disambiguation

1. Introduction

The emergence of corpora had influenced the research work of lexicographers, linguists and language teachers more than a decade (Hanks 2012; Callies and Paquot 2015). Although corpora with a big data size can further provide more persuasive evidence for the above research fields, meanwhile it takes much more efforts to examine these large-scale linguistic data. Therefore, this paper aims to facilitate the

1 http://homepage.ntu.edu.tw/~d00142002/sense_sketch
work in examining data line-by-line such as concordance.

On the other hand, among the various kinds of information most corpora can offer, collocation information is useful in language learning and lexicography (Kilgarriff et al. 2014; Thomas 2014), and its extraction is also important for NLP (Natural Language Processing) tasks (Pearce 2001). In particular, a collocation resource with sense annotation could be valuable in disambiguating word senses. Therefore, this paper focuses on the collocation behavior of word senses.

2. Literature Review

Most collocation resources focus on the pattern of parts-of-speech (e.g., Oxford Collocations Dictionary 2002), or the identification of grammatical collocations based on these POS patterns (Kilgarriff et al. 2004; 2014) for a word form. Hanks (2004; 2013) uses CPA (Corpus Pattern Analysis) to build the Pattern Dictionary of English for natural language processing and teaching. Chang et al. (2012) associate collocations with WordNet supersenses using hybrid models. Here the author proposes a Chinese open resource that further considers the collocations of a specific word sense.

On the other hand, syntactic information has been useful in extracting collocations (Seretan 2011) and disambiguating word senses (Chaplot et al. 2015). The recent developing universal dependency parsers (Straka et al. 2016; Alberti et al. 2017) further facilitate multi-lingual information processing. A dependency parser produces dependency relations between words in a sentence such as the object relation between the verb *eat* and the noun *fish*. Dependency parses can be also generated from phrase structure parses (Xia and Palmer 2001; de Marneffe et al. 2006), and Chang et al. (2009) followed this approach to devise the Stanford Chinese dependency parser, which will be adopted in this paper.

This study investigates the use of the Stanford dependency parser to extract grammatical collocations from Chinese sense-annotated texts, and releases the results as a web service for users to upload texts to build their own online collocation corpora. The proposed resource also provides these collocations extracted from sense-annotated data, in order to sketch grammatical behavior for a specific sense.

3. Method

This paper examines the approach to build a language resource by automatically extracting syntactic information from sense-annotated texts. The author begins with a small Chinese sense-annotated dataset from SemEval-2007 Task 5: Chinese-English Lexical Sample (Jin et al. 2007) to extract collocations by the Stanford dependency parser (Chang et al. 2009). The SemEval-2007 data contains 2,686 training sentences annotated with senses in English for 40 Chinese lexical-sample words. Every lexical-sample word has two to nine senses annotated in the sentences as shown in the
following format:

<instance id="中医.1">

<answer instance="中医.1" senseid="traditional_Chinese_medical_science"/>

<context>西医出身的她，转而钻研<head>中医</head>理论，试图从前人的经验中吸取有益的东西。</context>

For example, one of the 40 lexical samples, zhong-yi (中医), has 43 sentences annotated with two senses represented in English: Twenty-nine sentences containing the word zhong-yi are annotated with the sense traditional_Chinese_medical_science, and fourteen sentences annotated with practitioner_of_Chinese_medicine.

The 2,686 sentences with words segmented and sense-annotation are first downloaded from the SemEval-2007 Task 5 website. Then the encoding of the XML file is converted from GB2312 to UTF8. These word-segmented sentences with sense annotation are parsed to obtain dependency relations such as dobj(eat, fish), which will be treated as a grammatical collocation in this study.

From the dependency parses for each sense, we can obtain a grammatical collocation list sorted by the number of the occurrence of each dependent word and corresponding relation. For example, the noun fish could occur 12 times as the object of the verb eat.

4. Results

The extracted grammatical collocation with sense annotation can be queried from the proposed resource, as shown in Figure 1:
senseid: eat
  nsubj: 小时 事 新兵 东西 婴儿 组织 奇妙 姑娘 头 城镇 居民 老人
dobj: 鱼 饺子 草 自助餐 游客 起来 吃吃停停 苹果 午饭 水 肉

senseid: live_on
  nsubj: 省里 学校 再也
dobj: 大锅饭 饭 湖

senseid: suffer
  dobj: 亏 苦头 家长 闭门羹 境界
  nsubj: 买家 咱 小平 兄弟 我们 地 悠敞

senseid: wipe_out
  dobj: 车 马

**Figure 1** Grammatical sketch of the four senses (*eat, live on, suffer, and wipe out*) of the word form *chi* (吃)

This online resource can not only sketch the collocation behavior of a word form (e.g., the noun *fish* as the *object* of the verb *eat*), but also compare the collocation difference among word senses (e.g., the noun *pain* as the *object* of another sense of the same word form). The proposed resource also serves as a simple pipeline web framework for all users to easily upload and process their texts, in order to build their own online corpus for others to query, as shown in Figure 2.\(^6\)

---

\(^6\) Currently the open Stanford CoreNLP toolkit under the GNU General Public License is used to process texts such as Chinese word segmentation and dependency parsing.
Figure 2 Proposed online Chinese resource building and processing pipeline framework

The preliminary results were evaluated by applying these sense-annotated dependency parses in the WSD (Word Sense Disambiguation) task, with a simple bag-of-words model by counting the frequencies of the corresponding dependency parses for a sense $s$ in the training data:

$$
\hat{s} = \arg\max_{s \in S} \text{count}(\text{sent_triples}_{\text{testing}} \cap \text{sents_triples}_{\text{training},s})
$$

The performance of this model is shown in Table 1. Although the current extracted
collocations seem not satisfactory, the micro-average precision\(^7\) of the results (516/935=55\%) over the SemEval-2007 baseline (40\%) still implies that the grammatical behavior of word senses may play an important role in language understanding. Furthermore, this simplistic dependency model can also be considered a higher baseline for future Chinese WSD tasks.

**Table 1** The comparison of the micro-average and macro-average precision of the proposed resource and baseline system

<table>
<thead>
<tr>
<th></th>
<th>Micro-average Precision</th>
<th>Macro-average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Resource</td>
<td>55%</td>
<td>60%</td>
</tr>
<tr>
<td>SemEval-2007 Baseline</td>
<td>40%</td>
<td>46%</td>
</tr>
</tbody>
</table>

**Table 2** Training feature distribution of nouns.

<table>
<thead>
<tr>
<th>Target Word</th>
<th>Sense #</th>
<th>Training #</th>
<th>Test #</th>
<th>Baseline</th>
<th>Macro Precision</th>
<th>Avg. Matched Triples #</th>
<th>Training Triples #</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 本</td>
<td>3</td>
<td>68</td>
<td>25</td>
<td>40%</td>
<td>72%</td>
<td>3.2</td>
<td>2,178</td>
</tr>
<tr>
<td>2 表面</td>
<td>2</td>
<td>53</td>
<td>18</td>
<td>61%</td>
<td>61%</td>
<td>3.3</td>
<td>2,012</td>
</tr>
<tr>
<td>3 菜</td>
<td>2</td>
<td>56</td>
<td>19</td>
<td>53%</td>
<td>63%</td>
<td>2.5</td>
<td>1,843</td>
</tr>
<tr>
<td>4 长城</td>
<td>3</td>
<td>48</td>
<td>21</td>
<td>48%</td>
<td>62%</td>
<td>2.0</td>
<td>1,444</td>
</tr>
<tr>
<td>5 单位</td>
<td>2</td>
<td>50</td>
<td>17</td>
<td>59%</td>
<td>65%</td>
<td>1.9</td>
<td>1,807</td>
</tr>
<tr>
<td>6 道</td>
<td>3</td>
<td>53</td>
<td>18</td>
<td>50%</td>
<td>61%</td>
<td>2.8</td>
<td>1,712</td>
</tr>
<tr>
<td>7 队伍</td>
<td>3</td>
<td>64</td>
<td>22</td>
<td>46%</td>
<td>55%</td>
<td>2.5</td>
<td>2,483</td>
</tr>
<tr>
<td>8 儿女</td>
<td>2</td>
<td>60</td>
<td>20</td>
<td>50%</td>
<td>80%</td>
<td>7.2</td>
<td>1,821</td>
</tr>
<tr>
<td>9 机组</td>
<td>2</td>
<td>38</td>
<td>14</td>
<td>71%</td>
<td>86%</td>
<td>5.6</td>
<td>1,007</td>
</tr>
<tr>
<td>10 镜头</td>
<td>2</td>
<td>45</td>
<td>15</td>
<td>53%</td>
<td>60%</td>
<td>3.1</td>
<td>1,595</td>
</tr>
</tbody>
</table>

\(^7\) In SemEval-2007 Task 5, two kinds of precisions are used to evaluate all participating systems: micro-average and macro-average precisions (see Jin et al. 2007 for details).
<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>面</td>
<td>3</td>
<td>67</td>
<td>23</td>
<td>44%</td>
<td>61%</td>
</tr>
<tr>
<td>12</td>
<td>牌子</td>
<td>2</td>
<td>44</td>
<td>17</td>
<td>35%</td>
<td>65%</td>
</tr>
<tr>
<td>13</td>
<td>旗帜</td>
<td>3</td>
<td>50</td>
<td>18</td>
<td>56%</td>
<td>83%</td>
</tr>
<tr>
<td>14</td>
<td>气息</td>
<td>2</td>
<td>39</td>
<td>14</td>
<td>71%</td>
<td>86%</td>
</tr>
<tr>
<td>15</td>
<td>气象</td>
<td>2</td>
<td>47</td>
<td>16</td>
<td>63%</td>
<td>75%</td>
</tr>
<tr>
<td>16</td>
<td>日子</td>
<td>3</td>
<td>88</td>
<td>32</td>
<td>31%</td>
<td>47%</td>
</tr>
<tr>
<td>17</td>
<td>天地</td>
<td>3</td>
<td>65</td>
<td>25</td>
<td>40%</td>
<td>40%</td>
</tr>
<tr>
<td>18</td>
<td>眼光</td>
<td>2</td>
<td>41</td>
<td>14</td>
<td>71%</td>
<td>79%</td>
</tr>
<tr>
<td>19</td>
<td>中医</td>
<td>2</td>
<td>43</td>
<td>16</td>
<td>63%</td>
<td>81%</td>
</tr>
<tr>
<td></td>
<td>2.4</td>
<td>1,019</td>
<td>364</td>
<td>53%</td>
<td>67%</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Table 2 and 3 show the training and test results for the 40 nouns and verbs respectively. Both the proposed resource and the baseline perform better on the noun set (macro-average precision of 67% and 53% in Table 2 respectively). Furthermore, in Figure 3 we can observe a slightly positive correlation between the precision rate and the average number of matched triples per noun instance. In other words, it is suggested that the more dependency triples the model can match from the test sentence to the training sentences, the better it can predict the sense of the target word in the test instance.
The verbs, however, are more difficult to disambiguate (macro-average precision of only 53% and 44% in Table 3 respectively), and there is even no positive correlation between the precision rate and the average number of matched triples per verb instance. In the future, the disambiguation of verb sense shall be further investigated.

Table 3 Training feature distribution of verbs.

<table>
<thead>
<tr>
<th>Target Word</th>
<th>Sense #</th>
<th>Training #</th>
<th>Test #</th>
<th>Baseline</th>
<th>Macro Precision</th>
<th>Avg. Matched Triples #</th>
<th>Training Triples #</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 补</td>
<td>3</td>
<td>63</td>
<td>20</td>
<td>50%</td>
<td>80%</td>
<td>11.3</td>
<td>2,336</td>
</tr>
<tr>
<td>2 成立</td>
<td>4</td>
<td>73</td>
<td>27</td>
<td>37%</td>
<td>48%</td>
<td>10.6</td>
<td>2,237</td>
</tr>
<tr>
<td>3 吃</td>
<td>4</td>
<td>69</td>
<td>23</td>
<td>44%</td>
<td>61%</td>
<td>11.6</td>
<td>2,113</td>
</tr>
<tr>
<td>4 出</td>
<td>8</td>
<td>222</td>
<td>77</td>
<td>13%</td>
<td>39%</td>
<td>7.9</td>
<td>6,439</td>
</tr>
<tr>
<td>5 带</td>
<td>8</td>
<td>197</td>
<td>67</td>
<td>15%</td>
<td>39%</td>
<td>12.7</td>
<td>6,193</td>
</tr>
<tr>
<td>6 动</td>
<td>4</td>
<td>58</td>
<td>20</td>
<td>50%</td>
<td>70%</td>
<td>11.8</td>
<td>1,702</td>
</tr>
<tr>
<td>7 动摇</td>
<td>2</td>
<td>47</td>
<td>16</td>
<td>63%</td>
<td>50%</td>
<td>13.1</td>
<td>1,706</td>
</tr>
<tr>
<td>8</td>
<td>发</td>
<td>5</td>
<td>105</td>
<td>36</td>
<td>28%</td>
<td>50%</td>
<td>11.1</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>-----</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td>9</td>
<td>赶</td>
<td>3</td>
<td>56</td>
<td>18</td>
<td>50%</td>
<td>56%</td>
<td>9.9</td>
</tr>
<tr>
<td>10</td>
<td>叫</td>
<td>4</td>
<td>106</td>
<td>39</td>
<td>26%</td>
<td>39%</td>
<td>12.9</td>
</tr>
<tr>
<td>11</td>
<td>进</td>
<td>5</td>
<td>132</td>
<td>44</td>
<td>23%</td>
<td>55%</td>
<td>7.2</td>
</tr>
<tr>
<td>12</td>
<td>开通</td>
<td>2</td>
<td>56</td>
<td>20</td>
<td>50%</td>
<td>55%</td>
<td>11.4</td>
</tr>
<tr>
<td>13</td>
<td>看</td>
<td>4</td>
<td>103</td>
<td>34</td>
<td>29%</td>
<td>35%</td>
<td>12.9</td>
</tr>
<tr>
<td>14</td>
<td>平息</td>
<td>2</td>
<td>20</td>
<td>8</td>
<td>50%</td>
<td>50%</td>
<td>8.6</td>
</tr>
<tr>
<td>15</td>
<td>使</td>
<td>2</td>
<td>46</td>
<td>16</td>
<td>63%</td>
<td>44%</td>
<td>11.9</td>
</tr>
<tr>
<td>16</td>
<td>说明</td>
<td>2</td>
<td>60</td>
<td>18</td>
<td>56%</td>
<td>67%</td>
<td>11.1</td>
</tr>
<tr>
<td>17</td>
<td>挑</td>
<td>3</td>
<td>40</td>
<td>14</td>
<td>43%</td>
<td>50%</td>
<td>10.5</td>
</tr>
<tr>
<td>18</td>
<td>推翻</td>
<td>2</td>
<td>29</td>
<td>10</td>
<td>60%</td>
<td>50%</td>
<td>15.1</td>
</tr>
<tr>
<td>19</td>
<td>望</td>
<td>2</td>
<td>37</td>
<td>13</td>
<td>77%</td>
<td>77%</td>
<td>5.5</td>
</tr>
<tr>
<td>20</td>
<td>想</td>
<td>4</td>
<td>110</td>
<td>37</td>
<td>27%</td>
<td>32%</td>
<td>15.2</td>
</tr>
<tr>
<td>21</td>
<td>震惊</td>
<td>2</td>
<td>38</td>
<td>14</td>
<td>71%</td>
<td>71%</td>
<td>6.7</td>
</tr>
<tr>
<td></td>
<td>3.6</td>
<td>1,667</td>
<td>571</td>
<td>44%</td>
<td>53%</td>
<td>10.9</td>
<td>48,704</td>
</tr>
</tbody>
</table>
5. Conclusion and Future Work

This paper has shown the importance of the sketch of word senses because several senses of a word form could demonstrate an entirely different collocation behavior. In SemEval-2007 the Chinese-English Lexical Sample task, the best performed SRCB-WSD system (Xing 2007) uses a Maximum Entropy classifier trained with various features including the syntactic feature from a simple rule-based shallow parser, to obtain a precision rate of 71%. In the future, we will also explore the use of state-of-the-art models such as deep learning in order to perform better than the above WSD system.

A fully-automated extraction system of grammatical collocations, if a quality dependency corpus is available, will have various applications, such as exam sentence generation for teaching purpose, and thesaurus generation based on the collocation behavior. In the future, we will increase the size of a quality dependency treebank, investigate and evaluate these related applications.

References


---

8 http://github.com/mhshih/sense_sketch
Computational Linguistics, Pittsburgh.


A Dictionary and Corpora based Study of the Synonyms: “Contain” and “Include”

Congyi Joy Qu
The Chinese University of Hong Kong
qcy200818@163.com

1. Definition of the Research Topic

L2 learners of English often find it difficult to differentiate English synonyms. There are several common and traditional ways to learn synonyms, which may have some flaws: 1) teachers tell students the differences between synonyms according to their experience, which may be inaccurate; 2) students consult dictionaries and thesauruses whose explanations of synonyms may be unclear. Corpus, with authentic language resources complied, offers another way to learn synonyms.

Some previous studies have been done to analyze the similarities and differences between English synonyms, mainly from aspects of distribution, frequency, register, colligation, collocation, prosody, etc. Yu, (2010) used CLEC, FLOB (Freiburg-LOB
Some studies demonstrated how to use corpora to analyze and teach synonyms. Liu and Zhang (2006) demonstrated how to use ICE-GB (British Component of the International Corpus of English) to teach synonyms of “period” “phase” “stage” and “false” “artificial” “mock” “forged” “counterfeit”; Wang and Ren (2009) used LOB (Lancaster-Oslo-Bergen Corpus), BROWN (Brown University Standard Corpus of Present-Day American English Corpus) and BNC to analyze and teach “find” vs. “discover”, “buy” vs. “purchase”, “persuade” vs. “convince”, and “arouse” vs. “provoke”; Chen (2015) used BNC and COBUILD (Collins Birmingham University International Language Database) to compare “selection” vs. “option” and “ill” vs. “sick” and drew some pedagogical implications. Chen and Yuan (2011) and Jafarpour et al (2013) conducted experimental studies to examine the effectiveness of corpus-based approach to teaching the synonyms.

Some other studies have been done to compare native speakers’ and Chinese learners’ uses of English synonyms, with different corpora and research foci. Wang (2016) used LOB and CLEC and found that Chinese learners overuse “begin to infinitive” and underuse “begin” and “start” as intransitive verbs. Ren (2008) used Locness and CLEC (ST5, ST6, essays written by English majors) and discovered that advanced Chinese learners of English tend to mix the meanings and connotations of “common” and “ordinary”.

The writer of the paper is a secondary school English teacher in China. According to the writer’s experience, “contain” and “include” are a typical pair of synonyms which are difficult for Chinese secondary school students to differentiate. Thus, the study is expected to discover similarities and differences between the synonyms, and to find out what kinds of knowledge learners are lacking in their understanding and differentiation of the synonyms.

The paper covers the following contents: 1) an analysis of the synonyms in terms of meanings, uses and collocations based on a native English speakers’ corpus; 2) a comparative study of the uses of the synonyms by native English speakers and by low intermediate Chinese learners of English based on the native English speakers’ corpus and a Chinese learners’ corpus to find out the gaps between the 2 groups.

2. The Corpora

The native English speakers’ corpus is Locness (The Louvain Corpus of Native English Essays), which contains 288,177 words of argumentative and literary essays written by American and British university students and British A-level essays, with sub corpora including alevels 1-9, BRSUR 1-3, USARG and USMIXED.

The learners’ corpus is a sub corpus of CLEC (Chinese Learner English Corpus, Gui & Yang, 2003)—ST2, which contains 208,088 words of essays written by Chinese secondary school students.

Locness and CLEC (ST2) are chosen, for the following reasons: 1) the scope of the research is in the uses of the synonyms in essay writing; 2) a native English speakers’ corpus is needed for analysis of native uses of the synonyms; 3) a low intermediate
Chinese learners’ corpus is needed to be compared with that of native speakers so as to find out possible gaps and further help Chinese learners correct frequently made errors.

3. Methodology

To obtain results, the writer used Antconc to search for all forms of “contain” (contain, contains, containing, contained) and “include” (include, includes, including, included) in the 2 corpora, calculated the normalized frequencies, categorized the meanings, uses and collocations, and compared the findings from the 2 corpora.

4. Results

4.1 Frequencies of “Contain” and “Include” in Locness & ST2

The total number of words in CLEC (ST2) is 208,088. All forms of “contain” occur 35 times in ST2, and all forms of “include” occur 30 times in ST2. The total number of words in Locness is 288,177. All forms of “contain” occur 45 times in Locness (2 sentences with “contained” as an adjective were excluded), and all forms of “include” occur 109 times in Locness (1 sentence with “included” as an adjective was excluded). Sentences with “contained” and “included” as adjectives were excluded, because the study does not focus on the use of adjectives. Table 1 below shows the normalized frequencies (frequency per 100,000 words) of “contain” and “include” in the 2 corpora.

<table>
<thead>
<tr>
<th>corpora</th>
<th>contain</th>
<th>include</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST2</td>
<td>16.8</td>
<td>14.4</td>
</tr>
<tr>
<td>Locness</td>
<td>15.6</td>
<td>37.8</td>
</tr>
</tbody>
</table>

The figures show the following: 1) within groups, low intermediate Chinese learners use “contain” more frequently than “include”, and native speakers use “include” more than twice the frequency of “contain”; 2) between groups, low intermediate Chinese learners use “contain” more frequently than native speakers do, and native speakers use “include” more than twice the frequency of “include” by low intermediate Chinese learners.

The frequencies of “contain” by low intermediate Chinese learners, of “include” by low intermediate Chinese learners, and of “contain” by native speakers are similar, while the frequency of “include” by native speakers is significantly higher—more than twice the frequencies of the other 3. Therefore, low intermediate learners need to learn more to use “include” so as to approach native-likeness.

4.2 Meanings & Distributions of Different Meanings of “Contain” and “Include” in Locness & ST2

4.2.1 “Contain”

Meanings of “Contain”
Based on Collins Cobuild English Dictionary (2000), Cobuild English-Chinese Dictionary (Sinclair & Lu, 2006) was edited with Chinese translations of word meanings as complement. It provides the following meanings of “contain”, and representative sentences with “contain” in the corresponding meanings from Locness are extracted below accordingly.

1) If something such as a box, bag, room, or place contains a particular thing, it has that thing inside it=hold
   Eg. Some bags that contain bread, pasta… (USARG)

2) If something contains something else, it has the other thing among its parts or ingredients=have
   Eg. Marijuana stalks contain fibres. (USARG)

3) If a piece of writing, speech, or painting contains a statement, idea, or image, it has that statement, idea, or image in it=have
   Eg. For television movies say that it may contain violent scenes. (USARG)

4) If an account, idea, system, etc. contains a particular quality or character, it has this quality or character=have
   Eg. Thought about the meaning? It contains a common truth that…(USARG)

5) To contain something means:
   to prevent it from spreading or becoming more serious=control, curb;
   to keep it within a particular area or boundary=hold back;
   to enclose or form the boundary of an area (formal use)
   no example from Locness

6) If you contain a feeling such as excitement or anger, or if you contain yourself, you prevent yourself from showing your feelings=control, retain
   Eg. Find a way to contain human emotion. (USARG)

Distributions of Different Meanings of “Contain” in Locness & ST2

Each sentence with “contain” from Locness and ST2 is allocated to one of the above meanings, and the frequencies and percentages of the meanings are calculated and summarized in Table 2 below.

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Meanings of “contain”</th>
<th>1)</th>
<th>2)</th>
<th>3)</th>
<th>4)</th>
<th>5)</th>
<th>6)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST2</td>
<td>Frequency</td>
<td>0</td>
<td>34</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Percentage</td>
<td>0%</td>
<td>97.1%</td>
<td>2.9%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Locness</td>
<td>Frequency</td>
<td>6</td>
<td>16</td>
<td>16</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Percentage</td>
<td>13.3%</td>
<td>35.6%</td>
<td>35.6%</td>
<td>13.3%</td>
<td>0%</td>
<td>2.2%</td>
<td>100%</td>
</tr>
</tbody>
</table>

As can be seen from the table, “contain” with meaning 2) (have something as its parts or ingredients) is predominant in low intermediate Chinese learners’ writing. Only one
sentence occurs with “contain” with meaning 3) (writing, speech, etc. have an idea or image): “Every song contains a story.” Sentences with “contain” with the other 4 meanings cannot be found in low intermediate Chinese learners’ writing.

In native speakers’ writing, meaning 2) is also the most popular one, but not so predominant as in Chinese learners. Meaning 3) has the same percentage, and other meanings are also used, though with much lower percentages. No sentence can be found to use meaning 5). Locness is a corpus of native university students’ academic and literary writings, and the meanings offered by Cobuild English-Chinese Dictionary (Sinclair & Lu, 2006) are summarized from a corpus which includes all kinds of text types, so meaning 5) may be found in other text types.

The data show that low intermediate learners tend to use only the most common meaning of a word. Therefore, their knowledge of other meanings needs to be expanded.

4.2.2 “Include”

Meanings of “Include”

Cobuild English-Chinese Dictionary (Sinclair & Lu, 2006) provides the following meanings of “incude” and representative sentences with “include” in the corresponding meanings from Locness are extracted below accordingly.

1) Something that includes a smaller thing has it as one of the parts that make up the whole thing=embrace
Eg. Such issues include Leibniz’s theory of optimisim…(BRSUR)

2) If you include something in a whole thing, you make it a part of the whole thing=add
Eg. When an author is attempting to make a strong argument in order to persuade an audience to agree with them, they should include a little more than opinion and one way thinking. (USARG)

Distributions of Different Meanings of “Include” in Locness & ST2

Each sentence with “include” from Locness and ST2 is allocated to one of the above meanings, and the frequencies and percentages of the meanings are calculated and summarized in Table 3 below.

Table 3 Frequencies and Percentages of Meanings of “Include” in Locness & ST2

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Meanings of “include”</th>
<th>1)</th>
<th>2)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST2</td>
<td>Frequency</td>
<td>30</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Percentage</td>
<td>100%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Locness</td>
<td>Frequency</td>
<td>94</td>
<td>15</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td>Percentage</td>
<td>86.2%</td>
<td>13.8%</td>
<td>100%</td>
</tr>
</tbody>
</table>

The table shows that in low intermediate Chinese learners’ writing, “include” in meaning 1) (embrace) is predominant, while no sentence can be found with “include” in meaning 2) (add).
In native speakers’ writing, both meanings are used, but meaning 1) is also much more popular than meaning 2)—more than 6 times the percentage of meaning 2).

The data confirm what has been shown in Table 2—low intermediate learners tend to use only the most common meaning of a word, and measures should be taken to expand their knowledge.

4.3 Differences between “Contain” and “Include”

Section 4.1 presents different frequencies of “contain” and “include” in Locness and ST2 and finds out that learners underuse “include”. Section 4.2 presents meanings of “contain” and “include” provided by a corpus-based dictionary, lists corresponding example sentences from Locness, examines the distributions of different meanings of “contain” and “include” in Locness and ST2, and discovers that learners predominantly use the most common meaning and largely underuse other meanings. This section—section 4.3, is intended to analyze the differences between “contain” and “include” in terms of their meanings, collocations and non-finite verb forms with reference to Locness, to investigate learners’ errors that appear in ST2, and to further draw some pedagogical implications.

4.3.1 Differences in Meanings

Comparing “Contain” and “Include” in Meanings

When the 6 meanings of “contain” and 2 meanings of “include” are compared, differences and similarities can be found. For “contain”, meaning 1) 3) 4) 5) 6) are distinctive, and for “include”, meaning 2) is distinctive.

What makes synonyms difficult to differentiate is their similarities and the subtle differences between the similarities. Meaning 2) of “contain” (if something contains something else, it has the other thing among its parts or ingredients=have) and meaning 1) of “include” (something that includes a smaller thing has it as one of the parts that make up the whole thing=embrace) are quite similar. The 2 explanations and the example sentences from Cobulid English-Chinese Dictionary (Sinclair & Lu, 2006) do not seem to provide clear distinctions. Reviewing the sentences from Locness with these meanings can help reveal the subtle differences.

It can be inferred from sentences from Locness that when describing the relationship of “A has B”, “contain” means that “A” is an entire whole, and “B” is a component inside A; “include” means that “A” is a group of entities, and “B” is one of the entities. For example,

The agricultural industry will contain a substantial amount of cattle… (alevels 7. Locness)
Other contributory factors may include increasing vegetarians & people on a low fat diet…(alevels 7. Locness)

Learner’s Error: Mixing Meanings

The writer read through the sentences with “contain” and “include” from ST2 and found that learners may mix the subtly different meanings of the 2 words—specifically, may apply the meaning of one word to the other. For example,
He wrote a book which included advice…(ST2)

In this sentence, “book” is an entire whole and “advice” is a component within it, so it is preferable that “included” be replaced by “contained”.

**4.3.2 Differences in Collocations**

**Comparing “Contain” and “Include” in Collocations**

**The Collocations of “Contain”**

The writer picked out from Locness the 16 sentences in which “contain” is used in meaning 2) (have), and then used the “Collocates” function of Antconc to search for collocations of “contain” among the 16 sentences. It was found that in 13 out of the 16 sentences, “contain” collocates with words of chemicals or substances: chromosome, BSE (bovine spongiform encephalopathy), offal, Cadmium, lead, steroids, carcinogens, influenza, insulin gene, etc. For example,

Most refrigerators contain Freon…(USARG, Locness)

This use of “contain” is overwhelmingly used by low intermediate learners—in all the 34 sentences with “contain” in meaning 2) (have), “contain” collocates with words of chemicals or substances: vitamins, fat, fibre, flour, fruit, proteins, salt, vegetables, etc. For example,

Butter, cream, chocolate, sweet and so on, which contain a lot of sugar…(ST2)

**The Collocations of “Include”**

Similarly, collocations of “include” were examined among the 94 sentences from Locness and among the 30 sentences from ST2 with “include” in meaning 1) (embrace).

66 out of the 94 sentences from Locness have collocates referring to abstract concepts: measures, factors, aspects, issues, theory, power, institutions, policy, etc. For example,

The relationships created in business include the union between the producers…(USARG)

By contrast, 27 out of the 30 sentences from ST2 have collocates referring to concrete people or things: public holiday, me, courses, teachers, children, etc. For example,

We visited many places of interest, which included the Tian an Men…

The difference in collocations of “include” in Locness and ST2 may be due to that Locness consists of academic and literary writings of American and British University students, while ST2 consists of writings of Chinese secondary school students. The latter group are younger, so they may not have ideas of the sophisticated abstract concepts, or even if they have, they may not have adequate lexical resources to express them.
Learner’s Error: Inappropriate Collocates

In ST2, the following sentences were found:

So the Feng Jia Society was the longest time in Chinese history, which contained Qin, Han…Periods…
The first slavery country…it disappeared in about 770-476 BC, which contained Xia, Sheng, Zhou Periods.
Spring Festival is the Festival of Chinese. It includes three days.

The uses of “contain” and “include” in these sentences are literal translations from Chinese to English. The search results from Locness do not show any example of an era “contains” or “includes” some periods, so the uses of “contain” and “include” here may be inappropriate in terms of collocations. It is also possible that the size of Locness is not big enough to find such examples, and it may be advisable to search a larger corpus such as BNC.

It can be seen that low intermediate learners are not aware of collocations, and teachers can help them raise the awareness by showing and analyzing authentic examples from corpus.

4.3.3 Differences in Non-finite Verb Forms

“Include” is special in the uses of “including” and “included” as non-finite verbs, while “containing” and “contained” do not have such uses.

“Including” and Learners’ Error

“Including”

When the frequency of “include” in Locness was calculated, 36 sentences with “including” used as a preposition were counted in, which accounts for 1/3 of the 109 sentences with “include”.

The preposition “including” has the same meaning as the verb “include”—a group of entities have a certain entity. “Containing” cannot serve as a preposition. This is a significant difference in inflection in the synonyms. For example,

Computer generated pictures, including ‘fractal pictures’…(alevels 6. Locness)
Many disadvantages in store, including the loss of sovereignty…(BRSUR 3, Locness)

Reviewing the sentences, it can be found that “including” can equal to “for example”—to list example(s) among the group of entities, and the 2 sentences above can be understood as:

Computer generated pictures, for example, ‘fractal pictures’…
Many disadvantages in store, for example, the loss of sovereignty…

Learner’s Error: Lack of Knowledge of Grammar

In ST2, there are some sentences where learners should have used “including”, but
actually used “include” or “included”. For example,

Every official include the emperor couldn’t see anything.
People include the old, the young, the men and the women, are interested in taking an active part in sports and games.

These examples show that the learners were not aware of the need for inflectional change and just translated from Chinese to English “include”. More explicit instruction in part of speech and syntax should be provided.

“Included”

According to the writer’s working experiences and observations, English language education in secondary schools in Mainland China attaches much importance to grammar teaching. One grammatical structure which students find difficult to learn is Independent Genitive Structure, or in other words, Absolute Construction, and one typical example of the structure is the use of the non-finite verb “included”:

Everyone has to obey the rules, the leaders included. (Zhu et al, 2015)

Unlike “including”—a preposition which occurs before nouns that serve as objects, “included” occurs after nouns. “Contained” is not used in this way.

The writer wanted to search for such a use of “included” in Locness, but no match was found. This can indicate that, although such a structure does exist in grammar and theory, it may be rarely found in native speakers’ actual writing output.

4.3.4 Pedagogical Implications

In the sections above, some pedagogical implications have already been drawn. In this section, some additional suggestions are provided.

Implications for Teaching Meanings and Collocations

A corpus-based dictionary can provide more comprehensive explanations of words, but it may not provide enough example sentences to illustrate the meanings. Referring to a native speakers’ corpus can help find out much more authentic examples to complement. Besides, to design some exercises can help students consolidate the knowledge of differences between synonyms. Based on the corpus analysis of “contain” and “include”, exercises can be designed as follows:

Please use proper forms of “contain” or “include” to fill in the blanks:
1. The book ___ 64 maps, ___ 3 maps of China.
2. Healthy diet should ___ some nutrition, ___ sugar, protein and calcium.

As to the first sentence, in the first blank, “book” is an entire whole, and maps are components of the whole, so “contains” should be filled in the first blank. For the second blank, “3 maps” are examples of 64 maps, so “including” should be filled in the second blank.

As to the second sentence, in the first blank, “diet” is an entire whole, and the collocate “nutrition” is a chemical substance and a component of the entire whole, so
“contain” should be filled in the first blank. For the second blank, “sugar, protein and calcium” are examples of nutrition, so “including” should be filled in the second blank.

**Implications for Teaching Non-finite Verbs**

Given that very few examples of “included” used as a non-finite verb in Independent Genitive Structure can be found in Locness, teachers may consider putting less emphasis on the structure, since it is rarely used in communicative output. Meanwhile, it does not mean that teaching of the structure should be discarded. “Focus-on-Form” approach (Doughty & Williams, 1998) is necessary to guide students to pay attention to grammatical structures in communicative context in order to acquire metalinguistic knowledge.

**5. Conclusions**

To sum up, the following conclusions can be drawn from the corpus-based study.

Firstly, when it comes to the frequencies of learners’ and native speakers’ uses of the synonyms “contain” and “include”, the frequency of low intermediate Chinese learners’ use of “include” is remarkably lower than that of native speakers, so the learners should receive much more input in using “include”.

Secondly, when it comes to distributions of different meanings of “contain” and “include” in the 2 groups’ writing, the learners tend to predominantly use the most common meanings of the two words—have/embrace, and largely underuse other meanings, so their knowledge of other meanings should be expanded.

Thirdly, the synonyms, though similar, still differ in meanings, collocations and non-finite verb forms—“contain” means an entire whole has an entity, and “include” means a group of entities has a certain entity; “contain” often collocates with words of chemicals or substances, and “include” often with abstract concepts; the preposition “including” is widely used, and “included” can appear in Independent Genitive Structure, but “containing” and “contained” are not used in such ways. Learners are likely to mix the meanings, use inappropriate collocates or improper inflectional forms.

To solve the above problems, it is advisable that, in addition to dictionaries, teachers search native speakers’ corpora for more authentic and idiomatic example sentences, so as to provide more input for the underused word, underused meanings, and analysis of the subtle differences between the synonyms as well.

Locness may be too small for this research, as examples of “contain” in meaning 5) (to prevent something from spreading; to keep something within a boundary; or to prevent one’s feelings), and examples of “included” used as a non-finite verb form in Independent Genitive Structure cannot even be found.

A further question is: some of the example sentences from Locness do not seem very standard in terms of syntax, so should some modifications be made to them before they are presented to the students? For future research, more thorough analysis of learners’ errors in using the synonyms is needed.
References


Locness (The Louvain Corpus of Native English Essays).


The Lexicographical Treatment of Semantic Network Information in Online English Learner’s Dictionaries

Xiangming Zhang
Guangdong University of Finance & Economics
gdufezxm@126.com

Abstract

In the process of language use, various semantic networks are constructed among the semantically-related words. Language learners acquire semantic network information in many ways. A learner’s dictionary is an important resource from which dictionary users obtain the semantic network information of lexical items. In the age of internet, online dictionaries, compared with paper dictionaries, are increasingly more often consulted, and have more capabilities and measures to present semantic network information.

The aim of the paper is to evaluate the methods and measures for presenting semantic network information which are used in online versions of three ELDs: LDOCE6, OALD9 and MED2. In these dictionaries, semantic network information is presented in both the macrostructural and the microstructural levels. A comparative investigation is made to the treatment of semantic network information in the two structural levels of the above three online ELDs. In the macrostructure level, a descriptive analysis is made to the design of the start screen and the arrangement of the overall options to present semantic network information. While in the microstructure level, a case study is made to the noun entry ‘agreement’, analysing the devices employed inside the entry to convey semantic network information.

The results of the comparative study indicate that disagreement occurs among lexicographers over what kinds of semantic network information should be offered in ELDs and what measures should be employed to present such information. A further analysis is made to the current problems and suggested solutions concerning the presentation of semantic network information in online ELDs.

Keywords: Semantic network, learner’s dictionary, online, treatment.

1. Introduction

Semantic network information is a basic type of lexical information for language learners to obtain in the process of language acquisition. The lexical meaning is among the most widely studied core topics in the fields of linguistics and lexicography. In the process of language use, various semantic networks are constructed among the
semantically-related words by means of lexical meanings. On the other hand, a learner’s dictionary is a reference tool to help language learners solve lexical problems, and at the same time serves as an important resource from which dictionary users acquire the semantic network information of lexical items.

Compared with their paper counterparts, online learner’s dictionaries can do much better in treating semantic network information. Online learner’s dictionaries incorporate semantic network information among their relevant words or terms. This kind of information can be accessed by navigating the dictionary’s hyperlinks or by searching directly the keywords. In the search by navigation, the user accesses synonyms and antonyms of a word that are linked from within an entry, and which can lead to other synonyms and antonyms. In the direct search with keywords, the dictionary searches for the query lexical items in the semantic fields, and retrieves a list of entries that include those semantically-related words in their semantic network fields.

2. Literature review

In the present study, the research object is how semantic network information is treated in the online ELDs. Therefore, the literature review is first made to what semantic network is. Then a distinctive analysis is made between semantic network and semantic relations, a pair of confusing terms. Finally a explorative discussion is made to the relations between semantic network and lexicography.

2.1 Semantic network

The concept of Semantic Network was firstly proposed by J . R . Quillian, an American psychologist, in 1968 in his doctoral dissertation concerning semantic memories, where semantic network was regarded as the dominant mental model of associative memory and the memory was realized by the connection between the concepts. In his philosophy, knowledge is stored in human’s brain in the form of network and vocabulary learning is the connection of previous knowledge to new knowledge in the network. Collins & Quillian (1969) further proposed that semantic memory could be described as an intersecting network where concepts are represented as nodes and interconnections among concepts are based on relationships between the nodes. In this view, the individual concept networks form hierarchies with subordinate and super ordinate nodes.

Semantic network is an important type of semantic information, “the associations of related words that come to mind when a certain word is thought of” (Richards & Schmidt, 2002:478). For example, a person may associate ‘large’ with ‘big’, ‘long’ with ‘short’, and ‘house’ with ‘window’. “The lexical semantic representation of a word is not an isolated expression, but is in fact linked to the rest of the lexicon.” (Davis & Gillon, 2004:377) The network is made up of nodes and lines. The node can
express thing, object, concept, behavior, nature, state and so on. The line can express
the semantic relation between two nodes. Semantic Network is also a network of
interconnected elements which can be used to represent knowledge. “The elements are
nodes or concepts, which are connected to one another by virtue of having various
relations with one another.” (Carroll, 2000: 110)

Semantic network is explored and developed not only theoretically but also
empirically. The researchers worldwide have constructed a lot of computer-based
semantic networks, such as WordNet (Miller et al., 1990), EuroWordNet (Vossen,
1998), MindNet (Richardson et al., 1998), and HowNet (Dong et al. 2010), etc. In this
section, the relations between semantic network and semantic relations, between
semantic network and vocabulary acquisition, and between semantic network and
lexicography will be further described and demonstrated.

2.2 Semantic Network and Semantic Relations

Semantic network is realized mainly by the presentation of semantic relations. Lexicon
in the human mind is highly interlinked through various semantic relations to form a
semantic network. “A semantic network is a structure incorporating knowledge about
all possible semantic relations between words.” (Ceglarek, 2013: 26) Many researchers
have realized the important role of semantic relations in forming a semantic network.
“Most semantic networks have been created using semantic relations which are not
fully defined logically to allow unambiguous inference to be performed.” (Cassidy,
network and semantic relations: “The concepts create a semantic network linked by the
relations of hypernymy, meronymy, instantiation, synonymy, antonymy, association
and various other relations describing argument frames. The goal of the semantic
network is to provide description of the links between concepts, that exist in the human
languages and minds, and make it as objective as possible.”

Semantic relations are also important in the comprehension and analysis of the
whole discourse. Morris & Hirst (2004: 47) regard semantic relations as “the building
blocks of lexical cohesion”, and declare that “a clear understanding of their nature and
behavior is crucial”. In the viewpoint of Chiu et al (2007: 33), semantic relations are
“the fundamental building blocks that allow words to be associated with each other and
linked together to form cohesive text”. Furthermore, semantic relations help the users
understand better the words in the lexicon of a language. “A particular lexeme may be
simultaneously in a number of these relations, so that it may be more accurate to think
of the lexicon as a network, rather than a listing of words as in a published dictionary.”
(Saeed, 2008: 63) Polguère (2014: 399) even states the role of semantic relations like
this: “Each lexical unit is no doubt connected to the rest of the lexicon through
semantic relations it entertains with other lexical units.”

The study of semantic relations has a long history in the linguistic world. Semantic relations have been a subject of interest of various disciplines since ancient
times. The 20th century structural semantics has fostered new perspectives on semantic
relations as the basis for lexicon organization. More recently, semantic relations have
become a major theme of interest of Computational Linguistics, as they present a convenient and natural way to organize huge amounts of lexical data in ontology. WordNet and other machine-readable lexical resources (Mititelu et al, 2010: 1). Murphy (2003b:4) states the wide application of the semantic relation research: “Semantic relations among words have captured the interest of various brands of philosophers, cognitive psychologists, linguistics, early childhood and second language educators, computer scientists, literary theorists, cognitive neuroscientists, psychoanalysts – investigators from just about any field whose interests involve words, meaning, or the mind.”

There are at least two problems in the field of semantic relation research. First, semantic relations have not obtained adequate attention from researchers and users. “Despite their importance, lexical semantic relations are severely underrepresented in current linguistic ontology.” (Chiu et al, 2007: 33) Second, which is also more serious one, different classes of researchers have different views on the definite number of semantic relations and on the way to present these semantic relations. “There is no general agreement on the best method to represent the relations between concepts, nor on which is set of relations adequate to represent linguistic and world knowledge.” (Cassidy, 2000: 181-182) When researchers try to demonstrate the semantic relations between lexical items, they will most probably feel frustrated. “There are numerous suggestions about which sets of semantic relations will be useful for knowledge representation, but as soon as one attempts to define words and concepts in a manner that captures the nuances of how people use such words in language, the inadequacy of existing sets of semantic relations quickly becomes obvious.” (ibid: 182) Cruse (1986: 85) even declares the role of semantic relations in the field of lexical semantics like this: “Sense relations of the more specific sort are obviously too numerous and too idiosyncratic to form the basis for a general study of lexical semantics.”

2.3 Semantic Network and Lexicography

Language learners acquire semantic network information in various ways, among which is consulting dictionaries. Lexicographers have employed, or have tried to employ, different methods to present semantic network information in dictionaries, while this work is far away from being satisfactory, because an obvious problem is “that the learners find little information about the semantic network or syntactic behavior of words in bilingual dictionaries” (Thompson, 1987: 282). Lexicographers recognize that semantic network or semantic relation information is an indispensable part for a dictionary, especially a learner’s dictionary. This information is also regarded as an important type of semantic information. “In a broad sense, semantic information can be subdivided into the categories (a) content specification, (b) semantic relations, (c) field or subject classification and (d) equivalence.” (Danlex & Madsen, 1987: 41)

Many lexicographers have tried to present semantic network information in a dictionary. Ahlswede and Evens (1988: 217) propose to deal with semantic relation information in the way of definition: “Our working hypothesis is that a dictionary definition can be analyzed into a set of relations between the defined word on one hand,
and various elements of the definition proper on the other hand. We propose to use definitions to generate word-relation-word triples which may then be used as data to build the lexicon.”

Fontenelle (1997, 2000, 2012) also endeavors continually to explore the relationship between dictionaries and semantic networks, and assets for many times to create semantic network with a bilingual dictionary. Fontenelle (1997) deals with the construction of a large bilingual lexical-semantic database from the machine-readable version of an English-French dictionary, the Collins-Robert dictionary. He shows how a lexical-semantic database compiled from a bilingual dictionary can be used to identify some general frame elements which are relevant to a frame-semantic approach. He also states that an obvious problem in the current dictionaries is that “there is no explicit way of specifying that boar and sow refer to male and female pigs respectively”, so in order to make such distinctions explicit he proposes to “label the 70,000-odd pairs of semantically-related items with lexical relations” (Fontenelle, 2000: 237). With the aim of analyzing the relationship between semantic networks and lexicography, Fontenelle (2012) makes a comprehensive review of the application of WordNet, FrameNet and other semantic networks in the dictionary compilation, and assesses the impact of the various networks, described in International Journal of Lexicography issues over the last 25 years, upon dictionaries and lexicography, indicating that the current epoch is “a time when the word ‘network’ is on everybody’s lips”.

With the development of decades of years, users can seek more semantic network information when consulting dictionaries, especially learner’s dictionaries. Sterkenburg (2003: 412) defines ‘semantic network’ as ‘a coherent group of representational lexemes that share meanings’, and distinguishes ‘semantic relation’ and ‘semantic relationship’: semantic relation is “a relation between the senses of polysemous words that can be made explicit and described in dictionaries by labeling or grouping”, while semantic relationship is “a relationship between lexemes based on certain common features of meaning”. This distinction may be too rigorous for common users. In the broad sense, these two terms can substitute for each other. “Sense relations exist onomasiologically between words, but they also exist semasiologically within a word.” (Geeraerts, 2003: 85)

It is widely accepted among western lexicographers that semantic network information should be presented in the dictionary in an appropriate way. Vaquero et al (2006: 168) explains clearly the relationship between a dictionary, semantic relation and semantic network: “One of the important features that is present in any dictionary is the presence of semantic relations linking lexical items, and thus forming a semantic network.” Semantic network is more important for pedagogical or learner’s dictionaries, because they are expected more probably to offer users semantic network information. “Semantic relations are relevant for pedagogical lexicography from a variety of perspectives. An extensive presentation of semantic relations within a word entry is important in productive L2-tasks, when a language learner has to choose the semantically correct word out of many possibilities.”(Kremer & Abel, 2010: 380) As an academic monograph on systematic lexicography, Apresjan (2008) spends half of
the attention and energy on the problems of synonymy, the most basic and important semantic relation. Ogilvie (2011: 401-402) describes how semantic network is constructed in a dictionary: “By creating a semantic network view, the lexicographer presents the user with a network in which words in the dictionary that are semantically related are connected together by colored lines – each color represents a different relationship e.g. same meaning or alternate forms.”

The domestic researchers, although a bit later than the western peers, also show their interest on the relevant study of the application of semantic network in the field of lexicography. Zhao Yanchun (2003), as an academic monograph on cognitive lexicography, analyzes semantic network in a cognitive perspective, asserting that semantic network is part of lexical competence that language learners should acquire, and exploring semantic network in the aspects of semantic trains and semantic fields (ibid: 167-172). Yong Heming (2003), as an academic monograph on communicative lexicography, analyzes semantic network in a communicative perspective, and states that a dictionary should follow “relevance principle”, indicating that the various parts of a dictionary should be connected to each other (ibid: 128-130). Yong & Peng (2007: 44) further explores the importance of semantic relation treatment in a monolingual dictionary and also in an active dictionary: “The stem from which the lexical item is derived, the grammatical class to which it belongs, and the semantic relation it has to the stem can all affect its positioning in the monolingual dictionary structure…. But for the active type, the morphological form and the semantic relation, which may easily excite morphological and semantic associations with the stem to which it is related, should take precedence over the grammatical classification.”

Professor Zhang Yihua (Zhang Yihua & Yong Heming, 2007; Zhang Yihua, 2009; and Zhang & Guo, 2010) claims the lexicographical treatment of semantic network information in various situations. On the one hand, research is made to the importance and necessity of presenting semantic network information in a dictionary. Compared with the other key components of a dictionary, semantic network has not been given equal attention. “In conventional dictionaries, macrostructures and microstructures are given a lot of importance by lexicographers, but there are few studies on the lexical and semantic network.” (Zhang & Guo, 2010: 181) It becomes more explicit among the lexicographers why semantic network is important for a dictionary. “The systematic semantic relations in a semantic field are formed in the process of natural language communication. They generally connect together those lexical items sharing the same semanteme. However, during the process of dictionary compilation, the natural semantic field is torn apart by the lexicographical arrangement of entries. It is necessary to create efficient defining models and integral cross-reference systems, to reconnect the semantically-related lexical items, and to construct organic semantic network systems, so that users can consult the dictionary in a systematic way.” (Zhang Yihua & Yong Heming, 2007: 272) Comparing with the general dictionaries, active dictionaries need to treat more semantic network information. “How to connect semantic network and dictionary structure is a main task for active dictionaries.” (Zhang Yihua, 2009: 265)

On the other hand, effort is made to the various ways of presenting semantic
network information in a dictionary. Zhang Yihua (2009: 265) proposes a multidimensional semantic network pattern: “Based on mediostructure, a multidimensional semantic network is constructed in a dictionary, so as to offer a larger background for language learning, and put any lexical item into the systematic relations of a language. The semantic network of a dictionary connects not only the microstructure, but also macrostructure and outside structure. This connection is constructed on the basis of various systematic relations of language cognition, and it reflects the psychological process of human thinking and language cognition.” Zhang & Guo (2010: 182) further explains the ways to realize multidimensional semantic network pattern: “How to represent those relations in dictionaries is worthy of further study. For the time being, the construction of the lexical and semantic network in printed SLDs can be based on definition, illustrative examples, cross-reference, annotation, indication and related notes or special columns, which may establish necessary links among relevant information at different places.”

The presentation of semantic network information is more important in learner’s dictionaries than in general dictionaries, which also gradually becomes consensus among lexicographers and users. “It is a direction for the future lexicographical study and dictionary compilation to consider how to present lexical relations through semantic network, and reflect the individual meaning and usage through lexical relations, so as to make the definition more profound, more comprehensive, and more accurate.” (Zhou Pin, 2005) Feng Xuefang (2014) compares the English and Chinese word association patterns of Chinese EFL learners with those of English native speakers, and analyzes six types of semantic relations: synonyms, antonyms, hyponyms, collocations, semantic neighbors, and semantic remote relatives.

3. Research questions

In order to analyze in detail the presentation of semantic network information in online English learner’s dictionaries (hereinafter referred to as online ELDs), the present research tries to find out the possible answers to the following questions:

1) What are the similarities and differences between the presentation of semantic network information in the mainstream online ELDs?
2) What problems occur in the treatment of semantic network information in the online ELDs?
3) What measures can be taken to solve the problems concerning the presentation of semantic network information in the online ELDs?

4. Research methodology

In order to explore deeply the above-mentioned two questions, the present research is developed mainly in the methods of case study and contrastive research.

A case study is developed to the treatment of semantic network information in the
three online ELDs: OALD9 online, LDOCE6 online and MED2 online. The three dictionaries are chosen for investigation and analysis not only for their wide reputations among users, which indicates the representativeness of the top level of ELDs, but also for their latest versions, which guarantees the timeliness of the present study.

A contrastive analysis is made between the lexicographical treatment of semantic network information in the three online ELDs, evaluating their merits and demerits, exposing the existing problems and finding out corresponding solutions. The semantic network information is presented with various devices in the perspectives of macrostructure and microstructure.

5. Results and analysis

In terms of structure, a paper dictionary can be roughly classified into three parts: macrostructure and microstructure. Macrostructure refers to the component part outside an entry, while microstructure is the component part within an entry. This classification can also apply to an online dictionary.

In the present study, a detailed exploration is made to the semantic network treatment in OALD9 online, LDOCE6 online and MED2 online. For each dictionary, the exploration is carried out in the two perspectives: macrostructure and microstructure. In the macrostructure perspective, a bird-view description is made to the common measures employed to present the semantic network information of the lexical items. In the microstructure perspective, a case study is made to the semantic network treatment of the noun ‘agreement’ entry in the dictionary.

5.1 OALD9 Online

5.1.1 Macrostructural considerations

In the macrostructural perspective, OALD9 online employs the following measures to make semantically-related lexical items list together, thus constructing semantic networks.

1) Topics

On the top left of the screen interface, there are 6 item guideposts, among which is ‘Topics’, as is illustrated in Figure 1. The clicking of the guidepost will lead to ‘Topic dictionaries’, where you can see 24 mini-dictionaries, i.e. groups of words related to common subject areas. This part is very useful to build users’ topic vocabulary. These topics are:

conflict, Work

Take ‘Animals’ as an example. Under this name there are 2 sub-headings: Features (Animal homes, Groups of animals, and Parts of animals) and Types (Amphibians and reptiles, Birds, Domesticated mammals, Fish, Invertebrates, and Wild mammals). Each heading or sub-heading can be classified for several levels, till at last the user will be led to a bottom lexical item as an entry in a paper dictionary.

2) Vocabulary expanding

Also on the top left, there is another guidepost, ‘FAQ’ (Frequently Asked Questions), as is also illustrated in Figure 1. There are detailed answers to 12 questions, among which is ‘How can Oxford Learner's Dictionaries help me expand my vocabulary?’ Within the dictionary, vocabulary expansion can be realized in the following 3 aspects. Firstly, in the ‘Wordlists’ section, users can make full use of the following 4 tools: Oxford 3000, the Academic Word List, entries with pictures, and entries with usage notes. Secondly, the ‘topic dictionary section’, as is mentioned above. Thirdly, with the tool ‘My wordlist’, users, especially the premium users, can create their own topic lists and test themselves on the meanings of the words in each list.

![Figure 1](an_extract_of_the_home_page_interface_in_oald9.png)

**Figure 1** An Extract of the Home Page Interface in OALD9

3) Other results

For each headword, there are 3 guideposts on the right side of the screen interface. They are ‘Other results’, ‘Nearby words’ and ‘Explore our topics’, as are illustrated in Figure 2. In the section ‘Other results’, users can find a list of expressions, including phrases and idioms, with the headword as a constituting element. For example, if the word ‘ask’ is typed as a headword, then in the section ‘Other results’, users can find the following 3 groups of results: a) ‘All matches’, all the commonly-used expressions of the headword ‘ask’, i.e. ‘ask’, ‘ask back’, ‘ask for’, ‘ask after’, ‘ask around’, ‘ask for it’, ‘don’t ask’, ‘I ask you’, ‘don’t ask me’, ‘if you ask me’, and ‘ask, cry, etc. for the moon’. b) ‘Phrasal verbs’, a list of the phrasal verbs of ‘ask’ chosen from the above ‘All matches’, i.e. ‘ask back’, ‘ask for’, ‘ask out’, ‘ask after’, and ‘ask around’. c)
‘Idioms, a list of the idioms of ‘ask’ also chosen from the above ‘All matches’, i.e. ‘ask for it’, ‘don’t ask’, ‘I ask you’, ‘don’t ask me’, ‘if you ask me’, and ‘ask, cry, etc. for the moon’.

4) Nearby words

Also on the right of the interface, there is another guidepost ‘Nearby words’, where users can find a list of lexical items which share the same root with the headword or have the similar spelling. For example, the ‘Nearby words’ for ‘agreement’ are: agreeable, agreeably, agreement, agri-, and agribusiness.

For the third guidepost ‘Explore our topics’, this part is the detailed arrangement of the 24 topics mentioned above. It is more impressive and convenient for the users. What is worth mentioning is that each item can be led to sub-types or the whole entry article of the headword, so that users can achieve better comprehension for the relevant lexical items.

Figure 2 Another Extract of the Home Page Interface in OALD9

5.1.2 Microstructural considerations

In the microstructural perspective, a detailed description is to be made to the following various measures to present semantic network information inside a specific entry of the noun ‘agreement’.
1) Wordfinder

The column ‘Wordfinder’ is located at the end of a sense section, to list a number of words which are semantically related to the headword. As is illustrated in Figure 3-1, the ‘Wordfinder’ column for the first sense collects the following words which are semantically related to ‘agreement’: agreement, armistice, ceasefire, disengage, negotiate, peace, reparations, surrender, treaty, and truce.

2) Collocation patterns

Collocation is a common way to show syntagmatic relations. Collocation patterns are located between before the column ‘Wordfinder’, together with a relevant example for each pattern, as is also illustrated in Figure 3-1.

3) See also and see related entries

The two columns ‘see also’ and ‘see related entries’ are located at the end of the sense section, as is illustrated in Figure 3-1. The former is used to collect several expressions consisting of the headword, e.g. ‘gentleman’s agreement’ and ‘prenuptial agreement’ for ‘agreement’; while the latter is the collection of other entry words semantically-related to the headword, e.g. ‘peacekeeping’ for ‘agreement’.

Figure 3-1 An Extract of the Search Result of ‘agreement’ in OALD9
4) Synonym and opposite

The two labels ‘synonym’ and ‘opposite’ are also located at the end of the relevant sense section, as is illustrated in Figure 3-2. As the terms suggest, the former leads to the synonym of the headword, as ‘concord’ for ‘agreement’, while the latter leads to the antonym, as ‘disagreement’ for ‘agreement’.

5.2 LDOCE6 Online

5.2.1 Macrostructural considerations

In the macrostructural perspective, LDOCE6 online employs the following measures to present semantic network information:

1) Cross reference in ‘How to use’

On the home page of the dictionary interface, as is illustrated in Figure 4, there are 3 guideposts, among which is ‘How to use’. The click of the button will lead to step-by-step procedure in consulting the dictionary, together with 9 FAQs and practical answers. One of the questions is ‘How do I look up a cross-reference in an entry?’ Cross-reference is a useful way in present semantic network information. In the dictionary, cross-references are displayed in blue, and linked in the same way as any other word in an entry - just double-click on it and the entry will appear.
2) Juxtaposition of some semantically-related words with the headword

When a word is typed for searching, a list of semantically-related words with the headword will be juxtaposed and appear in the interface, as is illustrated in Figure 5-1. The clicking of each word will be linked to its own entry. Thus users can have a general overview of all the semantically-related words.

3) ‘Explore our topic dictionary’


Secondly, in LDOCE6, the topic is labeled, if there is any, after the corresponding headword, so that users can read further the topic section when consulting a headword, such as the topic label ‘LAW’ for the headword ‘agreement’, as is illustrated in Figure 5-2.
5.2.2 Microstructural considerations

In the microstructural perspective, LDOCE6 has done less work in presenting semantic network information, compared with OALD9. As is illustrated in Figure 6, only the collocation patterns are provided in the relevant sense sections. There is no information about synonymy or antonymy, though such information is provided in the paper version of LDOCE6. What is worth mentioning is that in the whole entry article, there is no hyperlink or cross-reference information, thus failing to construct semantic relations with other words, which makes the dictionary as simple as a paper version, or even worse.
5.3 MED2 Online

5.3.1 Macrostructural considerations

In the macrostructural perspective, MED2 employs the following measures to provide semantic network information:

1) Thesaurus usage in FAQ

On the top side of the home page interface, there are 7 menu buttons, as is illustrated in Figure 7, and the clicking of the last button ‘More’ will lead to a section ‘FAQ’. Among the 18 question is one about thesaurus usage: “How do I use the Thesaurus?” Next to the definition, there is generally a thesaurus symbol ‘T’ indicating thesaurus. The clicking of the symbol will lead to the free thesaurus, and give users the most frequent and useful synonyms or related words for the exact meaning of the headword.

2) Related words

On the right side of the entry article interface, there is a button ‘Other entries for this word’, the clicking of which will lead to a column ‘Related words’, as is illustrated in Figure 8. Inside the column, a dropdown list appears with a number of words or expressions which are semantically-related to the headword, such as trade agreement, license agreement, be in agreement, gentleman’s agreement, etc. Each word or expression has a hyperlink to its whole entry, thus constructing a semantic network.
Figure 7 An Extract of the Home Page Interface in MED2

Figure 8 An Extract of the Search Result Page of ‘agreement’ in MED2
3) Collocates

Just below the above-mentioned ‘Related words’ column is another column ‘Collocates’, the clicking of which will lead to the collocation information, as is also illustrated in Figure 8. For instance, in the ‘Collocates’ of ‘agreement’ are the verbs frequently used with ‘agreement’ as the object: conclude, forge, negotiate, ratify, reach, secure, and sign.

5.3.2 Microstructural considerations

In the microstructural perspective, the following measures are employed to offer semantic network information in MED2:

1) Explore thesaurus

At the end of a sense section is located a button ‘Explore Thesaurus’, as is illustrated in Figure 9, where, when the button is clicked, a group of synonyms and related words are grouped and contrasted with their respective definitions. For instance, following the second sense of ‘agreement’ are the following relevant synonyms and related words: Chatham House Rule, agreement, contract, deal, alliance, bargain, accord, common ground, unanimity, and understanding. Moreover, each word or expression is hyperlinked to its own entry article. What is worth mentioning is that this types of information is led by a red capital letter ‘T’, short for ‘Thesaurus’, and the guideword ‘Agreement and agreements’, as is also illustrated in Figure 9. Such devices make the information more noticeable and topic-prominent.
2) Collocation patterns

Collocation patterns are located after the definition in the sense section, to show how
the headword uses in a collocation structure, as is also illustrated in Figure 9. For
instance, 3 collocation patterns are provided for the second sense. They are:
‘agreement about/on’, ‘reach agreement’, and ‘there is general/wide agreement that
(=most people agree that)’. For each pattern there is an example sentence to show
dictionary users how the collocation pattern is used.

3) Collocates

At the end of a sense section, there is sometimes a box ‘Collocates’, to list some words
of a certain POS (part of speech) which are generally used with the headword as
collocation, as is illustrated in Figure 10. This box is the detailed explanation of the
‘collocates’ column in the macrostructure, as is mentioned above.

![Figure 9 An Extract of the Entry Article Page of ‘agreement’ in MED2](image)

![Figure 10 ‘Collocates’ in the Entry Article Page of ‘agreement’ in MED2](image)

6. Discussion and implications

The above explorative description indicates that various measures and devices have
been employed in the mainstream online ELDs to present semantic network
information. Nevertheless, many problems have been found by the fastidious users in the perspective of presenting lexical units’ semantic network information in these dictionaries. Correspondingly, some suggestions are to be made to solve these problems.

6.1 Current problems

Until currently, the online ELDs have not yet fulfilled their potential functions in presenting lexical items’ semantic network information although they have possessed large technological capabilities. Many researchers have found out some problems occurring in the current paper-based EDs in presenting semantic relation information. “In the electronic versions of paper-based dictionaries semantic relations are implicit and are difficult to detect by a language user, because the information contained in these dictionaries is garbled in some way. This makes these resources near to useless for vocabulary learning, mainly because the rich associative semantic network that a dictionary can contain remains inaccessible to them, and because that same network is so flawed that even a state-of-the-art NLP application could never straighten it up.” (Vaquero et al, 2006: 168)

What’s more, agreement has not yet been reached among linguists and lexicographers, as well as language users, on what semantic network information need to be presented in the dictionaries. For instance, the antonym information ‘disagree’ of the entry word ‘agree’ is mentioned in OALD9 online and LDOCE6 online, but not in MED2 online.

Moreover, consensus has not yet been made among dictionary compilers on what measures should be employed to offer semantic network information in online ELDs. Various labels and devices are employed in different learner’s dictionaries to convey semantic network information, which to a large extent makes users confused and frustrated.

Comprehensively speaking, the online ELDs are still in their early stage, and many potential functions have not realized and put into practice. Homogeneity is a big problem for the current mainstream online ELDs. “The last few years has seen free versions of British monolingual dictionaries for advanced learners appear online, one by one. On the whole, the major British MLD’s have followed a pattern of remarkable similarity, perhaps as part of the competitive drive, and this is also reflected in the features offered in their online versions. There is also a more down-to-earth reason for the similarities found in a number of British MLD’s: they tend to use the same software dictionary production platform from IDM.” (Lew 2011: 234)

6.2 Suggested solutions

Sufficiency is the first requirement which must be reached by the online ELDs in presenting semantic network information. In other words, dictionary compilers should come to an agreement on what semantic network information need to be presented in the dictionaries. The following types of information should be included as semantic
network information.

1) Semantic relation information

The lexical information of semantic relations is the most basic type of semantic network information. The semantic relation information basically the following 9 types: synonymy, antonymy, hyponymy, hypernymy, co-hyponymy, meronymy, homonymy, co-meronymy, and troponymy.

2) Semantically-related lexical items

The semantically-related lexical items are the words or expressions which share the same semantic fields. They are of great help in broadening users’ vision and enlarging their vocabulary. These lexical items are considered to have paradigmatic relations with the entry words.

3) Collocates

The collocates are the collocation patterns of the certain entry words. A good use of the collocates for a headword will make the language idiomatic and natural. They are essentially important for ESL learners. These collocates are considered to have syntagmatic relations with the headwords.

Efficiency is another requirement which must be reached by the online ELDs in treating semantic network information. In other words, dictionary compilers should reach consensus on what measures and devices need to be employed to present semantic network information in the dictionaries. Various structural components in the levels of macrostructure and microstructure can be used to treat semantic network information.

In online ELDs, semantic network information is generally presented in the following three perspectives:

A) Outside matter

In online dictionaries, outside matter is represented in the form of menu buttons on the top of the dictionary interface page. Generally there is a part ‘FAQ’ (frequently asked questions), inside which is a section ‘How can the dictionary help me expand my vocabulary’, to guide the user how to find a lexical item’s semantic network information in different ways. This part has the similar function with the front matter of a paper dictionary.

B) Within an entry

Within an entry article of an online dictionary, the following devices can be employed to present semantic network information:

a) Synonym and opposite
The two labels ‘synonym’ and ‘opposite’ can be located at the end of the relevant sense section, to build connections by hyperlinks with the synonyms and antonyms. The user can also be led to a synonym discrimination column. For instance, in the entry of ‘ask’, a column is used to distinguish ‘demand, require, expect, insist, and ask’.

b) See also and see related entries

The two columns ‘see also’ and ‘see related entries’ can be located at the end of the sense section, to indicate the semantically-related words. For example, in the entry of ‘agreement’, ‘see also’ is used to link ‘gentleman’s agreement’ and ‘prenuptial agreement’, while ‘see related entries’ is employed to refer to ‘peacekeeping’.

c) Idioms

At the end of the entry, there can be a column ‘Idioms’ to collect all the idioms of the headword, together with the explanations and examples. For instance, in the entry ‘ask’ there is a list of idioms: ask for it, be asking for trouble, be asking for it, don’t ask, don’t ask me, for the asking, I ask you, and if you ask me.

d) Phrases or collocations

At the end of the entry, there can also be a column ‘Phrases’ or ‘Collocations’, where nearly all the phrases or collocations of the headword are collected, and the clicking of the phrase or the collocation pattern can be linked by hyperlink to its own explanation and its illustrative example. For instance, in the entry ‘ask’ there is a list of phrasal verbs: ask after somebody, ask around, ask somebody back, ask for somebody, and ask somebody out.

C) Outside an entry

Outside an entry of an online dictionary, generally on the right side of the search page, the following devices can be employed to convey semantic network information:

a) Related words

A list of words or phrases which are semantically-related to the headword can be collected, so that users can search for detailed information. For example, for the headword ‘ask’, the following words or phrases can be collected as related word: ask around, ask for, ask out, ask after, don’t ask, don’t ask me, I ask you, if you ask me, etc.

b) Nearby words

In the column ‘Nearby words’, a list of words which are alphabetically arranged with the headword will be collected together, so that users can recognize more words which are similar in spelling with the headword. For instance, for the entry ‘agreement’, there is a list of nearby words: agreeable, agreeably, agribusiness, etc.
c) Topic information

Generally, an online dictionary can provide topic information as a special column. This information is located on the right side of the search page, so that users can search all the words about a certain topic. For example, under the topic ‘animals’ the user can find our scores of words in the sub-category level. Under each sub-category word is the further classification information, so that users can understand a large number of words which are semantically-related under a certain topic.

7. Conclusion

Lexical items’ semantic network information is a very essential part in a learner’s dictionary. Meanwhile, a learner’s dictionary is an important resource from which dictionary users acquire the semantic network information of lexical items. Due to the wide capacities of the internet technology, the online ELDs can do much better in presenting semantic network information. However, until present such work is far from satisfactory. The online ELDs have not done much better than their paper version. “It is surprising to see so many of the online dictionaries (including some from respectable publishers) still largely constrained by the paper model, with access mechanisms to lexicographic data often being substandard for today’s technology.” (Lew 2011: 247)

The insufficiency of the types of semantic network information offered in the dictionary and the inefficiency of the methods to present semantic network information are the two essential problems for the current online ELDs. In the present paper, a deep analysis is made to the classification of the semantic network information, aiming to solve the problem of insufficiency. Then an explorative description is made to the presentation of semantic network information in online ELDs, aiming to answer the problem of inefficiency.

References

Dictionaries


Other literature


Gruyter GmbH & Co. KG.
A Sociocognitive Approach to English Online Dictionaries

Ying Lv
Southwest University of Science and Technology
1092210169@qq.com

Congmei Chen
Southwest University of Science and Technology
632218086@qq.com

Abstract

In the 1990s, online dictionaries appeared as machine readable dictionaries began to take Internet as their medium due to the rapid development of computer and Internet technology. By virtue of huge advantages of the Internet, online dictionaries have demonstrated such merits as convenient retrieval, on-demand updatability, etc., in comparison to traditional paper dictionaries since inception and become increasingly popular among a large number of users. Observing current studies, the writer finds that both social and cognitive feature of dictionaries have been illustrated respectively, but few attempts have been made to approach online dictionaries from the integration of the two perspectives, i.e. the sociocognitive perspective. Since the social and the cognitive are not diametrically opposed to each other, it is both feasible and necessary to study online dictionaries from the sociocognitive perspective. Therefore, this paper attempts to study English online dictionaries from this perspective in terms of compilers and users. Based upon some propositions of sociocognition and the cognitive theory of multimedia learning, it first investigates how the social and the cognitive interact with each other in promoting the emergence and development of online dictionaries. Then some common characteristics such as dynamic repository of knowledge, hyperlinks, customization, etc., are presented, together with detailed analyses of the interplay of both social and cognitive factors. For example, online dictionaries have the capacity to quickly include the neologism, thus timely mirroring the social cultural phenomenon at a particular time, meanwhile, a comprehensive understanding of users’ cognitive habits, cognitive rules and cognitive needs is also required to better satisfy users’ multiple needs. It concludes with limitations of this paper and some constructive suggestions for future studies.

Key words: Sociocognitive perspective, English online dictionary, cognitive need, cognitive theory of multimedia learning.

1. Introduction

Computers are playing an increasingly common and important role in the “preparation, printing, and consultation of dictionaries” (cited in Bejoint 2001: 89) and the Internet “seems to have consolidated itself as the dominant platform for dictionaries” (Tarp
The emergence of online dictionaries (also called “Internet dictionaries”) is just consistent with this development trend. An online dictionary refers to “a dictionary or other reference work available via a computer network, such as the INTERNET” (Hartmann and James 2000: 100). By virtue of huge advantages of the Internet, online dictionaries have demonstrated such merits as convenient retrieval, on-demand updatability, etc., in comparison to paper dictionaries since inception and become increasingly popular among a large number of users. Though it is no more than thirty years since the birth of online dictionaries, still they have achieved prosperous development and vigorous popularity. It is estimated that “the number of people online will approach 3 billion by 2016 and is projected to reach 5 billion by 2020” (Thaker 2015: 73). With the growing popularity of the Internet, there might be more people accustomed to employing online dictionaries in the future. Therefore, it is both necessary and meaningful to study online dictionaries from various perspectives.

English online dictionaries have become a necessity for English learners to acquire vocabulary. Driven by the advanced computer and information technology in the modern society, online dictionaries appear and have the capacity to quickly collect the neologism, thus timely mirroring the social cultural phenomenon at a particular time. In this respect, they are social. Meanwhile, online dictionaries insist on user-orientation and go a step further by enabling the general users to participate in the compilation. On the other hand, the compilation of online dictionaries is built on the basis of a comprehensive understanding of users’ cognitive habits, cognitive rules and cognitive needs, thus mirroring their cognitive characteristics. Therefore it can be concluded that online dictionaries are both social and cognitive.

2. Research on Online Dictionaries at Home and Abroad

Before reviewing previous studies on online dictionaries at home and abroad, we will first discuss the relation between electronic dictionaries and online dictionaries.

2. 1 Electronic dictionary and online dictionary

There is a seemingly confusing yet common phenomenon that in some papers whose title includes “electronic dictionary” (for short ED), the phrase “online dictionary” frequently appears in the text or even becomes the research target. It gives rise to one question: what is the relationship between electronic dictionaries and online dictionaries. In order to answer this question, let us have a look at the definition of electronic dictionary first.

An electronic dictionary is “a type of reference book which utilizes computers and associated technology to present information on-screen. Examples include spelling checkers and thesauruses built into word processors, multi-volume dictionaries and encyclopedias on CD-ROM, multilingual terminological databanks and translation systems, research corpora, hypertext and the Internet” (Hartmann and James 2000: 47).

As can be seen from the definition above, online dictionaries are part of electronic
dictionaries, which can also be vividly demonstrated in the typologies of dictionaries of some scholars (e.g. Martin 1992; Lehr 1996; Nesi 2000; Li Ming and Zhou Jinghua 2000) in which electronic dictionary is regarded as the umbrella term of various kinds of electronic dictionaries. Online dictionaries have been further divided into different types by different scholars (e.g. Lu Gusun 2006; Gao Yongwei 2012; Wang Meng 2010; Lew 2011b; Tarp 2012; Zhang Yihua 2013) and based on the present classifications, the writer attempts to categorize English online dictionaries into four types in this paper: 1) online version of the paper dictionary; 2) meta-online dictionary; 3) dictionary websites; 4) semantic dictionary.

2.2 Online dictionary research

The art and craft of online dictionaries is referred to as “cyberlexicography”, proposed by Carr, who defines it as “employing the Internet to compile or create a dictionary” (1997: 209). Some other scholars (e.g. Hartmann and James 2000; Zhang Jingwen 2001; Zhang Yihua 2013) have also attempted to define online dictionary. Among various studies on online dictionaries, a great many scholars (e.g. De Schryver 2003; Gao Yongwei 2012) demonstrate the obvious advantages of online dictionaries as well as their current shortcomings; some concentrate on studying the effects of online dictionary on vocabulary acquisition and retention in comparison to other types of dictionaries. For instance, Granger and Paquot (2012) discover that online dictionary and CD-ROM dictionary are more beneficial for comprehending, producing and memorizing vocabularies but some other scholars (e.g. Schmitt & McCarthy 1997; Kent 2001) claim that the retrieval ease of online dictionary use will result in the shallow processing of the consulting word, thus being detrimental to vocabulary retention. Research into the use of online dictionary also attracts certain attention of some scholars (e.g. Wiegand 1998; Welker 2008; Tarp 2009). While the analysis of log files has become a common way to achieving the desired goal, C. Müller-Spitzer, A. Koplenig, and A. Töpel (2011) believe that it cannot reflect actual user demands. Anna Dziemianko even conducted a study to investigate whether the color of functional labels in online dictionaries exerted influence on the speed and effectiveness of dictionary search and the result shows that it improves vocabulary retention. P. A. Fuertes-Olivera (2013) proposes some recommendations for compiling high-quality online dictionary for translation.

However, observing the current studies, the writer finds that few attempts have been made to approach online dictionaries from the sociocognitive perspective even though this perspective has been used for language use and learning which is closely related to dictionary use. Therefore, this paper attempts to study English online dictionaries from this perspective in terms of compilers and users.
3. Theoretical Framework

3.1 Some propositions of sociocognition

Sociocognition is based on the view that neither language use nor language learning can be adequately defined or understood without recognizing that they have both a social and a cognitive dimension which interact (Batstone 2010). According to Batstone (2010), sociocognition can be divided into holistic sociocognition and analytic sociocognition.

3.1.1 Holistic sociocognition

Holistic sociocognition emphases “the inseparability of the social and the cognitive at every level and any attempt to examine them as separate entities is destined to fail, since by definition it will misrepresent the nature of the very phenomenon it aims to illuminate” (Batstone 2010). One of the representative scholars who agree with this view and have carried out several studies based on it is Dwight Atkinson (Atkinson 2002; Atkinson et al 2007) who observes that the social and the cognitive interweave in linguistic description, in first and second language acquisition and in language use. According to Atkinson (2002), language never occurs apart from a rich set of situational/sociocultural/historical/ existential correlates, and to separate it out artificially is to denature it.

The inseparability between the social and the cognitive can also be understood in terms of the unity between the mind, the body and the world. It is believed that cognition is not only situated exclusively in the minds of individuals, but also linked to and often assisted and enhanced by social factors in the environment such as the use of various kinds of gestures, the relation between two entities, etc.. For example, the mental processes of language learning cannot meaningfully be separated from the support a learner receives from another interlocutor (Lantolf 2000). Embodied cognition, situated cognition, theories of learning as participation and ecological perspectives on language use and language learning has laid theoretical foundation for the emergence of this view.

3.1.2 Analytic sociocognition

Analytic sociocognition can be distinguished from holistic sociocognition in two ways. Firstly, “although the social and the cognitive are interdependent whenever they are in process, they can usefully be analyzed at a theoretical level as separable elements” (Batstone 2010: 10). Secondly, “particular elements or combinations of elements act upon one another lead to partially predictable outcomes, and at last some of these elements can be influenced by careful pedagogic intervention” (11).

As far as relationship between mind, body and world is concerned, analytic
conceptions of sociocognition are quite different from those of holistic ones. Mind and body are regarded as entirely separated entities, which is also referred to as ‘mind/body dualism’. One area of research which vividly reflects analytic sociocognition is variationism.

It is worth mentioning that the writer assumes that the sociocognitive perspective in this paper adopts the thoughts of holistic sociocognition. This means social factors and cognitive factors should be integrated to study English online dictionaries. The lack of either side will fail to comprehensively illustrate the key points.

3.1.3 Alignment

According to Atkinson, “alignment is the complex means by which human beings effect coordinated interaction, and maintain that interaction in dynamically adaptive ways” (2007: 169). Atkinson regards alignment as a fundamental tenet of the sociocognitive approach to second language acquisition. But it is worth mentioning that alignment is by no means confined to the field of second language acquisition. It can also be applied to the process of learning on a broad level.

The mind-body-world environment is constantly changing, so individuals have to adapt themselves to it accordingly in order to survive. The ongoing alignment of human organism to its changing environment can also be deemed as learning. In fact, “learning is the default process of continually aligning oneself with one’s sociocognitive environment” (171). This kind of learning is accomplished through participation in and interaction with the environment. Without aligning to one’s environment, however, one will be soon eliminated from the world.

3.2 Cognitive Theory of Multimedia Learning

The cognitive theory of multimedia learning (CTML) was proposed by Richard E. Mayer, a famous American educational psychologist in his Multimedia Learning. Mayer (2005) and some other cognitive researchers believe that people learn more deeply from words (spoken words or written words) and pictures (e.g. illustrations, photos, animation, and videos) than from words or pictures alone, known as “multimedia principle” and that multimedia learning happens when learners build their mental representations from both words and pictures (Mayer 2005).

3.2.1 Three assumptions of the cognitive theory of multimedia learning

In order to maximize effective learning, the designers must have a clear picture of how people learn in the process of designing multimedia message. In other words, the design of multimedia messages is shaped by the designers’ conception of how human mind works. According to Mayer and Moreno (1998) and Mayer (2005), CTML is composed of the following three assumptions: the dual-channel assumption, the limited capacity assumption, and the active processing assumption.
a. Dual-channel assumption

The dual-channel assumption holds that humans possess separate information processing channels for visually represented material and auditorily represented material (Mayer 2005: 33). This assumption is based on the proposal that the human information-processing system contains two channels, namely, the auditory/verbal channel and a visual/pictorial channel. Humans process the information presented to the eyes such as illustrations, animations, videos, on-screen texts in the visual channel while they process the information presented to the ears such as narration, nonverbal sounds in the auditory channel.

However, learners may be capable of converting the processing of the information initially presented to one channel into the other channel. For instance, illustrations may first be processed in the visual channel as it is presented to the eyes, but an experienced learner may have the ability of mentally converting words into sounds, which are usually processed in the auditory channel.

b. Limited capacity assumption

The limited capacity assumption maintains that humans are limited in the amount of information that can be processed in each channel at one time (Mayer 2005: 35). When the learners are presented with a large amount of information, they can only be able to process part of the information in the particular channel. For instance, after listening a piece of English news such as BBC or CNN news for about two or three minutes, few except for the excellent professional interpreters could be able to remember all the information word by word. Most people can only remember some important information that is sensitive to their auditory channel. Therefore, knowing how much information can be processed in each channel at one time is of great significance. In spite of individual difference, on average, memory span is relatively small—approximately five to seven chunks (ibid). Other things being equal, if a learner could constantly enlarge the breadth of a chunk, he or she could be able to remember more information at a time.

c. Active processing assumption

The active processing assumption is that humans actively engage in cognitive processing in order to construct a coherent mental representation of their experiences (Mayer 2005: 36). There are three processes that are important for active learning, namely, selecting relevant material, organizing selected material, and integrating selected material with existing knowledge (Mayer 1996, 2001; Wittrock 1989). When a learner attempts to keep in mind of the incoming information, the active learning occurs, because in this process, the learner takes the initiative to process the information. The information is processed until a coherent mental representation is formed. This process is also called the process of model building. “A mental model (knowledge structure) represents the key parts of the presented material and their
relation” (ibid). Process, comparison, generalization, enumeration, and classification are generally viewed as the basic knowledge structure (Chambliss & Calfee 1998; Cook & Mayer 1988).

3.2.2 Three memory stores in Multimedia learning

Three memory stores involved in multimedia learning are sensory memory, working memory, and long-term memory. Sensory memory holds the incoming information whether in the form of pictures or in the form of printed words or spoken sounds for a very brief time period; working memory is employed for temporally holding and manipulating knowledge in active consciousness; long-term memory, as its name implies, can hold large amounts of knowledge over a long period of time (Mayer 2005). The following cognitive model of multimedia learning could vividly demonstrate the whole process of dealing with incoming information:

Figure 3.1 The cognitive model of multimedia learning (cited in Mayer 2005: 37)

4. Compilers’ Sociocognitive Approach to English Online Dictionaries

4.1 Compilers’ Sociocognitive Approach to the Emergence of English Online Dictionaries

The emergence of English online dictionary is the result of the interplay of both social and cognitive factors.

4.1.1 The emergence of English online dictionaries from the social perspective

First, the emergence of English online dictionaries can be analyzed from the social perspective. Moreover, this analysis can be approached from the following three angles.

a. Technological advancement

It is undoubted that the emergence of English online dictionaries is driven by the “technological innovation, widespread and generalized full access to the Internet” (de
Schryver 2003: 150) in modern society without which online dictionaries would never have the chance of coming into being not to mention gaining its present popularity. Therefore, it is no exaggeration to say that the English online dictionary can be deemed as one of the social products of technological development which acts as the driving force in the advancement of dictionaries.

b. Language, dictionary and society

The social side of dictionaries can be reflected from the relationship between language and society as dictionaries are generally considered as the record and carrier of language. It should be noted that there is a profound relationship between language and society. On one hand, human beings acquire and use language in society, so language can be regarded as a part of society. Also, language is “the principal means whereby people conduct their social lives” (Kramsch 1998: 3). What is more, language evolves with social changes. For example, the change from the old English to the modern English occurred against the background of hundreds of years’ social development. Furthermore, social culture, social development as well as the improvement of social life can be indirectly observed from the selection of entries in dictionaries, just as Bejoint mentioned in his The Lexicography of English, “the worldviews of a society can be mirrored through the selection of words, the wording of definitions, the choice of examples and the use of usage labels or comments” (2010: 76). For example, the first version of the New English-Chinese Dictionary included many words and examples which could reflect the characteristics of the socialist revolution and socialist construction at that time. There are some relevant examples listed as follows:

(1) Social-imperialism is capable of anything foul and base.
(2) People’s communes are fine.
(3) The old poor peasant’s face bears the stamp of his sufferings in preliberation days.

It is obvious that the underlined word and phrases in the above three sentences carry the typical political characteristics of the 1970s when this dictionary was compiled and published. In 2009, the fourth edition of the New English-Chinese Dictionary came out, and compared to the first edition, it indeed underwent certain changes. For example, it incorporates entries ranging from the field of sociology, culture, science and technology to that of daily life, etc., such as “water footprint”, “eco-management”, “plastic soup”, “netbook”, “facebook”, “mobisode”, “guerilla marketing”, “helicopter parent”, “barefoot luxury”, so on and so forth. Furthermore, it also includes several cyberspeak in the appendix such as “aak” (asleep at keyboard), “aslmh” (age, sex, location, music, hobbies), “kiss” (keep it simple and stupid), “yyssw” (yeah, yeah, sure, sure-whatever), etc.. It is also clear that all these cyberspeak included into this dictionary reflects some aspects of people’s modern life: This kind of change in printed dictionaries is definitely not incidental; instead, it is an inevitable trend. It can therefore be concluded that “dictionary can serve as the mirror both of the conscious (or unconscious) social values and judgments of its compilers, and of the chief social
and ideological concerns of the period of the compilation” (Cowie 1995: 294).

Though slightly different from printed dictionaries in the way of establishing relations with society, the English online dictionary is linked to society by its rapid upgrade rate. It can quickly absorb the widespread yet unstable words. For example, in 2015, “duang” had enjoyed a wide popularity since it appeared in a TV interview in March, however, when the writer consulted the Youdao online dictionary and iCIBA online dictionary, two well-known domestic English online dictionaries, in May, she found that both online dictionaries had already taken in this word which is demonstrated respectively as follows:

![Figure 4.1 The incorporation of “duang” into Youdao online dictionary](image1)

![Figure 4.2 The incorporation of “duang” into iCIBA online dictionary](image2)

It can be seen from the explanation of this word, especially the one in Youdao online dictionary, that language originates from society and reflects the social phenomenon. From the above analysis, one can draw the conclusion that as the medium of language, online dictionaries are also socialized.

On the other hand, the use of language such as the meanings of words, the situation to which the uses of proper words are subject, as well as the language variations, should be in conformity with social conventions. This implies that language bears inseparable relationship to culture. Language not only expresses cultural reality
but also symbolizes cultural reality (Kramsch 1998). In terms of the relation between language and dictionaries, it is unavoidable that dictionaries are also engraved with cultural elements. Dictionaries, as one of the cultural products, record languages which have been acknowledged by a specific society and they exert certain influence on the social culture at the same time. This is also true for English online dictionaries since it is the common feature of dictionaries. Therefore, it can be concluded that the English online dictionary uptakes nutrients from social culture and also reflects the changes of it.

c. Competition and cooperation

First, competition and cooperation coexist between printed dictionaries and their online counterparts. On one hand, the increasing popularity of online dictionaries poses an obvious threat to paper dictionaries causing the decreasing of the number of their sales (Lew 2011b; Gao Yongwei 2012). On the other hand, the online version is viewed as the best advertisement for its printed counterpart thereby promoting its sales (Lew 2011b; Zhang Jingwen 2001) and printed dictionaries will further popularize their online equivalents (Lew 2011b).

Second, competition and cooperation are intertwined among online dictionaries. For one thing, it is the internal competition that promotes publishers of the “Big Five” to bring out their online versions successively, and constantly perfect them. For another, online dictionaries are made available in a variety of ways such as free access, a paid subscription for partial content (often extended or more professional one) and a paid-only service, therefore they are more popular among users compared to paper dictionaries which always charge users of certain money. We can conclude that online dictionaries cooperate with one another in the face of the competition from other types of dictionaries. Since both competition and cooperation are very common phenomena in society and they are often the driving force of new things, it is sound to analyze the emergence of online dictionaries from the social perspective.

4.1.2 The emergence of English online dictionaries from the cognitive perspective

This section concentrates on exploring the role that cognitive factors play in the emergence of English online dictionary.

First, as has been mentioned above, the emergence of English online dictionaries is driven by the social development reflected in its technological advancement. However, if lexicographers/compilers had not recognized the technological development in society not to mention consider taking advantage of it in dictionary-making, online dictionaries would probably never emerge in the history. Fortunately, the social development is sensed by compilers and they might have thought about how to utilize it to better compile dictionaries or to compile better dictionaries. Confronted with changes in society, they either choose not to be influenced and thus to remain unchanged or embrace the development and make some changes accordingly. It is clear that the latter is a wise choice, as without aligning to
one’s environment, one will be eliminated from society sooner or later (Atkinson 2007). It seems that compilers were acutely aware of this point and made the right choice. It is on the basis of this that dictionaries are able to exist on the Internet as it does nowadays and be called online dictionaries.

Second, as stated before, language, society, culture and dictionaries are closely related to each other. Dictionaries can record language and reflect the social culture of a specific community at a particular time. As is known to all, the upgrade rate of traditional paper dictionaries is rather slow, and some new words, phrases and cultural phenomena cannot be absorbed into dictionaries timely. But new words and phrases are constantly coined due to the rapid development of society. How to take in them more quickly in order to better satisfy users’ punctual needs is a problem in front of compilers who will certainly figure out a way to resolve it. It turns out that humans are indeed “highly attuned co-cognizers” (Atkinson 2010: 5). Thanks to their cognitive efforts, the online dictionary which makes this become a reality and effectively improves the situation, comes into being.

Third, it is true that both competition and cooperation occur between paper dictionaries and online dictionaries. But it should not be ignored that human mind is also involved in these two typical social behaviors. After all, dictionaries, whether paper dictionaries or online dictionaries, are the outcome of human thought and human action. Dictionaries themselves cannot compete or cooperate with each other, because they do not have thoughts. It is because of the fact that they can be viewed as the record of the thoughts of compilers that competition and cooperation can coexist between paper dictionaries and online dictionaries or among different types of online dictionaries.

From what has been analyzed above, it has been noted that English online dictionaries can also be analyzed from the cognitive perspective in terms of compilers.

4.1.3 The emergence of English online dictionaries from the sociocognitive perspective

As can be seen from the above, English online dictionaries can be approached from both social perspective and cognitive perspective. Therefore, the combination of both perspectives -- sociocognitive perspective, a new perspective, can be employed to explore the emergence of English online dictionaries. In a word, English online dictionaries would not be able to exist without the impetus and interaction of technological advancement and the active human cognition. In fact, these two factors cannot be investigated separately as any attempt to examine them as separate entities will “misrepresent the nature of the very phenomenon it aims to illustrate” (Batston 2010: 6).

4.2 Transformation of Modes of Compilation

It is evident that the mode of the compilation of dictionaries has gone through several changes.
4.2.1 Transformation of media of compilation

As pointed out by Wei Xiangqin and Zhang Boran (2001), the compilation of dictionaries went through four periods in terms of the medium and carrier: the written period, the printing period, the computer period and the Internet period. Also, dictionaries can be classified into paper dictionaries and electronic dictionaries and the latter can further be subdivided into CD-ROM electronic dictionaries, pocket (hand-held) electronic dictionaries and online dictionaries. In this regard, the media have been changed from the original paper to the present Internet.

It goes without saying that the development of dictionaries is closely linked to that of society. Each progress in technology in society becomes the driving force of the advancement of dictionaries. For instance, it is because of the wide popularity of the Internet and the mature development of computer and information technology that dictionaries are able to be presented on the Internet. But this cannot be accomplished without the cognition of this development on the side of compilers who recognize the social development and are interested in taking advantage of it to the compilation of dictionaries.

In addition, it is also worth noting that mental activities are involved in the whole process of compiling dictionaries. Compilers are supposed to think about the layout of the microstructure and macrostructure of dictionaries including how to define an entry, how to organize the sequence of different meanings of an entry, and how to arrange the structure of the whole dictionary.

Therefore, when observing the transformation of the media of compilation of dictionaries, one should not consider the social side and the cognitive side separately but to take both of them into account so as to have a comprehensive perspective.

4.2.2 Transformation from compiler-orientation to user-orientation

It is well-known that dictionaries have always been regarded as the authority, and they are “often admired and revered with awe, and the status of ‘the dictionary’ in some countries could be linked to that of the lay Bible” (Lew & de Schryver 2014: 341). This is largely because compilers of dictionaries are always authoritative lexicographers and linguists and the compilation of a dictionary often goes through several years. At that time, great importance was attached to the compilers or lexicographers who were in charge of compiling dictionaries from the very beginning to the very end. Lexicographers have always determined the form and the contents of dictionaries by posing as dictionary users themselves while users can only passively wait to receive the information presented in these dictionaries. It is not exaggerated to say that compilers who seldom consider what do users really need, and who are accustomed to compiling dictionaries according to their own needs and experience (or intuition), play the most important role in the triangle consisting of compilers, texts and users (see figure 4.3). Thus the mode came into being and gradually became fixed: first, compilers are responsible for compiling dictionaries, arranging the content of the texts and finding out the way to organizing the texts, and then the texts are presented to
users once they buy one dictionary, however, it is worth mentioning that compilers rarely interact with users in the whole process as what can been shown by the dotted line in the left triangle of figure 4.3. As a result, it is not surprising that the compiler-oriented dictionaries lead to low efficiency in deciphering codes because users were burdened by the often “cryptic lexicographic content” (341) of dictionaries. Why dictionaries compiled by authoritative compilers cannot bring the expected effects in relation to language learning and language use? One of the reasons lies in the fact that the compilers often “systematically overstate the metalinguistic skills of the users” (Lew 2011a: 1).

Fortunately, however, an increasing number of compilers (e.g. Cowie 1999; de Schryver 2003; Bejoint 2010) have gradually come to realize the importance of realizing and satisfying users’ practical needs, specifically speaking, the decoding and encoding needs in the process of compilation. To echo with this trend, the word “user-friendly” or “user-friendliness” has been frequently mentioned in various papers with regard to lexicography. The former “entered the English language in 1977” (de Schryver 2003: 182), while the latter was originally used as a computer term. According to Bogaards and van der Kloot, “‘user-friendliness’ is used to describe efforts on the part of dictionary-makers to present information in ways that are assumed to be most easy for intended users” (2001: 98). This is especially true when it comes to the numerous advantages of online dictionaries through which users can not only enjoy the great convenience brought by the use of advanced digital technology and the increasing popularity of the Internet, but also interact with the compilers by commenting on the quality of definitions of some words or some given examples, sometimes even the quality of the whole dictionary and participate in compiling certain words, especially some terminologies. In this way, the mode of compilation has shifted from the traditionally dominant compiler-orientation to the currently prevalent user-orientation. And the relationship among the three parts, namely compilers, texts and users, has also changed correspondingly, which can be explicitly demonstrated by right triangle of the figure 4.3.

![Figure 4.3 The change of relationship among compilers, texts and users](image)

In the traditional printed dictionaries, it is often taken for granted that compilers are supposed to compile the texts and present them to users, and users seldom have the opportunity to play a certain role in the compilation of texts, therefore the arrow between texts and users is unidirectional with the former acting as the starting point and the latter the ending point while in online dictionaries, users are capable of
contributing to the compilation of dictionary texts, in this way the arrow between them becomes bidirectional; Also, unlike the printed dictionaries in which there are few interactions between compilers and users, online dictionaries enable users to communicate with the compilers by virtue of various modern tools such as emails, blogs, websites, etc., as de Schryver (2003: 160) puts it, “a ‘letter to the publisher’ is much more easily written in an online electronic medium than when one has to actually pen down comments”, therefore in the right triangle the writer uses bidirectional arrow to replace dotted line in the left one to mark the change of relationship between compilers and users. Here, two questions naturally arise: why such kind of transformation could happen and what causes these changes? The two questions can be answered from the following three perspectives.

a. User-orientation

It might be possible that this change in lexicography is more or less influenced by the literary theory of “Reader-response criticism” which emerged in the 1960s and 1970s with Norman Holland, Stanley Fish, Wolfgang Iser, Tonathan Culler and David Bleich as its representatives who advocated to lay more emphasis on readers instead of authors and texts because “meaning is not an inherent property of texts, rather, meaning is made by a reader, in the act of reading” (Fowler 2008: 127). It is universally acknowledged that the theoretical development of lexicography relatively lagged behind that of other fields, so it often borrows and absorbs theories from various fields. Though “a dictionary is meant to be consulted, not read” (Bejoint 2001:10), the content of a dictionary is often regarded as a text, and thus the person who consults the content, to a large extent, can be viewed as a reader. Seemingly, user-orientation and reader-orientation are different in terms of the subject, but due to the fact that the concept of “user” and “reader” overlaps here, it is really difficult to completely differentiate one from the other in the realm of lexicography. Therefore, it might be safe to conclude that the transformation from compiler-orientation to user-orientation in lexicography cannot be analyzed separately without taking the external influence of “reader-response criticism” into consideration.

b. From “top-down” to “bottom-up”

“Before the Internet, dictionaries were typically created ‘top-down’ from editors, through publishers, to readers” (Carr 1997: 214). Then with the advent of the Internet and the development of computer technology, the online dictionary appeared in the last decade of the 20th century thus gradually yet dramatically transforming the “top-down” model into “bottom-up” one as far as dictionary making is concerned. By the so-called “bottom-up collaborative editing” (de Schryver 2003: 160), users are able to approach dictionaries directly through the Internet in which they can also participate in the compilation, thereby increasing the share of their contributions to the compilation of dictionaries. Urban Dictionary, for instance, is a bottom-up initiative. By proposing that “URBAN DICTIONARY IS WRITING BY YOU”, it not only actively encourages
users to define a word by themselves, but also calls upon users to vote on the best
definition of a word compiled by other users.
Bottom-up creation demonstrates “considerable communicative and sociological
potentials” (Carr 1997: 215). With the increasing involvement of users in dictionary
compilation, online dictionaries show their social attributions in a more obvious way.
In other words, social elements cannot be ignored in this type of dictionary owing to its
special way of compilation. It is worth noting that this transformation can also be
employed to explain why users are able to gain its present significant status.

c. From compilers’ compilation to crowdsourcing

When paper acts as the main or exclusive medium of dictionaries, compilers have
played a dominant role in the compilation. It is either a group of professional
lexicographers or the cooperation of a large number of expert lexicographers and
linguists who are gathered together to compile a dictionary, since compiling a
dictionary is usually assumed to be the kind of work that can only be accomplished by
these groups of people in terms of competence and professional knowledge. It usually
took a great many years to compile one paper dictionary and once the dictionary was
published, it is almost impossible to make any change even if some obvious mistakes
were found and they had to wait to be corrected until the second or revised edition
came out.

Fortunately, however, the emergence of online dictionaries has effectively
improved the situation since crowdsourcing has been exploited to compile dictionaries.
Different from the traditional model, crowdsourcing makes it possible for more people
to participate in the compilation of dictionaries.

The new term “crowdsourcing” which is usually regarded as portmanteau of
“crowd” and “outsourcing” was first coined by Jeff Howe in 2006. Though the word is
new, the idea behind it is not in the field of lexicography. Despite today’s largely
Internet-based crowdsourcing projects, the Oxford English Dictionary is arguably the
first reference work to be crowdsourced. In 1879, OED’s first editor, James Murray,
made an appeal to the “English-speaking and English-reading public” to ask volunteers
to note down unusual words as well as the sentences they appeared from periodicals,
pamphlets, works of literature, etc., and submit their findings to this dictionary’s
editors. This process is just what crowdsourcing currently advocates. Nowadays,
crowdsourcing has been more widely exploited in the compilation of dictionaries. For
example, the OED has launched “OED Appeals” (http://public.oed.com/appeals/),
which is a major online initiative set to involve the public in tracing the history of
English words.
Wiktionary, whose name is the blend of words “wiki” and “dictionary”, is another typical example. It is also called a “crowdsourced wikipedia style dictionary”, which relies on the “collective intelligence or wisdom of crowds” (Meyer & Gurevych 2012: 259), not only because it is mainly written collaboratively by an enormous number of volunteers, but also because it is an Internet-based multilingual and interactional dictionary. Through crowdsourcing, users of Wiktionary can be both compilers, the “Wiktionarians” (2012: 271), and consumers. Here, in fact, it should be noted that the nature of aforementioned bottom-up model and crowdsourcing model is almost the same, both of which “put dictionary users in the shoes of lexicographers” (cited in Muller-Spitzer & Wolfer 2015: 4). In both models, users are able to edit entries, playing the role of compilers. They just approach the same phenomenon in different terms; therefore some scholars (e.g., Meyer & Gurevych 2012) choose to mix the use of them to refer to this phenomenon instead of strictly distinguishing one from the other. For example, in terms of its way of compilation, Wiktionary is not only regarded as a crowdsourced dictionary (e.g., Muller-Spitzer & Wolfer 2015) but also deemed as a dictionary compiled in the bottom-up approach (e.g., Meyer & Gurevych 2012).

To conclude, all these three aspects can independently explain the transformation from compiler-orientation to user-orientation. They are, however, delineated and analyzed respectively in this part in order to provide more perspectives to address this phenomenon. No one can definitely say which perspective on earth exerts the decisive influence on the change in lexicography. But according to the the time they (i.e. Reader-response criticism, bottom-up approach and crowdsourcing) appeared, it might first be influenced by Reader-response criticism which causes it to shift its focus from the compilers to the users and start to pay attention to the actual needs of users, while it is through the bottom-up and crowdcouring models that users obtain the opportunities of participating in the practical compilation which further develops the concept of user-orientation and strengthens users’ status in the field of lexicography.

It can thus be inferred that the development of the theory of lexicography is not isolated; rather, it is closely related to and influenced by the social development, not just in terms of its technological advancement, but also in relation to its theoretical development trends.
5. Some Characteristics of Online Dictionaries from the Users’ Sociocognitive Perspective

When compared to the traditional paper dictionary, it is obvious that online dictionaries demonstrate a large number of advantages which can be mirrored through some of their unique and innovative characteristics that can be regarded as the outcome of the interplay of the cognitive and the social factors.

5.1 Convenient retrieval system

Equipped with a convenient “retrieval system” (Nesi 2000: 839), online dictionaries provide great convenience in accessing the content of dictionaries from the users’ perspective.

5.1.1 Multiple carriers

Different from traditional paper dictionaries whose carrier is merely confined to the paper form, online dictionaries possess multiple carriers such as the desktop computer, the laptop computer, various cellphones, iPad as well as other electronic devices available nowadays. Due to this kind of flexibility in the carrier, the online dictionary outweighs its paper counterpart in terms of its portability. It is also because of the flexibility in the carrier and the fact that the content is usually accessed through the Internet that an online dictionary can enable hundreds or thousands of users from different regions even countries to do the consultation at the same time and satisfy their different needs whereas if one person is using a paper dictionary to look up words, the other person cannot use it at the same time period.

Paper dictionaries are visible because whenever you intend to consult a new word, you have to turn to a brick and mortar dictionary at hand for help which is quite inconvenient when traveling outside. However, the case is different when it comes to the online dictionary. Online dictionaries are somewhat invisible as they are usually embedded into electronic devices and you are able to use them anywhere and at any time as long as you own an electronic device and have the access to the Internet. Since the modern electronic devices tend to be small and exquisite, it is much easier and more convenient to bring anyone of them, and users only need to download one or more dictionary APP which will sometimes satisfy their needs even offline. In brief, the online dictionary in a variety of carriers greatly benefits the users in relation to its portability. Then what is the driving force behind the utilization of various carriers?

It is notable that with the rapid development of information and technology and with the improvement of people’s living standards, electronic devices like cellphones have almost become an essential part of people’s life. It is also a trend that people pay increasing attention to the ecosocial development and advocate to use less paper in various fields so as to reduce the consumption of limited natural resources. Against the background, the carrier of dictionaries drastically changes from merely depending on paper to also relying on some electronic devices thus gradually achieving paper-less
progress. This development trend is in conformity with the social development. In fact, it is the result of human action and it involves the cognition of and the alignment with external environment. It is no exaggeration to say that the development of any field is the continuous alignment with the ecosocial environment, as “without aligning to one’s environment-without, that is, functioning as part of a larger, interactive mind-body-world ecology-human beings quickly cease to exist” (Atkinson 2007: 185), not to mention the development of areas at the disposal of human beings. It is because of the fact that the cognitive process is invisible that its participation and significance are often ignored. As can be seen from the analysis, the social factors and the cognitive ones are intertwined in the process, the coexistence of which forms a sociocognitive environment and the ignorance of any one fails to fully explain the reasons why the online dictionary could be able to exist in multiple carriers.

5.1.2 Efficient access

As one of the main benefits of online dictionaries, the multiplicity of access has “liberated users from the straitjacket of the printed page and alphabetic order” (Atkins 1996: 516). Therefore, users obtain more choices among a wide range of search options including fuzzy search, incremental search, all text search, etc., besides the traditional one. These make it possible for quick and easy access through online dictionaries. In addition, the “copy-and-paste facilities” (de Schryver 2003: 169) also quicken the retrieval process.

The speed of retrieval is usually viewed as an important testimony of a good dictionary, as de Schryver (2003: 173) puts it, “no matter how outstanding the contents of a dictionary, if the contents one needs at a particular point in time cannot be accessed in a quick and straightforward way, the dictionary (de facto) fails to be a good dictionary”. The multiplicity of access of online dictionaries enables users to retrieve information in a quicker way than ever before. When consulting the paper dictionaries, users have to know exactly the spelling of the consulting words, and then look through the dictionary word by word in an alphabetic order which will take up a certain time. This is just the case of consulting a single word. It takes longer time when users intend to consult the meaning of multi-word expressions (MWE’s) such as a phrase, an idiom or a proverb, because they have to make a judgment about which one is the headword at first and then obtain the desired result through looking up a dictionary for the headword. However, in the case of online dictionaries, the situation is considerably different. It has been noted that even when users are not clear about the whole spelling of a specific word, still they will successfully obtain the desired information at a relatively rapid speed. This is because, in the process of consulting online dictionaries, when users type one or two letters, they are presented with several words starting with these letters and can observe among them whether there is the very word that they are looking for and then make the corresponding choice quickly. This way of searching is called incremental search. Also, even when users carelessly type in a misspelling, they will be given a list of suggestions. For example, when one intends to look up “pursuit” in the CALD Online (see http://dictionary.cambridge.org/), but wrongly types in
“pursuit”, he or she will be given the following information:

**Search suggestions for “pursuit”**

We have these words with similar spellings or pronunciations:

1. pursuit
2. pursue
3. pursuant
4. pursuits
5. pursued
6. pursuer

As can be observed from the above, confronted with the wrong spelling of the word “pursuit”, instead of presenting no information at all, CALD Online provides users with more choices such as “pursuit”, “pursue”, “pursuant”, etc., according to the given information. In this way, the user can find out his or her own typing mistake, choose the appropriate word on one hand and acquire more words with the similar spelling at the same time thus helping broaden his or her vocabulary on the other hand. Being able to rapidly abstract the most relevant information from a large number of data is one of the great advantages of online dictionaries. But this cannot be accomplished without the support of the advanced computer and information technology brought by the social development. Moreover, the design of gathering words with the similar spelling together when the searching information given is vague or incorrect is built upon the adequate understanding of users’ cognitive rules and habits. The designers have perceived users’ learning habits, kept them in their minds and endeavored to make their design conform to them. Thus, as one of the products of “cognition-for-action” (Atkinson 2010: 4) and integrated social advanced technology with this cognition, the online dictionary can not only improve the retrieval speed but also better users’ vocabulary learning.

Furthermore, besides the above-mentioned situation of the consultation of a single word, in the consultation of information more than one word such as a phrase, an idiomatic expression or a proverb, users are able to obtain the results within seconds after they type in one to be consulted without the necessity to consider which one is the headword among them. With only typing in the target word or phrase and a click, the result is soon presented in front of users, which is made possible by the use of the computer technology, because “the computer has a great capacity for rapidly looking up specified words, parts of words, groups of words, or information categories” (Svensén 1993: 253).

Last but not least, it is noted that “advance in speech recognition now make voice research a viable option” (Lew & de Schryver 2014: 350). This can benefit those who only know the pronunciation but are not clear about the spelling of a word. Even who know both pronunciation and spelling can also use this way to consult the word at a
rapid speed without typing in the word in the search box.

c. Specific user interface of information presentation

Different from the traditional paper dictionary in which the target content is usually surrounded by some other irrelevant information, the online dictionary presents the target content exclusively in a specific user interface.

For instance, when consulting the word “emerge” in both OALD 7 and OALD online, one can obtain the results respectively as shown in the following:

![The consulting result of OALD7](figure5.1)
![The consulting result of OALD online](figure5.1)

**Figure 5.1** The comparison of information presentation between OALD7 and OALD online

As can be seen from the above, besides the target word, the presented page of OALD7 also includes many other words, and because the word “emerge” lies in at the bottom of the right side of the page, one has to scan all the words in this page one by one in order to find this target word while in the case of the OALD online, it only provides users with an interface in which only the information about the target word is presented, therefore, users are able to focus on the consulting result thus avoiding distraction of their attention. It is clear that the online dictionary liberates paper dictionaries from the restrictions of the linear texts on a superficial level. But apart from benefiting from the computer and information technology, there might be a crucial reason for this change on a deep level.

As mentioned in chapter three, humans are limited in the amount of information that can be processed in each channel at one time (Mayer 2005: 35). Here, since the
information is presented either on screen or in page, it is processed in the visual channel. The provision of much more information than required by the user in a given consultation will cause the information overload, “a phenomenon which may hinder and even obstruct the retrieval of the information needed” (Tarp 2012: 255). Therefore, it is obvious that the specific user interface of information presentation in the online English dictionary can effectively avoid the overload of information that might be caused by the presentation mode of the paper dictionary. Chances are that the designers of online dictionaries realized this shortcoming of paper dictionaries and attempted to change the situation, and that is why the online dictionary could present information in the present way. In addition, it should be noted that the computer and information technology plays an important role in putting it into practice. Thus it can be concluded that the social and the cognitive are interwoven with each other and jointly contribute to the specific presentation mode of online dictionaries.

In addition, considerable online dictionaries have flexible control buttons through which users can “control the presentation of lexicographical data” (Lew 2011b: 242), choosing to see more or less information. Macmillan English Dictionary Online, for instance, provides two presentation modes which can be chosen by flipping the Show more/Show less control button next to the lemma sign. This kind of design is the outcome of taking different levels of users’ different needs into account. If some users just want to know the pronunciation and the meaning of a specific word, for example, they only need to click the relevant button through which the needed information is displayed while other irrelevant information remains unseen. The reason for such kind of design lies in the fact that different individual users have different consultation needs and it is unwise to put all the same information in front of them as too much information at one consultation will confuse some users and increase their memory burden.

5.2 Dynamic repository of knowledge

Unlike paper dictionaries whose content becomes fixed once got published, the online dictionary can be regarded as a dynamic system, a dynamic repository of knowledge, which is reflected by its rapid incorporation of neologism, the wide application of multimedia and quick corrections of mistakes. Therefore, theoretically speaking, it has great potential of never going out of date.

5.2.1 Freedom from the space constraints

“Space limitations have always been one of the main frustrations of dictionary professionals” (Granger 2012: 12), and it is because of this, in the process of dictionary-making, compilers of paper dictionaries have to make a careful selection of entries according to the type and function of the dictionary and the specific needs of the target users. As a result, the number of entries covered in one paper dictionary is quite limited, therefore it is not rare that one cannot find a terminology in a specific field in a general-purpose dictionary and in order to finish a translation involving
several subjects, one has to look up more than one dictionary which is a great cost in both time and money and which is quite inconvenient. Also, constrained by the space, it is usually impossible to incorporate too many pictures into a paper dictionary even though “people learn more deeply from words and pictures than from words alone” (Mayer 2005: 47). It is also notable that because of the characteristics of stability, a printed dictionary usually does not include neologism.

However, they are no longer problems at all in the case of online dictionaries. Freedom from the space constraints and on-demand updatability are viewed as true advantages of web-based dictionaries (Nichols 2010). Firstly, without the constraint of space, online dictionaries are able to cover more vocabulary compared to paper dictionaries. For example, the Oxford English Dictionary online (OED online) covers 600,000 words and 3 million quotations which is impossible for one paper dictionary. To have a better understanding of the number listed above, it is necessary to have a look at one paper dictionary with a relatively large vocabulary size. For instance, as one of the well-known dictionaries in China, The English-Chinese Dictionary (the second edition) claims that it covers 220,000 words or so. This dictionary has more than 24,000 pages and weighs about 1.5 kilograms which means that it is of great inconvenience for users to carry it around whenever they need. But it is totally possible to bring more than one online dictionary with oneself, because online dictionaries are “virtual” (Atkins 1996: 516), and “they might only exist at the time of access” (de Schryver 2003: 163). On the other hand, online dictionaries often have the capacity to include vocabularies from various fields in order to meet diverse needs of different levels of users. Oxford Dictionaries online, for instance, advocates that they create content for general users, students, professionals, writers, game puzzlers, teachers, language learners, researchers, technology companies, and software developers. The unlimited space has also made it possible for online dictionaries to present as many examples as possible to better illustrate the usage of a specific entry (Gao Yongwei 2012).

Secondly, free from space constraints of the paper dictionary, pictorial illustrations are more likely and cheaply to be included in electronic dictionaries (Lew 2010). The visualized pictures together with the definitions of words play a better role in enhancing the learning efficiency of users. In addition to pictures, online dictionaries can also include audios, videos, animations, etc.

Thirdly, due to the large storage capacity and the great advantage of the Internet itself, instead of being static, online dictionaries have the potential of dynamically updating vocabulary bank and including new yet unstable words and phrases at a relatively quick pace.

There is no denying that it is the combination of the advanced computer and information technology and the great advantage of the Internet that make all these possible. Moreover, it is also based on the very recognition of the deficiencies resulting from the space constraints of paper dictionaries. In this regard, users’ punctual needs become the driving force of these designs of online dictionaries. Again, the social and the cognitive aspects are interacting with each other in the process.

b. The application of multimedia
Apart from the static pictorial illustrations which have been used for a long time by paper dictionaries, online dictionaries distinguish themselves by also including a range of dynamic multimedia content such as audio recordings of entries, definitions and examples, animations, videos, etc.

One of the most obvious innovations of online dictionaries in comparison to paper dictionaries is that the former provide audio recording of entries. In this way, both verbal form and non-verbal one are included into online dictionaries, thus enabling users to listen to the pronunciations of a specific word in addition to its spelling, definition, examples, etc.. Generally speaking, most English online dictionaries have provided both British and American pronunciations, and users only need to “click on the appropriate icon to hear the word” (Miller 2012: 46) in either pronunciation according to their preferences. It should be noted that the “audio presentation would afford dictionary users an opportunity to get more exposure to so-called real language” (Lew 2010: 296). According to the “dual-channel assumption” (Mayer 2005), the sounds presented to the ears are processed in the auditory/verbal channel and the visual images (e.g. spellings, definitions and examples) presented to the eyes are processed in the visual/pictorial channel. But the sound of a word can be converted into its visual image, and in this way, it is possessed in the visual channel. The mental conversion of a visual image into a sound can also happen. The double stimulus from both ears and eyes will incur double processing of the information thus facilitating the formation of long-term memory in which the information is stored as the prior knowledge. On the basis of this, when a user hears the pronunciation of the same word next time, he or she will unconsciously conjure up the spelling and the meaning of this word and when seeing the spelling of the word, he or she could select relevant information from the prior knowledge and thus organize it into the verbal model. This can also be used to explain why learners who skillfully grasp a large number of vocabularies and their corresponding pronunciations can always do better in understanding a piece of spoken news than those who only acquire considerable vocabularies and are not familiar with their pronunciations.

On the other hand, it is notable that with the globalization and the increasing communication among people from different countries, a higher standard of learning languages is required. For example, in the process of learning English, instead of being satisfied by the acquisition of the meanings of words, learners attach great importance to practicing and improving their listening abilities, because acquiring English is not just to obtain a decent mark on the paper, but to better communicate with people from western countries in the real life and in order to achieve this, learners should at least understand what their partners say. As a matter of fact, significant emphasis has been laid on the investigation of learners’ practical language skills including both speaking and listening capabilities when one applies to study abroad. The typical example is that the three famous language proficiency tests such as TOEFL, IELTS and GRE all have their due share in testing listening and speaking abilities. The prerequisite for speaking well is to know how to correctly pronounce each word. In reality, not everyone has studied how to pronounce each phonetic alphabet in a systematic way. Therefore, the provision of pronunciations of words is of great significance for this group of users, as
they can listen, imitate and gradually acquire the correct pronunciation. Even for those who have already known the pronunciation, the presentation of pronunciation is important in helping them check whether they have pronounced the specific word right through listening to the standard British and American pronunciations and if it turns out that they do make a mistake, they can make a correction in time. In addition, the input of both verbal and visual information will establish relations with prior knowledge thus further consolidating the already known knowledge.

In a word, the presentation of pronunciation in English online dictionaries is consistent with the social trend of laying increasing emphasis on the cultivation of English listening and speaking capabilities and the existence of it is indeed playing an undeniable role in facilitating English learning in practice.

5.3 Hyperlinks

The multimedia nature of the Web (Word Wide Web) has made it possible for online dictionaries to “present some of their micro-structures in a multimedia way through the hyperlinks to audio, graphic, or video files” (Gao 2012: 427). As a particular characteristic of online dictionaries, hyperlinks in general exist in two forms: the internal hyperlinks and the external hyperlinks.

5.3.1 Internal hyperlinks

Internal hyperlinks here refer to links to other information within the same online dictionary such as a word, a phrase, an idiom, as well as other information that are related to the original consulted information and can be found quickly through a simple click. It can be seen that the internal hyperlinks is, to some extent, similar to cross-reference in paper dictionaries but in fact they are just means of achieving cross-reference. There is no doubt that “the use of hyperlinks has made cross-referencing a much easier task” (Gao 2012: 427), because one does not have to find the cross-referenced information manually, instead, only by a click, the system will present the required information in front of users within a second. For instance, when consulting the word “hyperlink” in the OALD online, one will obtain the result as follows:
The above figure shows the consulting result of the word “hyperlink” and there are some nearby words on the right side. As can be seen, the words listed all start with the prefix “hyper” and they are in blue color. When pointing at any one of them, hypermarket for example, there would appear an underline and a click in it, will bring the following result:

As can be clearly seen from the above figure, the information obtained through the hyperlink also comes from OALD online. Therefore, hyperlinks as such are called internal hyperlinks. It is worth noting that most words in the online dictionary have more than one internal hyperlink.

5.3.2 External hyperlinks

External hyperlinks here refer to links to information in corpora, other reference works, the Internet, etc., beyond one online dictionary which can be quickly retrieved by a
click. The data in an online dictionary “may not be sufficient or accurate enough to meet users’ information needs in a specific consultation” (Tarp 2012: 262). Therefore, in order to provide more and better data satisfying users’ cognitive needs, it is necessary to link to additional data stored in the external sources. According to Rundell, “Online dictionaries could link directly to general and specialized corpora, allowing users to search for examples of any word, pattern, or linguistic feature they are interested in” (2010: 174). In addition, it should be noted that links to these external information sources are beneficial for “the comprehensiveness of the information sources” (Bothma 2011: 80) provided.

It should be noted that the information obtained through both internal and external hyperlinks is not the recreation of new information but the re-presentation of already existing data. There is usually a backtrack option that could keep the history of the hyperlinked items and enables users to go back to any of them (de Schryver 2003). Moreover, The hyperlink, no matter which type it is, is undoubtedly a faster and easier way to achieving cross-reference. Also, computer and information technology lays foundation for the implementation of hyperlinks.

Observing the hyperlinked items, one may find that hyperlinks do not randomly exist among items. Instead, items cross-referenced with one another through hyperlinks always have certain semantic, syntactic, phonological or pragmatic relations. It is impossible and impractical to present all kinds of information in relation to the given item, as the provision of too much information will possibly lead to information overload for users. The hyperlinks enable users of different levels to make their own choices on whether to open up the information they are interested in according to their practical cognitive needs. Humans are likely to establish relations among some seemingly irrelevant information so as to better memorize and extract when needed. So humans’ cognitive habits are taken into considerations in this design.

5.4 Customization

Customization refers to the process which “allows the dictionary to be adapted to users’ needs” (Granger 2012: 13). The situation is described by different scholars in different concepts. Rundell (2010) calls it “customization and personalization” while Tarp (2011) defines it as “individualization” and according to Sobkowiak (1999), dictionaries with such characteristics are flexible dictionaries.

There is no doubt that the large amount of data in online dictionaries and the ease with which it can be accessed are nothing at all if the information that users have obtained does not correspond to their true needs. Therefore, whether users’ needs have been met and to what extent their needs have been satisfied are indeed playing a decisive role in the quality of dictionaries. In fact, users’ needs have become a central issue in both lexicographic practices and theories. It is not unusual that “an individual user with individual information needs occurring in an individual situation decides to make an individual lexicographical consultation in order to access the concrete and individual data that may satisfy his or her individual needs” (Tarp 2012: 260). Nowadays, the real challenge is to “to look at the needs of dictionary users, of every
language, and every walk of life, users as diverse as people themselves, and give them the kind of information they need for whatever they are using the dictionary for” (Atkins 1996: 516). Tono (2011) once conducted a study on dictionary use exploiting the eye tracking technique, and the result shows that dictionary users differs greatly in terms of consultation habits and strategies. This implies that different users have different needs and expectations. In a word, specific users, users’ specific needs and the specific user situations should be taken as the starting point for all the lexicographic theories and practices.

According to Gamper and Knapp (2002), there are two types of customization in general, namely, adaptability and adaptivity. For one thing, dictionaries can be adaptable which means that users can manually customize their needs, and this is just the very meaning of the design of control buttons in online dictionaries. For another, dictionaries can also be adaptive. In this respect, online dictionaries automatically adapt to the specific user based upon the dictionary log files, as “no lexicographers, however well prepared, is capable of dealing with each and every one of the infinite number of individual needs that an infinite number of individual users may have in an infinite number of situations” (Tarp 2012: 260). To be specific, once an individual user submits some specific data on user profiling to the system, the system will automatically select, filter and present the specific data according to the user’s practical needs.

5.5 Hybridization

One of the most striking features of online dictionaries is that it blurs the original clear-cut dividing line among different types of dictionaries. It is not uncommon that an online dictionary contains not only lexical knowledge but also the etymological and encyclopedic information. This trend can be properly described by the word “hybridization” given by Hartmann, who further illustrates it as the “combination of one or more types of reference work in a single product” (Hartmann 2005a: 195).

As a matter of fact, the paper dictionary has demonstrated such kind of trend long before. For instance, a wide range of information such as grammatical information, etymological information, onomastic information, encyclopedic information, etc., has gradually been added to the general-purpose dictionary. But due to the fact that increased content means increased price and weight both of which are unlikely to be accepted by users as far as the cost and portability is concerned, it moves at a quite slow pace and makes a little progress which is so small that even can be ignored in comparison to the huge one achieved in online dictionaries.

Moreover, it is an irreversible trend that users increasingly expect to obtain various kinds of information they desire in a single reference work. In this regard, the conventional pure dictionary definitely cannot satisfy these diverse needs of users. Without the space constraints and the worry for increasing cost, online dictionaries take full advantage of the computer and Internet technology to include as much information as possible covering a wide area. In this way, users’ multiple levels of needs are better met in just one online dictionary.
Last but not least, by virtue of the inherent advantages of computer and information technology, users are in a better position to switch different dictionary types (i.e. monolingual, bilingual and multilingual dictionaries) within one system.

In brief, the occurrence of hybridization in online dictionaries also cannot be analyzed separately without considering the interplay of the social and the cognitive elements. First, there must be such kind of social need for this trend; secondly, this social need should be recognized and taken into account in designing online dictionaries.

The above-mentioned are just the analysis of some of the shared characteristics of online dictionaries from the users’ sociocognitive perspective. However, it is worth noting that besides these shared advantages, different types of online dictionaries also demonstrate their specific characteristics which will be illustrated in detail in the following section.

6. Conclusion

Generally speaking, online dictionaries can be regarded as one of the sociocognitive products. For one thing, online dictionaries are the sociocognitive products of social technological development, as they benefit significantly from the cognition and exploitation of the great technological advancement in society, the computer and Internet technology, to be specific. For another, online dictionaries can also be deemed as the sociocognitive products of the development of theories in various fields. For example, lexicography itself has been deeply influenced by theories in other subjects such as reader-response criticism, bottom-up model and crowdsourcing, thus gradually transforming from compiler-orientation to user-orientation. Users, user needs and user situation are taken as the starting point of dictionary making. Both social and cognitive factors are merged in this process.

Moreover, dictionaries can be studied from the sociocognitive perspective in terms of their relationship with society through language. Online dictionaries have the ability to quickly include new words or phrases and the action itself can not only better satisfy users’ cognitive needs, but also record language timely thus quickly reflecting the social phenomenon at a particular time period. In this regard, dictionaries are no longer static once published, instead, they are dynamically upgrading and thus deemed as a dynamic repository of knowledge due to the huge storage capacity without the restrictions of space which, coupled with advanced information technology, also makes it possible for the wide application of various multimedia such as audio recordings, videos and animations. The application of these multimedia might be useful for consolidating vocabulary acquisition and retention.

Dictionaries are essential for users in second or foreign language acquisition. Compilers must bear users’ cognitive habits and cognitive needs in mind in dictionary making. On the basis of this, more ways should be explored to improve consulting services in order to effectively enhance users’ learning efficiency. It is no exaggeration to say that without the consideration of both social factors and cognitive elements,
online dictionaries would not come into being, not to mention enjoy its present popularity.

References


Bothma, T. J.D. 2011. Filtering and adapting data and information in an online environment in response to user needs. In E-Lexicography: the Internet, digital initiatives and lexicography, eds. P. A. Fuertes-Olivera and H.


Abstract

The influence of foreign languages in the development of the South African indigenous languages, like Northern Sotho, leads to borrowing or foreign acquisition, which ultimately leads to the increase in the vocabulary of these languages. The languages which influenced the developments in the vocabulary of Northern Sotho, or Sesotho sa Leboa, include inter alia, the following languages which are spoken entirely in the Southern part of Africa: English, Afrikaans, Xitsonga, Tshivenda and the Nguni languages (i.e. IsiZulu, Isindebele, IsiSwati and IsiXhosa). The borrowing of lexical items from these languages is manifested either through direct borrowing or indirect borrowing. Direct borrowing, which is the direct conversion or the Sothoization of the foreign lexical items, leads to the formation of loan words while indirect borrowing, which is usually preferred by the purists, is done through the coinage system, where only the meaning or the sense of the lexical item is loaned and not the word itself. This newly acquired meaning is usually associated with that of the indigenous Northern Sotho lexical item. This system is called neologism and is always leading to the creation of polysemy and ambiguity. The main objective with this research is, therefore, to give a critical analysis of the effects of purism in the development of the Northern Sotho language, especially with regard to the adoption of foreign lexical items and the lemmatization of vocabulary. When confronted with the pressure of foreign acquisition, the purists always prefer to use coinage, or neologism, rather than direct conversion of foreign words into the vocabulary of Sesotho sa Leboa or Northern Sotho. The critical analysis of this purist policy reveals the following discrepancies:

1. Both the quantitative and qualitative developments of the vocabulary of the language is compromised because there is little or no increase in the vocabulary of the language, due to the fact that only the meanings are added to the existing equivalent indigenous word, thereby causing the word to be polysemous or ambiguous.
2. Neologism also leads to the coining of new words to avoid adopting or converting foreign words into the vocabulary of the language. This system does not only lead to ambiguity but also to the creation of words which are not usable, the so-called ‘Book Terminologies’. These coined words are named ‘Book Terminologies’ because the words are only kept in our books and the lexicons, but are never used by the people in any normal speech or conversations.

**Keywords:** neologism, purism, foreign acquisitions, polysemy, coinage

1. **Introduction**

Language development pertains mainly to the increase and development of the vocabularies and meanings in the language. This also pertains to the scientific development of spellings and orthography rules to record the vocabularies and structures of the language. The development of vocabulary is mostly facilitated through foreign acquisitions by way of borrowing or conversions of foreign vocabularies and meaning from other languages. The common method of foreign acquisition is done through direct borrowing or the conversion of the foreign word or through neologism, which is usually associated with indirect borrowing.

2. **Neologism**

Neology or neologism is defined by WhatIs.com (2017) as:

*the use of a new word or the use of an existing word but given a new meaning.*

In Accordance with this system, the Northern Sotho purist linguistic scholars use the existing Northern Sotho lexical items, which are more or less semantically equivalent with the foreign word, to refer to the newly acquired meaning.

3. **Purism**

According to dictionary.com (2017), Purism is a rigid adherence to or insistence on purity or nicety in the use of words, especially a word, phrase, or sense. **www.dictionary.com** defines purism as:

*Strict observance of or insistence on purity in language, style, etc.*
Purism is one of the dominating factors when it comes to the standardization of the South African indigenous languages like Northern Sotho. The majority of the Northern Sotho, or Sesotho sa Leboa standardizing officials and purist linguists use purism in several ways with the following two as the most common methods, i.e. (i) dialectal exclusions and (ii) severe linguistic restrictions on foreign acquisition.

a) Dialectal exclusions

With regard to dialectal exclusions, purism is manifested through intensive marginalization of other dialects by those in authority. In most cases, the standardizing officials tend to standardize their own dialects at the expense of those dialects which are not adequately represented in the standardizing bodies. As a result, most of the vocabularies from the side-lined dialects are not included in the standard vocabulary and, in most cases, are not even included in the language orthography. In this case most of these side-lined vocabularies are doomed to disappear from the Northern Sotho corpus instead of enriching the language. For instance, the Northern Sotho language has 30 dialects, but all these dialects are dominated by very few dialects like Sekone, Sepedi, and the dialects spoken in the central areas of Mookgophong, Polokwane, Ellisrus and Mokopane. The vocabularies of the majority of the dialects, which include Khelobedu, Setlokwa, Sehananwa, Seroka and Sepulana, do not form part of the Northern Sotho official vocabulary because the standardizing bodies side-lined these dialects on the grounds that these are impure dialect. The following are examples of the marginalized dialectal terms which, even though these dialects are spoken by the majority of the Northern Sotho communities, these dialects are still side-lined by the language:

<table>
<thead>
<tr>
<th>Dialectal</th>
<th>Standard</th>
<th>English meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khepepe</td>
<td>Bene</td>
<td>‘small truck’</td>
</tr>
<tr>
<td>Malobane</td>
<td>maabane</td>
<td>‘yesterday’</td>
</tr>
<tr>
<td>Melothi</td>
<td>Sejagobe</td>
<td>‘greedy’</td>
</tr>
<tr>
<td>Pikiri</td>
<td>komiki</td>
<td>‘cup’</td>
</tr>
<tr>
<td>Mnankhoṭo</td>
<td>Leribiši</td>
<td>‘an owl’</td>
</tr>
<tr>
<td>Mokhope</td>
<td>Morula</td>
<td>‘Marula beer’</td>
</tr>
<tr>
<td>Kotapeni</td>
<td>Afokhatha</td>
<td>‘avocado’</td>
</tr>
<tr>
<td>Mmidi</td>
<td>Lefela</td>
<td>‘mealie’</td>
</tr>
</tbody>
</table>

b) Restrictions on foreign acquisitions

Foreign acquisition, through which foreign words are converted and borrowed directly into the borrowing language, is the major system used by all developed and developing languages world-wide to bring the language up-to-date with the
developed international languages and cultures. Scientific research has positively proved that without foreign acquisition the language will rarely develop. Without foreign acquisition the language will never develop to be used as a language of education, language of commerce, language of science, language of research, etc. This simply shows that without foreign acquisition the language will never develop to be a complete official language to cater for its speakers in a developed world. Therefore, it is important to realize that linguistic scholars and the language standardizing officials should know that the language develops faster when there is sufficient contact with foreign languages, especially the developed languages. It is important for the language researchers to know that the indigenous language will never have equivalent terminology to express all the newly acquired items and concepts which originate with foreign acquisition. This is apparently because semanticists have proved the fact that there are no two words which might have exactly the same meaning even though they might be regarded as equivalents. Most South African indigenous languages borrow words from English and Afrikaans languages. Other languages like French, Latin and Greek contributed to a lesser extent, but these contributions mostly went through English.

With regard to foreign acquisition, purism is manifested in the form of avoidance for a directly adoption of the newly acquired terms into the vocabulary of the language. Most South African linguistic standardizing authorities prefer the use of neologism, i.e. the use of the existing related indigenous term to express the meaning of the new foreign word. In this way, only the meaning expressed by the foreign lexical item is accepted and adopted and not the word itself. This is usually done as a way of falsely demonstrating that the indigenous languages have their own terminology to name or to express new objects or concepts.

Another type of neologism is the coinage system where new terms are coined or formed to express the meanings of a foreign term. These coined terms are usually formed by using the indigenous terms which are more or less semantically related or equivalent to the meanings of the foreign terms and in most cases these are compound words. The main reason given by the purists in this regard is to keep the language as pure as possible from contamination by foreign terminology to keep the indigenous languages as pure as possible. The use of neologism aims to keep the original cultures of the indigenous communities as safe as possible from being swallowed by other foreign cultures.

4. **The effects of purism in the developments of the languages**

Purism is one of the important systems which help to purify the languages, especially the indigenous languages in Southern Africa. Purism, as it is applied in the standardization and orthography rules in some of the South African indigenous languages like Northern Sotho is usually done without considering the quantitative
vocabulary growth of the languages. Excessive use of purism through the use of neologism in the South African indigenous languages contributed much to a very low and poor growth in the development of the vocabularies of the languages. Neologism, as it is used in these languages, pertains to the avoidance of adopting or converting foreign terms into the structural vocabularies of the indigenous languages. As a result, purism leads to the following negative effects in the development of the languages:

a) **Ambiguity**: The use of neologism means that the foreign term itself is not adopted as a loan word, but only its meaning. The adopted meaning is included as one of the meanings of an existing indigenous term, which has a more or less equivalent meaning to the newly acquired sense. This means that the word which originally had its own meaning is now going to undergo meaning expansion to refer to another equivalent meaning. Research by semanticists like Leech (1978) has effectively proved that no two words can have exactly the same meaning, and as a result, the transference of the meaning of a foreign term to an existing indigenous term leads to the excessive rise of ambiguities in the meaning of words in the languages. In this way, the vocabulary is not growing, but only the meanings of the existing words are growing and the word which referred to one meaning will come to refer to several related meanings, e.g. the Northern Sotho word ‘makhura’ originally referred to ‘fats’ in English. Due to neologism, the word has come to refer to the following meanings which result from foreign acquisitions:

<table>
<thead>
<tr>
<th>Makhura</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Fats’</td>
</tr>
<tr>
<td>‘Cooking oil’</td>
</tr>
<tr>
<td>‘Petrol’</td>
</tr>
<tr>
<td>‘Ointments’</td>
</tr>
<tr>
<td>‘An exhilarator’</td>
</tr>
</tbody>
</table>

The word has five meanings and the exact meaning of the word can only be understood when used in context with a qualificative. This means that if I say: *mphe makhura* ‘give me fats’ one will not know whether I want cooking oil, petrol or an ointment until I qualify my sentence, e.g., as follows: *Mphe makhura a go tlola mmele* ‘give me a body lotion’.

b) The explanation in 4.1. tries to demonstrate that the use of neologism in Northern Sotho has not only resulted in the creation of ambiguity and polysemy, but also forms a barrier to the quantitative growth and development of the vocabulary of the language. The direct loan words which are commonly used by the communities in the languages are avoided by the standardizing authorities as
impure and unwanted forms of terminology in the language. For instance, most people use the following words when speaking in order to be unambiguously understood: phetrolo for ‘petrol’, ointmente for ‘ointment’ ekselereitha for ‘exhilarator’, which refer directly to what they mean, instead of makhura, which needs to be first qualified to be understood.

c) In most cases neologism leads to the creation of the vocabulary which may be referred to as ‘book terminology’, i.e. the terminology which will always be in our books, but which will rarely, or never be used by the people. These include even the very same purists who created these terms. Most of these words are too figurative and ambiguous, as in the following examples:

<table>
<thead>
<tr>
<th>English</th>
<th>Afrikaans</th>
<th>Zulu</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Overtake’</td>
<td>go sega kgopu</td>
<td>instead of obatheika</td>
</tr>
<tr>
<td>‘Petrol’</td>
<td>makhura</td>
<td>instead of phetrolo</td>
</tr>
<tr>
<td>‘Cellular phone’</td>
<td>sellathekeng</td>
<td>instead of selfounu</td>
</tr>
<tr>
<td>‘Automatic’</td>
<td>morethaoitiriśa</td>
<td>instead of othomethiki</td>
</tr>
<tr>
<td>‘Window’</td>
<td>lehlabaphefo</td>
<td>instead of lefesetere (Afrikaans: ‘venster’)</td>
</tr>
<tr>
<td>‘Informer’</td>
<td>tshebi</td>
<td>instead of mpimpi</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(from Zulu Language)</td>
</tr>
<tr>
<td>‘ATM’</td>
<td>Motšhenewatšhelete</td>
<td>instead of just ATM</td>
</tr>
<tr>
<td>‘Marathon’</td>
<td>Mokato</td>
<td>instead of just marathone</td>
</tr>
<tr>
<td>‘Ship’</td>
<td>Leselawatle</td>
<td>instead of Sekepe (Afrikaans: skip),</td>
</tr>
</tbody>
</table>

For instance, the coined word: sellathekeng is a compound coined word combining an adverb, sella- ‘that which cries’ and the noun -thekeng ‘the waist’ to name a ‘cellular-phone’. This is due to the fact that the original cellular-phones were always tied in the waists because these were a bit bigger in size. Coined words like this one are recorded in our dictionaries and in the official vocabularies but the communities keep on calling a cell-phone selfounu, which the purists do not like.

5. Conclusion

In conclusion, it is important to realize that, even though purism is important to guard against a contamination of our languages:

- It is this system which contributes highly to the poor growth of our indigenous languages because purism forms a barrier to development of the vocabularies of the languages.
It is this system which contributes much to ambiguity and the development of polysemy in the languages.

The excessive use of neologism leads to the creation of terms, i.e. ‘the Book Terminologies’, which are accredited by the language authorities and the language standardization, but rejected or ignored by the communities in the practical application, including even those scholars who coined these words, i.e. the purists.

The excessive use of neologism leads to the creation of a gap between a spoken and a written language.

It is this system which directs us to write what we don’t speak, while speaking what we don’t write.

It is this system which directs us to standardize what we don’t speak, while speaking what we don’t standardize.

References

Abstract

Beginning with the year 1989, both the German language and Upper Sorbian (a minority language spoken in Germany) have experienced significant lexical changes as the result of powerful social changes, viz., German unification and the resulting “invasion” of “Federal Republic of Germany” (FRG) terminology with a concomitant wealth of neologisms. Upper Sorbian (USo), although moribund, still enjoys vibrant usage in the (Catholic) area northwest of Budyšin (Bautzen). Apart from unification, a process that has still not been fully completed across all the new states of the former GDR due to inequality of certain benefits like pensions, there was the full onslaught of the bureaucracy of the European Union (EU), adding the incorporation of English-language terminology in the German language as well as in USo in all spheres of society. Thus, as compared with the relative speed in the post-1989 process of incorporating FRG terminology, the incorporation of EU terminology in USo is taking a lot longer, as it also has in many of the Southern member-countries of the EU (to wit the long time it takes for translating EU documents into the languages of these countries). As compared with similar trilingual situations in Asia (for Yakan-Filipino-English, see Pack & Behrens 1978), North America (for Comanche – Spanish – English, see Rejón & Gelo [1865], 1995; for Inuktitut – English – French, see Schaarschmidt 2001), and Central and South America (for Emberá – Spanish – English, see Sara 2001) it is clear that what is needed in this situation for USo in Germany is a trilingual dictionary if only because of the frequent reference in English-Upper Sorbian dictionaries to German definitions in the explication of English lemmata (Schaarschmidt 1991; 1994; 2001).

Keywords: German unification, FRG terminology, EU terminology, trilingual dictionaries

1. Introduction

Beginning with the year 1989, both the German language and Upper Sorbian (a minority language spoken in Germany) have experienced significant lexical changes as the result of powerful social changes, viz., German unification and the resulting “invasion” of “Federal Republic of Germany” (FRG) terminology with a concomitant wealth of neologisms. Upper Sorbian (USo), although moribund, still enjoys vibrant usage in the (Catholic) area northwest of Budyšin (Bautzen). Apart from unification, a process that has still not been fully completed across all the new states of the former GDR due to inequality of certain benefits like pensions, there was the full onslaught of the bureaucracy of the European Union (EU), adding the incorporation of English-
language terminology in the German language as well as in USo in all spheres of society. Thus, as compared with the relative speed in the post-1989 process of incorporating FRG terminology, the incorporation of EU terminology in USo is taking a lot longer, as it also has in many of the Southern member-countries of the EU (to wit the long time it takes for translating EU documents into the languages of these countries). As compared with similar trilingual situations in Asia (for Yakan-Filipino-English, see Pack & Behrens 1978), North America (for Comanche – Spanish – English, see Rejón & Gelo [1865], 1995; for Inuktitut – English – French, see Schaarschmidt 2001), and Central and South America (for Emberá – Spanish - English, see Sara 2001) it is clear that what is needed in this situation for USo in Germany is a trilingual dictionary if only because of the frequent reference in English–Upper Sorbian dictionaries to German definitions in the explication of English lemmata (Schaarschmidt 1991; 1994; 2001).

2. The sociolinguistic basis of trilingualism

It is probably fair to say that trilingualism presupposes a stage of bilingualism: probably the most common development is that a bilingual community has faced integration into a larger political unit with a dominating language. We shall briefly exemplify here four such situations on an international basis (1.1.–1.3. below), concentrating on Sorbian – German – English trilingualism (1.4.) as it emerged following the integration of the German Democratic Republic into a united Germany and thus directly into the European Union.

2.1 Comanche - Spanish - English

Comanche is an Uto-Aztecan language spoken by less than 100 persons (with the number of speakers decreasing) - the ethnic population residing in Western Oklahoma is 8,500; the language status is “moribund” (Ethnologue 2016); courses in Comanche are offered at the University of Oklahoma. The age of speakers is 50 and older. Spanish was the second language prior to the establishment of the Kiowa-Comanche-Apache Reservation in 1867.

2.2 Inuktitut - English/French

In April 1999, a new territory was carved out in the Canadian Eastern Arctic region to be known as Nunavut (composed of the noun nuna “land” and the postpositive possessive vut “our”). The language of Inuktitut is spoken by roughly 85% of the 25,000 inhabitants of the self-governing territory covering two million square kilometres (about one fifth of Canada’s land mass). The Inuit (a plural word with the singular being Inuk) are part of what is known as the Eskimo-Aleut language family, and apart from Canada, they reside in Greenland, Alaska, and Russia. At the moment, the two official languages of Nunavut are Inuktitut and English but it has become clear that since Canada is officially bilingual (English-French) for officials in Iqaluit (the capital of Nunavut) to communicate with officials in Ottawa, they will require a knowledge of French. There are plans, at least in the capital, Iqaluit, to open a school for teaching French as a third language. At the moment there is a great need for comprehensive bilingual and trilingual dictionaries.
2.3 Yakan - Filipino - English

Yakan (also spelled Yacan) is spoken by some 106,000 people in the Southern Philippines including 86,000 people in Basilan Province (Ethnologue 2016). It is classified as belonging to the Austronesian language family and is taught in primary schools in Basilan. There are no monolingual speakers with L2 consisting of English, Filipino, Chavacano and Tausug. Before English (American) acquired its present-day role in the Philippines, there was probably a stage of Yakan – Filipino - Spanish trilingualism since the Spanish government had established a fort on Basilan’s northwest coast (joshuaproject.net/people_groups/15956/RP).

2.4 Sorbian - German - English

Long before German colonization of the Eastern region began in the 12th century, there were several groups of Slavic speakers inhabiting these areas. The only surviving group of Slavic speakers today are the Sorbs, and according to all indications, their language is moribund, with only a relatively small enclave of mainly Catholic speakers showing vigorous language use. Up until the creation of the European Union and the unification of the two Germanys, the Sorbs were bilingual Sorbian-German speakers, although many of them also spoke the language of their eastern neighbour, Polish, or their southern neighbour, Czech. As English is the lingua franca of the EU, and has also managed to replace much of the administrative-scientific life of Germany, it has now become virtually a third language for the Sorbs. Most dictionary writers have not paid sufficient attention to this trilingual situation in the Sorbian area. But one can see in the case of one bilingual dictionary not only that the writer cannot ignore English but that he cannot ignore German either because in more than one instance the knowledge of German helps in understanding the English definitions, for example, in the entry scout = scout [člon młodożinskeje organizacije, ně: Pfadfinder; or the entry wrecker = rjubežnik, wosebje pri pribroze, ně: Strandräuber (see Strauch 1995 [ně stands for němski "German"); interestingly, such explanatory German glosses are absent from Strauch 2000; see also Schaarschmidt 2001, 42) . Thus, a Sorbian - German - English dictionary of the type proposed by us in this paper will fill a great need. The situation of Sorbian in the German Democratic Republic was distinct from that of minorities in other Slavic regions or countries in that the sociolinguistic situation of Sorbian must be seen in two contexts: 1) the “invasion” of a new German language (that of the Federal Republic of Germany after unity) as described in our study of a dictionary in progress during that time (Schaarschmidt 1991:69-70; 1994:571-76). The completion of this dictionary coincided with the time of the unification events in November 1989. The authors of the dictionary decided to include this sudden influx of FRG lexicology by adding an appendix at the end of the second volume. The reflection of the social changes was therefore very quick by the inclusion of items, such as Ossi, USo osi “inhabitant of the former GDR” Wessi, USo wesi “West German”; Aussteiger, USo disident “dissident”; Callgirl, USo callgirl; Bluejeans, USo bluejeans; Cheeseburger, USo cheeseburger or calta z twarožkom; recyclen, USo recyclovač “recycle”. Also, in deference to the uninhibited FRG lexicology, as opposed to the puritanic attitude of GDR lexicology, the appendix included words like Scheiß, USo njerjad “muck, feces”; Kontraceptivum, USo kontraceptivum “contraceptive”; and ironically bumsen, USo pjezać “copulate”, notwithstanding the fact that the more common German f-word is not contained in the body of the dictionary (but it is in the Duden, the official German guide
for lexical usage and spelling). The first official dictionary of the new lexicology appeared only seven years later (Jentsch, Pohontsch & Schulz 2006). 2) The massive onslaught of English from the European Union still requires a detailed description. Hopefully, this double onslaught of both new German and English lexical items on Sorbian, moribund as this minority language group has increasingly become in the past 30 years, will be addressed in our trilingual German-Sorbian-English dictionary, a sample of which is part of the present presentation.

3. The case for trilingual dictionaries

In a paper submitted for the Second International Congress of Asialex at Yonsei University, Seoul, Korea, August 8-10, 2001 (and published in Schaarschmidt 2001), we made a case for trilingual dictionaries in trilingual situations not covered by bilingual dictionaries. The clearest case for the need of a trilingual dictionary is probably one in which reference needs to be made to a third language in a bilingual dictionary in order to define more precisely what a given lemma means in one or both of the two languages covered. This can be seen, for example in an English-Sorbian dictionary where the unambiguous definition of concepts in English has to be given via the third language, i.e., German (see Strauch 1995; Schaarschmidt 2001:42).

What was not clear at the time was the format of such dictionaries concerning the directionality of listing lemmata from three languages. It seems that the basic format for trilingual editions of dictionaries was that followed in the English-Spanish-Comanche Vocabulary where synonymy was defined by using parenthetical inclusions of lemmata from the third language in curly brackets, for example, in their English-Spanish-Comanche Vocabulary: “thigh – {muslo} – puicap” (Rejón & Gelo 1865:47). It should be noted here that Rejón & Gelo may have anticipated the third necessary column as introduced by Solomon Sara (see below) because the section English-Spanish-Comanche was then supplemented by a separate Comanche-English vocabulary. In their Yakan-Pilipino-English dictionary, Pack & Behrens (1978:15) basically followed this procedure except that the definition of synonyms was indicated by parenthetical insertions in colons: “sanglag : binusá : roast”. This, to say the least, proved to be very cumbersome. Both procedures apparently aimed at the conflation of information in a single entry. However, both methods are in actual fact pseudo-inflations. The transition from a bilingual to a trilingual dictionary in trilingual situations seems to create certain problems in the definition of synonyms. The various trilingual dictionaries discussed here use various ad hoc devices, such as inserting synonyms in curly brackets or colons. The use of three columns in six possible linear orderings, while not permitting the explicit definitions of synonyms, allows their implicit definition by the process of the repetition of lemmata in the left-most column for each given variant, as demonstrated in the sample below from our Sorbian-German-English dictionary. The entire problem probably does not seem to exist in online dictionaries because linearity is not an issue there. The breakthrough seems to belong to Solomon Sara with his Emberá-Spanish-English trilingual dictionary (2001). Sara proceeded from the fact that to arrange three languages in a dictionary format there are potentially six possibilities of listing the languages but he chose merely three possibilities (Emberá is spoken in several variants in Panama and Colombia). We have adopted the same format for our Sorbian-German-English trilingual dictionary (in preparation). In such a format there is no problem with...
the pseudo-conflation of lexical items, and synonymy relations can be easily incorporated by multiple entries.

4. Conclusion

On the basis of several sociolinguistic situations outside the former Communist bloc, we have shown that the imposition of another dominating power in an already bifurcated society of a domineering two-power situation paves the way towards trilingualism. Our main focus has been the situation of Sorbian in the former German Democratic Republic after unification as well incorporation into the European. That this dual incorporation of societies gave rise to a very stressful situation found a reflection in a complicated lexical situation with a relative fast beginning (unification and the assimilation of FRG language) and a much slower reflection of the EU lexicon in both German and Upper Sorbian (with extinct Lower Sorbian showing surprisingly more flexibility in reflecting social change in lexical structure).

Another conclusion to this study might be that “old habits are slow to die”, even in the world of dictionaries. Thus, in spite of Sara’s (2001) clearly more adequate trilingual dictionary, the same publisher has come out fourteen years later with the cumbersome so-called trilingual dictionaries, for example, for Tula and Tera in Nigeria. These are in reality bilingual dictionaries with a separate index for a third language (as in Mu’azu & Polo 2015; and Mu’azu & Magaji 2015), i.e., in the fashion of Rejón & Gelo 1865 and Pack & Behrens 1978, albeit with grammatical and lexical-semantic information (synonymy) more in the way of encyclopedic dictionaries (Russian *толковые словари*). It has also become obvious that traditional bilingual dictionaries are too cumbersome in allowing a triple referencing in trilingual situations, especially as far as synonymy is concerned. Thus, for example, in the Mu’azu/Magaji (2015) dictionary, the method of listing English meanings and synonyms looks very much like Pack & Behrens (1978) or Rejón & Gelo ([1865], 1995):

<table>
<thead>
<tr>
<th>Tera</th>
<th>English</th>
<th>Hausa</th>
</tr>
</thead>
<tbody>
<tr>
<td>biri</td>
<td>(pillow)</td>
<td>matashin kai</td>
</tr>
<tr>
<td>boka</td>
<td>(sorcerer, soothsayer)</td>
<td>boka, matsafiya</td>
</tr>
</tbody>
</table>

This pseudo-conflation is clearly inferior to our model of the two items listed with two more possible reorderings, see also the longer excerpt from the trilingual Sorbian-German-English dictionary (in preparation).

<table>
<thead>
<tr>
<th>Tera</th>
<th>English</th>
<th>Hausa</th>
</tr>
</thead>
<tbody>
<tr>
<td>biri</td>
<td>pillow</td>
<td>matashin kai</td>
</tr>
<tr>
<td>boka</td>
<td>soothsayer</td>
<td>boka</td>
</tr>
<tr>
<td>boka</td>
<td>sorcerer</td>
<td>boka</td>
</tr>
<tr>
<td>boka</td>
<td>soothsayer</td>
<td>matsafiya</td>
</tr>
<tr>
<td>boka</td>
<td>sorcerer</td>
<td>matsafiya</td>
</tr>
</tbody>
</table>

A trilingual dictionary has more plasticity in that respect and is probably the only adequate means of giving equal status to the three languages involved.
References

APPENDIX: Sample of “TRILINGUAL UPPER SORBIAN - GERMAN – ENGLISH DICTIONARY”

CONSULTED SORBIAN DICTIONARIES (in chronological order):


THE THREE VARIANTS (SAMPLE):

Notes on Methodology and Lemmata Research:

1. Word groups are only listed if one of the three languages has a compound formation.
2. The reflexive so/sich is only given for verbs that are “absolute reflexives”.
3. Some individual forms that cannot be easily predicted from a base form are listed in the dictionary, such as pój “komm!/come!” (infinitive póńć) or dźi “geh!go!” (infinitive hić).
4. For Sorbian verbs, there are forms with different prefixes/suffixes that denote aspectual relations, such as the completion (C), duration (D), repetition (R), and instanteneity (I) of an action. All such verb forms are listed separately in this dictionary but are marked with the appropriate abbreviation, i.e., C, D, R, or I, i.e., VR = a verb meaning “a repeated action”.
5. Adverbs in Sorbian are derived from adjectives by the suffix –je. Therefore only irregular adverbs are given in the dictionary. Similarly, the comparative degree uses the suffix –išo, and only irregular derivations are listed.
6. The grammatical gender of nouns in German is not usually predictable from their
phonetic shape/ending, so all of them are marked with gender specifications [f, sg; m, sg; f, sg; or pl for pluraliatantum]. Where gender distinctions are marked by suffixes (usually in both Sorbian and German, the difference is marked by a slash (/) as well as by [m/f]). In addition, where the phonetic shape of nouns in Upper Sorbian does not allow the correct prediction of gender, a gender marker is added for those, e.g., USo *hobby [m, sg]. As there are quite a few nouns in Upper Sorbian that end in –a but are masculine, gender is marked for these as well. Some geographical names are morphologically adjectives in USo. Thus, when a geographical name has the form of an adjective (marked “ADJ”), it is declined like an adjective, e.g. Awstrijska ADJ “Austria” – w Awstrijskej “in Austria”; z Awstrijskeje “from Austria”.

7. Other abbreviations: A = adverb, ADJ = adjective, D = diminutive, V = verb, INT = interrogative; REL = relative; Pro = pronoun; PN = place name.

8. The listing of synonymous items is a problem in a trilingual dictionary. In many previous trilingual dictionaries, synonyms were included in parentheses. This problem is resolved here by restricting each column to single-word or single-group expressions and listing synonymous words or word-groups as separate lemmata, following the approach applied by Solomon Sara (2001). *A Tri-Lingual Dictionary of Emberá-Spanish-English. LINCOM EUROPA. Sara’s approach may be conveniently called the “implicit definition” of synonymy in lexical items.

9. For both the Upper Sorbian and English lemmata native speakers were consulted; for some German lemmata, native German speakers were consulted as well in cases where this author’s native competence was deemed to be somewhat out-of-date.

The Upper Sorbian alphabet (and approximate pronunciation)
(Letters in parentheses occur only in foreign words)

<table>
<thead>
<tr>
<th>A</th>
<th>a</th>
<th>G</th>
<th>g</th>
<th>N</th>
<th>ń [palatal]</th>
<th>U</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>b</td>
<td>H</td>
<td>h</td>
<td>O</td>
<td>o</td>
<td>(V</td>
<td>v</td>
</tr>
<tr>
<td>C</td>
<td>c</td>
<td>[ts]</td>
<td>Ch</td>
<td>ch</td>
<td>[kʰ] or [χ]</td>
<td>Ó</td>
<td>ó [ʊ]</td>
</tr>
<tr>
<td>Ć</td>
<td>č</td>
<td>[tʃ]</td>
<td>I</td>
<td>i</td>
<td></td>
<td>P</td>
<td>p</td>
</tr>
<tr>
<td>Ć</td>
<td>ć</td>
<td>[tʃ]</td>
<td>J</td>
<td>j</td>
<td>(Q</td>
<td>q</td>
<td>(Y</td>
</tr>
<tr>
<td>D</td>
<td>d</td>
<td>K</td>
<td>k</td>
<td>R</td>
<td>r [uvular]</td>
<td>Z</td>
<td>ż</td>
</tr>
<tr>
<td>Dź</td>
<td>dź</td>
<td>[dʒ]</td>
<td>L</td>
<td>l</td>
<td>[w]</td>
<td>Ź</td>
<td>ź [like French j]</td>
</tr>
<tr>
<td>E</td>
<td>e</td>
<td></td>
<td></td>
<td>S</td>
<td>s</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ė</td>
<td>Ė</td>
<td>[ˈɛ]</td>
<td>M</td>
<td>m</td>
<td>Ś</td>
<td>ś [sh]</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>f</td>
<td>N</td>
<td>n</td>
<td>T</td>
<td>t</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

I.

<table>
<thead>
<tr>
<th>Upper Sorbian</th>
<th>English</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>*charakterizować</td>
<td>characterize</td>
<td>charakterisieren</td>
</tr>
<tr>
<td>*charterowy/šarterowy</td>
<td>charter</td>
<td>Charter-</td>
</tr>
<tr>
<td>*chemija</td>
<td>chemistry</td>
<td>Chemie</td>
</tr>
<tr>
<td>*chemikalije</td>
<td>chemicals</td>
<td>Chemikalien</td>
</tr>
<tr>
<td>*chemikar</td>
<td>chemist</td>
<td>Chemiker</td>
</tr>
<tr>
<td>*chemiski</td>
<td>chemical</td>
<td>chemisch</td>
</tr>
<tr>
<td>*chětro</td>
<td>quite</td>
<td>ziemlich</td>
</tr>
<tr>
<td>*chěža</td>
<td>house</td>
<td>Haus</td>
</tr>
<tr>
<td>English</td>
<td>German</td>
<td>Upper Sorbian</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>---------------</td>
</tr>
<tr>
<td>brave</td>
<td>mutig</td>
<td>*chrobły</td>
</tr>
<tr>
<td>characterize</td>
<td>charakterisieren</td>
<td>*charakterizować</td>
</tr>
<tr>
<td>charter</td>
<td>Charter-</td>
<td>*charterowy/sarterowy</td>
</tr>
<tr>
<td>chemical</td>
<td>chemisch</td>
<td>*chemiski</td>
</tr>
<tr>
<td>chemicals</td>
<td>Chemikalien</td>
<td>*chemikalije</td>
</tr>
<tr>
<td>chemist</td>
<td>Chemiker</td>
<td>*chemikar</td>
</tr>
<tr>
<td>chemistry</td>
<td>Chemie</td>
<td>*chemija</td>
</tr>
<tr>
<td>chips</td>
<td>Chips</td>
<td>*chipsy</td>
</tr>
<tr>
<td>choir</td>
<td>Chor</td>
<td>*chór</td>
</tr>
<tr>
<td>Christ</td>
<td>Christus</td>
<td>*Chryst</td>
</tr>
<tr>
<td>Christus</td>
<td>Christus</td>
<td>*Chrystus</td>
</tr>
<tr>
<td>cottage</td>
<td>Häuschen</td>
<td>*chěžka</td>
</tr>
<tr>
<td>Cottbus</td>
<td>Cottbus</td>
<td>*Choćebuz</td>
</tr>
<tr>
<td>except</td>
<td>außer</td>
<td>*chiba</td>
</tr>
<tr>
<td>flag</td>
<td>Fahne</td>
<td>*chorhoj</td>
</tr>
<tr>
<td>greyhound</td>
<td>Windhund</td>
<td>*chort</td>
</tr>
<tr>
<td>hide</td>
<td>sich verstecken</td>
<td>*chować so</td>
</tr>
<tr>
<td>horseradish</td>
<td>Meerrettich</td>
<td>*chrêń</td>
</tr>
<tr>
<td>house</td>
<td>Haus</td>
<td>*chěža</td>
</tr>
<tr>
<td>ill</td>
<td>krank</td>
<td>*chory</td>
</tr>
<tr>
<td>illness</td>
<td>Krankheit</td>
<td>*chorosć</td>
</tr>
</tbody>
</table>
### III.

<table>
<thead>
<tr>
<th><strong>German</strong></th>
<th><strong>Upper Sorbian</strong></th>
<th><strong>English</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>außer</td>
<td><em>chiba</em></td>
<td>except</td>
</tr>
<tr>
<td>charakterisieren</td>
<td><em>charakterizować</em></td>
<td>characterize</td>
</tr>
<tr>
<td>Charter-</td>
<td><em>charterowy/šarterowy</em></td>
<td>charter</td>
</tr>
<tr>
<td>Chemie</td>
<td><em>chemija</em></td>
<td>chemistry</td>
</tr>
<tr>
<td>Chemikalien</td>
<td><em>chemikalije</em></td>
<td>chemicals</td>
</tr>
<tr>
<td>Chemiker</td>
<td><em>chemikar</em></td>
<td>chemist</td>
</tr>
<tr>
<td>chemisch</td>
<td><em>chemiski</em></td>
<td>chemical</td>
</tr>
<tr>
<td>Chips</td>
<td><em>chipsy</em></td>
<td>chips</td>
</tr>
<tr>
<td>Chor</td>
<td><em>chór</em></td>
<td>choir</td>
</tr>
<tr>
<td>Christus</td>
<td><em>Chryst</em></td>
<td>Christ</td>
</tr>
<tr>
<td>Christus</td>
<td><em>Chrystus</em></td>
<td>Christ</td>
</tr>
<tr>
<td>Cottbus</td>
<td><em>Choćebuz</em></td>
<td>Cottbus</td>
</tr>
<tr>
<td>Fahne</td>
<td><em>chorhoj</em></td>
<td>flag</td>
</tr>
<tr>
<td>Gehen</td>
<td><em>chodzenia</em></td>
<td>walking</td>
</tr>
<tr>
<td>Gehweg</td>
<td><em>chódnik</em></td>
<td>sidewalk</td>
</tr>
<tr>
<td>Haus</td>
<td><em>chėža</em></td>
<td>house</td>
</tr>
<tr>
<td>Häuschen</td>
<td><em>chěžka</em></td>
<td>cottage</td>
</tr>
<tr>
<td>Kiefer</td>
<td><em>chójna</em></td>
<td>pine</td>
</tr>
<tr>
<td>krank</td>
<td><em>chory</em></td>
<td>ill</td>
</tr>
<tr>
<td>krankhaft</td>
<td><em>chorobny</em></td>
<td>pathological</td>
</tr>
<tr>
<td>Krankheit</td>
<td><em>chorość</em></td>
<td>illness</td>
</tr>
<tr>
<td>Meerrettich</td>
<td><em>chrěn</em></td>
<td>horseradish</td>
</tr>
<tr>
<td>mutig</td>
<td><em>chrobły</em></td>
<td>brave</td>
</tr>
<tr>
<td>preserve</td>
<td><em>chować</em></td>
<td>keep</td>
</tr>
<tr>
<td>sich verstecken</td>
<td><em>chować so</em></td>
<td>hide</td>
</tr>
<tr>
<td>Stall</td>
<td><em>chlěw</em></td>
<td>stable</td>
</tr>
<tr>
<td>Wimpel</td>
<td><em>chorhojčka</em></td>
<td>pennant</td>
</tr>
<tr>
<td>Windhund</td>
<td><em>chort</em></td>
<td>greyhound</td>
</tr>
<tr>
<td>ziemlich</td>
<td><em>chětro</em></td>
<td>quite</td>
</tr>
</tbody>
</table>

* The Upper Sorbian examples have been marked with an asterisk (*) as they require a different program for sorting them alphabetically.
Linguistic Purity? Profiling Lettered Words in Chinese Dictionaries

Lan Li
The Hong Kong Polytechnic University
eglilan@gmail.com

Qi Gong
Jinan University
tgongq@jnu.edu.cn

Abstract

Language use in the digital era has led to the thriving of lettered words in non-romanized languages. Some linguistic purists have concerns for these unwanted alien elements, but using letters directly has become an irreversible process in vocabulary expansion. Lettered words have become a significant part of modern Chinese lexicon. The omnipresence of lettered words has also created a challenge to dictionary compilation. Although lettered words have entered Chinese dictionaries, but what are the selection criteria? How can they fit into macro- and micro-structures of dictionaries? This paper, from social, cultural, and linguistic perspectives, probes into a list of 1000 Chinese lettered words from a recently published Chinese dictionary, analysing their origins, word class, and uses in modern Chinese. The findings show that lettered words are formed by ten word formation approaches and can be used as verbs, adjectives and adverbs though nouns taking up 75% of the total. The lettered words are widely used in different genres. The limitations of the lettered word list are also discussed. Drawing on the FUDGE factor analysis by Metcalf (2002), the authors suggest that the inclusion of lettered words in a Chinese dictionary should have practical guidelines for selection, definition, pronunciation and usage. Frequency, unobtrusiveness, diversity of use and endurance of lettered words ought to be taken into consideration.

Keywords: lettered words, linguistic purity, etymology, classification, selection criteria

1. Introduction

All languages show the presence of different kinds of borrowings. Rapid globalisation and English as an international lingua franca in the digital era have lead to the thriving of lettered words in non-romanized languages. Acceptance of foreign elements, however, varies from language to language. Lettered words in Chinese refer to the words consisting of English letters without Chinese translation, such as GDP, MBAUSB. Some linguistic purists have concerns for these unwanted alien elements, but using letters directly seems to have become an irreversible process in vocabulary
expansion. Lettered words are an obvious part of modern Chinese lexicon today. Lu (2013) reported that about 4000 lettered words are used in printed media, e-media and spoken language in Chinese. The official Xiandai Hanyu Cidian (The Contemporary Chinese Dictionary by Commercial Press) is very cautious in recording lettered words into the Chinese language league. In its 6th edition published in 2012, only 254 lettered words were documented in the appendix. This however triggered or rejuvenised a debate on linguistic purity. A number of scholars blamed the editors for recording non-standard language use in a standard Chinese dictionary, which may spoil the unity and beauty of the Chinese language (Su, 2013). They suggested substituting loan words with the creation of new words from Chinese characters and preventinglettered words from entering the language, unless they are used in some specialised areas.

Linguistic purism is a sociolinguistic issue concerning national identity therefore it is also known as linguistic protectionism. It is the practice of defining or recognizing one variety of a language as being purer or of intrinsically higher quality than other varieties. As Thomas (1991:12) observes “Purism is the manifestation of a desire on the part of a speech community (or some section of it) to preserve a language from, or rid it of, putative foreign elements or other elements held to be undesirable (including those originating in dialects, sociolects and styles of the same language). It may be directed at all linguistic levels but primarily the lexicon. Above all, purism is an aspect of the codification, cultivation and planning of standard languages”.

Linguistic purism varies in goals, forms and intensity, and is mostly carried out through language academies. Thomas (1991) noticed that some languages, such as English, Russian and Japanese, are more open to all sources of enrichment, at the same time characterized by a lack of intellectual digestion of foreign influxes. Other languages keep purism as a constant value-feature of the speech community, for example French, Arabic, Tamil and Icelandic. Most languages keep a moderate attitude to linguistic purism however with a reactive correction to a potentially dangerous trend in the development of a standard language.

Linguistic purism in Chinese writing system can be discussed at two levels: interlingual and intralingual. The interlingual is direct bowing from romanised languages mainly from English; the intralingual derives within the Chinese language, using Pinyin variants instead of characters. The following table can illustrate source of lettered words.

<table>
<thead>
<tr>
<th></th>
<th>Interlingual</th>
<th>Intralingual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Putonghua (Putonghua)</td>
<td>GDP, GPA, QA, USB</td>
<td>HSK, GB, RMB</td>
</tr>
<tr>
<td>Cantonese (Putonghua)</td>
<td>GDP, GPA, QA, USB</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PC (Police Constable)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IP (Inspector of Police)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CE (Chief executive)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 Nature of lettered words in Chinese
Using lettered words directly in non-romanised language can speed up communication in modern society and avoid misunderstanding caused by translations. Loan words are now used directly in most languages. However, display of foreign elements in language varies. For example, romanised languages employ the same orthographic system for loan words. Non-romanised languages, if they use alphabets, can use transliteration to avoid foreign codes, such as:

Arabic i-Pad: ماي باد
MBA: ماجستير في إدارة الأعمال

Korean i-Pad: 아이패드; MBA”경영학석사

Hindi i-Pad: आई-पैड; MBA: एमबीए

Japanese i-Pad: アイパッド; MBA: MBA

Loan words in these cases do not seem to affect orthographic purism of these languages.

Using ideographic characters, the Chinese language is in a very different writing system. Untranslated loan words or lettered words when mixed with Chinese characters look alien, undermining the orthographic purism. Some Chinese purists believe that with over five thousand years of civilization, Chinese characters can be and should be used to explain anything new, and neologisms should be introduced after translation (Lu, 2013). These scholars are concerned that the use of lettered words in Chinese would encourage non-standard use of the language and may leave the writing system in disarray. They advocate that foreign elements be banned from standard media, and Chinese dictionaries should not have non-Chinese elements.

It cannot be denied that the omnipresence of lettered words is the result of globalisation reflecting a rapid and unavoidable language change. Fig. 1 shows concordance lines of a lettered word GDP (Gross Domestic Product) used in Chinese news reports.
Figure 1 GDP in LeedsU online Chinese corpus

The data is from a Chinese online corpus compiled by LeedsUniversity in the UK. In this 26 million-character data set, GDP is used more often than the expression 国内生产总值. Su and Wu (2013), based on another Chinese corpus, compared the frequency of lettered words and their Chinese translations in official China Daily from 1991 to 2004.

Figure 2 Comparison of GDP and 国内生产总值 in China Daily over 14 years

It is obvious that the use of lettered word GPA increased significantly after the year 2000. This is probably because of the impact of globalization and the increased level of English in China. Su and Wu (2013) also compared another word DNA. Deoxyribonucleic acid is a substance that carries unique genetic codes of living things. Although it has got a standardised Chinese translation 脱氧核糖核酸, corpus data shows that 90% of its use is in English abbreviation DNA, which is easy to say and to remember.
The above examples prove that the use of lettered word is a choice by Chinese writers, a natural process of language use; it cannot be controlled by any big names or institutions. However, it should not be abused, either. This paper, based on different data, attempts to profile lettered words in Chinese, including its typology, its semantic features and lexicographic practice.

2. Data used in the study

The data used in this study is from three sources: 1) the Chinese language reports (2009-2013) by the National Language Resource Monitoring and Research Center; 2) 26 million words of online Chinese Corpus by Leeds University in the UK (http://corpus.leeds.ac.uk/internet.html); and 3) A list of lettered word in Xiandai Hanyu Da Cidian compiled by Gong Xuesheng (2015) published by Commercial Press. The reason to choose this list is that it has 1001 lettered words, the longest list in the published Chinese dictionary up to date.

The National Language Resource Monitoring and Research Center was set up in 2003 under the Ministry of Education's Language and Information Management Division. The aim is to provide scholars with more quantitative data of the Chinese language and its contexts, to use modern technology analyzing and managing the national language resources dynamically, and to improve the influence of the Chinese language in the world. It announced Chinese media neologisms from 2009 to 2012 but only included eleven lettered words:

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of new word</th>
<th>No of loan word</th>
<th>%</th>
<th>No of lettered word</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>396</td>
<td>12</td>
<td>3%</td>
<td>3</td>
<td>70 码</td>
</tr>
<tr>
<td>2010</td>
<td>500</td>
<td>20</td>
<td>4%</td>
<td>6</td>
<td>Phone 时代, TA 时代, i 理财, IN 词, DNA 喷雾, PE 腐败</td>
</tr>
<tr>
<td>2011</td>
<td>593</td>
<td>15</td>
<td>2.5%</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>585</td>
<td>42</td>
<td>7.19%</td>
<td>2</td>
<td>VD 症候群, H 族</td>
</tr>
</tbody>
</table>

Figure 3 Comparison of DNA and 脱氧核糖核酸 in China Daily
To have a clearer picture of lettered words in Chinese, we explored the list compiled by Gong (2015) which has 1001 lettered and numbered words.

3. Findings and discussions

3.1 Origin of lettered words

As expected, the majority of the 1001 lettered words are originated from English (84%) such as USB, CNN, BBC, API. About 10% are code-mixed with English and Chinese, for example, K 金, PH 值, IT 界. Some lettered words are shortened from Chinese Pinyin, for example, HSK, BD, GB, MM; others are English abbreviations of Chinese items such as CET (College English Test) and CNBA (Chinese National Basketball Association). There are also a number of Latin and French words directly used in Chinese now.

Word class of lettered words

Part of speech of lettered words is another observation we made. It is interesting to notice that all the lettered words are not nouns; they take up only 80%. 10% of lettered words are verbs such as CC (carbon copy), PK (player kill), PS (photoshop) and BFN (Bye for now). There are also 20 adjectives such as top, yr(your) and BT (bian tai), and 25 adverbs as WE (whenever), TID(three times a day) and TTBOMK(to the best of my knowledge).TBS (tablespoon) and TPS are quantifiers and haha, hehe are onomatopoeia.

3.2 Formation of lettered words

The creation of lettered words in Chinese has many forms, similar to the ten word formation approaches in English, except derivation. They are shown in Table 2.

Table 3 Word formation approaches of lettered words

<table>
<thead>
<tr>
<th>Word formation</th>
<th>Frequency</th>
<th>% of the total</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbreviation</td>
<td>751</td>
<td>75</td>
<td>GPS, GPA, VIP, HSK, CET</td>
</tr>
<tr>
<td>Acronym</td>
<td>82</td>
<td>8.2</td>
<td>AIDS, ASCII, SOHO, CEPA</td>
</tr>
<tr>
<td>code-mixing</td>
<td>133</td>
<td>13.3</td>
<td>H 股, N 股, X 光, AA 制</td>
</tr>
<tr>
<td>Compound</td>
<td>49</td>
<td>5</td>
<td>chatroom</td>
</tr>
<tr>
<td>Shortening</td>
<td>81</td>
<td>8</td>
<td>app, sync</td>
</tr>
<tr>
<td>Blending</td>
<td>8</td>
<td>0.8</td>
<td>G20, G8</td>
</tr>
<tr>
<td>Conversion</td>
<td>9</td>
<td>0.9</td>
<td>Cc, PS, VS</td>
</tr>
<tr>
<td>Eponym</td>
<td>2</td>
<td>0.2</td>
<td>ICQ, MSN</td>
</tr>
</tbody>
</table>
Lettered words are found to be used in a variety of genres. As Metcalf (2002) noticed, the more active an area is, the more neologisms will be generated in that area. The most eminent domains with lettered words are science and technology, business, education sociology and online communication.

3.3 Genre of lettered word

The new words in modern technology are fresh to the whole human race. The vast advances of computers and wireless communication have changed the life of human beings dramatically in the last twenty years. Many of them are used in everyday life but have not obtained Chinese equivalents. As a result, lettered words referring to science and technology have taken up about 20% the 1000 list. AI (artificial intelligence), GPS (global positioning system), LAN (local area network) and CPU (central processing unit) are convenient terms for particular technology; USB (universal series bus), LED, MP3, PSP and iTouch represent ever-changing electronic gadgets; DNA, MRI (magnetic resonance imaging), CT (computerized tomography) and B超 (ultrasonic B-scan) are high-tech medical terminologies which have become part of everyday vernacular of the Chinese people.

A number of lettered words have emerged in business sector as the result of new ideas and new business transactions, such as B2B (business to business), C2C (consumer-to-consumer), A股 (RMB stocks in China), B股 (foreign currency stocks in China), H股 (Chinese stocks issued in the Hong Kong market) and N股 (Chinese stocks issued in New York). Many business job titles also borrow English abbreviations directly such as CEO (Chief Executive Officer), CFO (Chief Financial Officer), and CPA (Certified Public Accountant).

Socio-political terms have played an important role in modern languages. Politics is no doubt a very prosperous area in which language is fully utilized as an indispensable tool. The abbreviations or acronyms of many international organizations are used directly in the Chinese language, such as WTO, OPEC, WHO, BBC, CNN, even though they have Chinese translations or transliterations. Like undergoing an extremely active process of metabolism, language of politics witnesses plenty of words come and go. As everyone has noted, the post-9/11 world has exposed us to a barrage of new terms, for example, WMD (weapon of mass destruction). Many political words and expressions are employed in a creative way. As a matter of fact, new expressions of politics are often endowed with sheds of subtle meanings and strong color since their purpose is mainly either to defend or to attack. C4SR (Center for Spatial Research) and C3I (Communication, Command, Control and Intelligence systems) are two new systems employed in militarily service.

Lettered words have shown a shift from online media to traditional media (Sue, 2013). The users of online media are mainly young people with better English. Education is one of the central topics, therefore many lettered words are used, such as
GPA, TOEFL, IELTS, GRE and SAT. The digital generation is also very open and active in language use. 181 lettered words in the list are categorised as online expressions. The majority are borrowings from English, for example, BFN (bye for now), BTW (by the way), FU (for you) and FYA (for your action/amusement).

There are also some online Chinese pinyin abbreviations such as GG, MM, DD and BD. Some are used to express intimacy such as MM, GG, others are internet thick or coarse language, for example, TMD, BD, SB. Whether these colloquial terms can have a place in dictionary is debatable; the official Xiandai Hanyu Cidian (The Contemporary Chinese Dictionary) excludes them all. However one thing is certain that they are a part of the modern Chinese now. Some will stay if they fit into Chinese cultural and linguistic conventions, and others will die out if they violate general rules.

In the next section we are going to discuss the challenges of recording lettered words in Chinese dictionaries.

**Challenges to the compilation of Chinese dictionary**

The value of dictionaries lies in their contribution towards eliminating obstacles in communication because the average dictionary user regards a dictionary as the most comprehensive authority on linguistic information.

The list of lettered word in Gong’s Chinese dictionary (2015) has 1001 entries, providing us with a base for analysis. However the selection criteria does not seem to be clear; quite a number of words can be ruled out. Some entries are repetitive, for example, DVD and DVD 机, IT, IT 界, IT 人才, IT 时代. Other words have proper or standard Chinese translations; it seems unnecessary to enlist them in the lettered words. The examples are Ctrl 键 (控制键), Delete 键 (消除键), Enter 键 (输入键).

Hartmann (2001) regards a dictionary has three types of structures: macrostructure, microstructure and mediostructure. Macrostructure refers to the selection of words of a dictionary. With influx of lettered words, the decision on what words to include seems harder than ever before. To create a word list of lettered words, FUDGE factor analysis by Metcalf (2002) can be employed. FUDGE factors are Frequency of use, Unobtrusiveness, Diversity of users and situations, Generation of other forms and meanings and Endurance of the concept (Metcalf, 2002: 149-166). The origin of this system was from clinical medicine, in which Dr Virginia Apgar introduced a set of Apgar Score to evaluate the health condition of a newborn baby. Metcalf made great endeavor to interpret the mysterious and complex process of survival for the new words into a feasible evaluation mechanism to predict whether a word shall die or survive. In every factor, there are three levels; a word can gain zero, one and two marks respectively. The higher the mark is, the better the word should survive.
Take the word DINK for example. The acronym emerged in the late 1980s, meaning Double Incomes No Kids. As a new life-style in modern society, DINK families are surging in Western world, as well as in Hong Kong and major cities in mainland China. ‘Frequency of use’ in FUDGE indicates the more frequently a word is used, the more likely it will be kept in dictionary. A Google search of DINK has 1,050,000 hits. However, the figure may not be reliable because many of them are proper nouns such as names of person or institution. A search for DINK family can ensure the meaning we are looking for. With 527,000 hits, the frequency of DINK deserves a score 2.

The unobtrusiveness of DINK can be rated 2 because it is not strong-looking and is in a word family with ink, think, link, sink. The word DINK also has a high productivity to generate other forms and diversified meanings:

- **DINKY**  Double Incomes No Kids Yet
- **DINKUM**  Double Incomes No Kids Unbelievable Mortgage
- **DINKEM**  Double Incomes No Kids with Excessive Mortgage
- **DIOK**  Double Incomes One Kid
- **DINKWAD**  Double Incomes, No Kids, With A DOG

Diversity of users and situations in the FUDGE requires the word to be used by a variety of people and not be restricted in a specific situation. DINK can be rated 2 because the corpus search reveals that it is used in a variety of genres: spoken, fiction, magazines, newspapers and academic. Endurance of the concept means if the object, event or even notion the word based on is still alive and popular. DINK, as many sociologists predict, will last for quite a number of years in modern society. The word DINK, therefore, scores 10 in the FUDGE system, and now it is the entry in 34 dictionaries.

**Meaning of lettered words**

Like many other words, a lettered word may also have more than one meaning. The dictionary compiler has to decide which meanings to be recorded in a Chinese dictionary and what sense order to decide.

For example, the word GG in Baidu has 11 senses; some are from English and some originated in Chinese:

1. In computer game: GoodGame
2. Online expression: Gege (Chinese: brother)
3. Brand name: GuccioGucci
4. Pet: gou gou (Chinese, dog)
5. Gan ga (Chinese, embarrassed)
6. Googlenetizen
7. Online TV: Goldenglobe
8. Girls' Generation
9. Game interactive platform
10. American soap opera: Gossip Girl
11. Hand-held gaming device: Gamegear

Senses 7 to 11 have limited frequency and are not widely used, therefore they should not be included in a dictionary entry.

Many English abbreviations are polysemic. The online Acronymfinder shows that PS can have 245 meanings, NA has 101, APP 120, TMD 40, TSP 131 and TBS 142. Which meanings can entre into Chinese can also be assessed by FUDGE factors.

**Pronunciation**

Pronunciation is another challenge when lettered words enter into Chinese. Is the borrowing orthographic or phonetic? The dictionaries with lists of lettered word seem to have avoided this issue by providing meanings only. Xinshiji Xinciyu Dacidian by Kang and Liu (2016, Shanghai Lexicographic Press) collected 158 lettered and numbered words occured during the years 2000-2015. Phonetic information is provided by adding pinyin to Chinese characters in code-mixed words. For example, E 时代 [E shidai], DNA 指纹 [DNA zhiwen], S 股 [S gu] and N 婚 (N hun). The lettered words without phonetic information seem to have the assumption that all dictionary users understand English or at least know English alphabet, but this is not likely the case. Is it necessary to provide phonetic symbols for the letters? English or Chinese? Or both? The following words are some examples which may cause confusion:

- **MP3**  em pi:san / em pi:ɔri:
- **APP**  ai pu / ei pi:pi:
- **VS**  vi es / vɔ:sɔ s
- **GGdʒi:**  dʒi:/ gege
- **MM**  emem / meimei / mumu / mama?

**Mediostructure**

Mediostructure is cross-reference in a dictionary. It can be used to establish relations among different components of a dictionary and create textual cohesion. For example, ATM can have a cross-reference to the Chinese entry 自动柜员机 and IELTS can be redirected to雅思。Various structural components can interact by mediostructure. There is no doubt that lettered words in printed Chinese dictionaries will face problems of indexing. How to count the number of strokes? How to index them in the Radical
List? Where to arrange them in the Phonetic Guide? These questionas could be answered in electronic media.

4. Conclusions

There is no absolute pure language in the globalized world today, therefore it is not surprising to see more and more lettered words entre into the Chinese language. We should have an open mind to this language change. Using lettered words in Chinese is a natural process of modern communication and cannot be stopped by anyone. However, given the fact that Chinese has a different orthographic system, lettered words should be carefully observed, rather than abused.

It is lexicographers’ job to provide dictionary information of lettered words to Chinese users. The challenges are at macro-, micro and medio-levels. FUDGE factor analysis is a good guide for word selection. Each lettered word can also follow lexicographic conventions with a proper microstructure. ‘Although no dictionary would encompass all the information a dictionary user would need or like to have, the lexicographer should try to satisfy the average dictionary user’ (Zgusta, 1989).

References

社科院(2012)现代汉语词典。北京：商务印书馆。
亢世勇, 刘海润 (2016)新世纪新词语大词典。上海：上海辞书出版社。
龚学胜(2015)现代汉语大词典。北京：商务印书馆。
宋子然(2016)汉语新词新语大辞典。上海：上海辞书出版社。
陆俭明 (2013) “汉语危机”论可以休也。中国语文, 3/1, 80 – 82。
国家语言资源监测语料库年度新词语报告
http://www.cssn.cn/zt/zt_xkzt/xkzt_yyxzt/hyddxcyyj/
苏新春,吴晓芳 (2013) 字母词的生命力与局限性。北华大学学报,14/2,11-15。
The Quantitative Patterns of Lexical Neologisms: A Comparative-diachronic Corpus Approach

Tao Ma
Shanghai Sanda University
taoma@sandau.edu.cn

Abstract

This study examines the quantitative patterns of lexical neologisms to their dictionary inclusions by analyzing the time series of frequencies in diachronic corpora i.e. the WebCorp and the BCC, and the time intervals between the initial corpus occurrences and the final inclusions in dictionaries i.e. the Oxford English Dictionary and the CiHai dictionary. It is found that Chinese and English lexical neologisms have convergent patterns, based on which an implicit rejection criterion is revealed.

Keywords: Lexical neologism, Diachronic corpora, Dictionary inclusion.

1. Introduction

It is always a deep problem for dictionary editors to establish consistent criteria of rejecting or accepting a neologism for the divergence between the design of a dictionary and the expectation of its users, for the inconsistency between its forecasted needs and its future performance, and for the conflict between the idealism of its exhaustiveness and the limitation of its search effectiveness. In an era of big data and internet search engines, these tensions could have been resolved if stakeholders of dictionaries had reconciled with each other in terms of their descriptive or prescriptive philosophies, their liberal or conservative interests and their formal or functional perspectives to language nature and its use. Nevertheless, it is not an easier task to reach the reconciliation than to expand the paradigm of rejection or acceptance rules, either of which serves for the purpose of sifting neologisms, but the objectives of various dictionaries at different levels attribute to the diversification of the principles which guide lexicographic and lexicological works in practice.

Since attention drawn to detecting and extracting neologisms in the field of neology (Roche 1999, Stenetorp 2010, Cook 2011, Lau 2012, Torres del Rey 2014) is as much divided as their qualitative acceptance rules for the ad hoc purpose of a specific dictionary in the field of lexicography (Landau 2001, Breen 2004, 苏新春
2003, 陶炼2004, Atkins and Rundell 2008), there is few generally applicable frequency-based rules of neologism inclusion, which are hardly reliable and practical when dictionaries of different languages are compared.

Though the necessity to establish stringent criteria of neologism inclusion is as debatable as the feasibility of an encyclopedic effort to exhaust all neologisms at a synchronic state, either rejection or acceptance rule constitutes one side of the same coin for neologism entries in published dictionaries. Therefore, by looking into accepted neologisms, the patterns of rejections can be inversely revealed so that the implicit criteria of neologism inclusions with reference to frequency data apart from experts’ insights can be explicitly examined and outlined in a more systematic and applicable way on the basis of their quantitative properties, and this is the objective of this study.

2. Literature review

It is proposed that in previous studies, there are two perspectives to neologisms, namely lexicographic versus lexicological, based on which the selection criteria and inclusion mechanisms of neologisms vary considerably. From a lexicographic standpoint (Atkins and Rundell 2008), a neologism is the new combination of old lexical forms, which can be syntactic or morphological. At the syntactic level, a new combination is used but no morphological gap is filled, which attributes to the semantic expansion of lexis, or a semantic neologism. At the morphological level, a new lexical item is created and a morphological gap is filled, by which morphological expansion accompanies with semantic expansion of lexis, and consequently a lexical neologism appears.

The key task in the lexicographic analysis of neologisms is to determine whether the frequency is sufficiently high and necessarily consistent, in other words, the threshold of neologism inclusion in terms of the validity and reliability of corpus frequency before their dictionary listing. However, when the life-cycle of a neologism is tracked, the reoccurrence of most recorded neologisms is highly volatile (Renouf 2007, 2013). The simple reason behind this phenomenon of past-future conflict is that the baseline index of a neologism candidate reported in a corpus provides little diachronic information, so the prediction for its future use can be inaccurate and the window period of frequency confirmation is short due to the limit of a dictionary project. Therefore, the relatively low probability of frequency consistency reduces its dictionary searchable value, which undermines the uncertainty of lexicographic effort of neologism candidate selection and inclusion. In essence, a lexicographic perspective captures the supply-side of neologisms for formal purposes in observation, which is replicable and applicable in practice for dictionaries of different languages, but comparatively, it overlooks the demand-side of neologism inclusion from a functional
approach, by which a so-called lexicological perspective avails.

The main difference between lexicographic and lexicological perspectives lies in the treatment of neologisms in terms of their formal and functional aspects. The formal observation of use based on frequency data can rarely indicate the functional observation of usage and usefulness. In other words, listing words is lexicographical but hardly lexicological (Considine and Iamartino 2007), as exhaustiveness is not equivalent to utility. A lexicological perspective focuses on the semantic ambiguity of a neologism for dictionary explanation while a lexicographic perspective emphasizes findability and occurrences for dictionary record. A sufficiently high frequent neologism of low ambiguity does not necessarily require a dictionary inclusion, whereas a sufficiently ambiguous neologism of low frequency may not necessarily require a dictionary inclusion, either, because sufficiency does not entail necessity. So, a neologism candidate for dictionary inclusion has to satisfy two conditions at the same time: the degree of ambiguity and the level of frequency, based on which the selection of a neologism pertaining to the design of a dictionary varies. A lexicographic approach probes into the supply-side of neologism generation and the empirical basis of a neologism in the light of the frequency data from a corpus for the sufficiency of its inclusion, while a lexicological approach, on the contrary, deals with the demand-side of a neologism to be explained in a dictionary which constitutes the necessity of its inclusion.

In practice, previous research indicates that for large monolingual dictionaries, a systematic approach (O’Donovan and O’Neil 2008) is recommended for the selection and the inclusion of neologisms in terms of their longevity and usage by combining lexicographic and lexicological perspectives. The actual compiling process involves two stages which are based on the empirical evidence from corpus data and the judgments of experienced lexicographers. In the first detection stage, a web-based program runs on an algorithm for automatic new raw data collection from targeted sources e.g. newspapers, and it checks a reference list of lemmas for the unreferenced lemmas which are filtered and stored. The list of filtered lemmas will be tracked and updated on a monthly or quarterly basis and their time series of frequency data will be reported for experts’ selection. In the manual selection stage, a lexicographer makes the decision of neologism selection with reference to the historical data report of frequency and the current register and genre distribution of the neologism candidate.

The Chambers Dictionary (2008) used the systematic approach above and in actual process, about 600 neologisms are reported each day but most of them are discarded for the orthographic hardline of a dumb logic in checking spelling for neologisms, or the orthographic neologism (Janssen 2005). Noticeably, even though the frequency and distribution data of a neologism candidate formulate its key selection criteria, from a lexicographic perspective, a neologism is more than a single lemma. Therefore, the stringent criteria of automatic selection and detection are reliable but insufficiently valid for exhaustiveness (Creamer 1995). For the spelling
filtering algorithm to be valid, collocations and syntactic strings must be written into
the reference list in addition to a lemma stop list, which means the program must run
on the basis of big data computation.

Noticeably, in the editing process of the Contemporary Chinese Dictionary the 5th
edition (2005), a monolingual dictionary of standard Chinese language, multi-layered
criteria are adopted to select and include neologisms in addition to the frequency-based
criteria, namely event-based, meaning-based, structure-based and derivation-based (陶
炼 2004). The event-based criterion is a socio-linguistic measurement to determine the
degree of social influence by which a neologism candidate related to a socially
recognizable event is judged. The meaning-based criterion is used to decide whether a
neologism candidate has constructed a new semantic feature in the lexicon of a
language while the structure-based criterion is referred to determine whether a
neologism candidate has been placed in a new syntactic combination of the existing
lexemes. The derivation-based criterion is applied to judge whether a neologism
candidate is derived from the existing morphological or syntactic structure. Two
criteria out of four mentioned above have to be met at the same time for the neologism
candidate to be included in the dictionary.

In fact, the multi-layered criteria are the modified version of composite definitions
for the classification of neologisms, that is: semantic versus lexical. The definition of a
semantic neologism identifies the change in the existing lexemes’ collocation and
co-text while the perspective to a lexical neologism focuses on the initial time of
occurrence in a diachronic corpus (Renouf 2013). To certain extent, the data source for
the classification and verification of neologisms to their definitions is largely built on
the proper methods of extracting information from diachronic corpora while the valid
measurable data of social influence from an event can hardly be collected other than
the frequency of the event-related lexemes, so the multi-layered criteria test ingret
returns to a frequency-based approach to neologisms again and the perspective to the same
issue of diffusion and consistency is changed. Therefore, the legitimacy of a neologism
inclusion and its proof is largely provided by the treatment of its frequency data by a
quantitative analysis.

3. Research Design and Methodology

This study examines the quantitative patterns of lexical neologisms to their dictionary
inclusion. It is hypothesized that there are unwritten rules in the dictionary inclusion of
lexical neologisms, and this study intends to reveal the rules by asking a research
question that whether there are statistically significant patterns of lexical neologisms to
their dictionary inclusions.

It measures two variables i.e. the time series of frequencies in diachronic corpora
and the time interval between the first corpus record and the final dictionary inclusion.
It uses a Chinese dictionary and an English one to randomly select the target lexical neologisms for measurement, and it collects data from a Chinese diachronic corpus and an English one so as to answer the research question and compare the quantitative patterns between different dictionaries and between different languages.

The two dictionaries referred to here are the Oxford English Dictionary and its Chinese counterpart, the Cihai Dictionary, for both of them being encyclopedia dictionaries to contain a substantial number of neologisms in principle. Importantly, the year 1989 when the OED 2nd edition was published coincides with the year when 5th edition of the Cihai Dictionary was also released, and this gives the starting point for observation. While the OED began to release its new word lists annually on the internet since 2000, the Cihai published its 6th edition in 2009 which constitutes the cutting point of observation for dictionary inclusion. Therefore, the window period for collecting pre-inclusion data of target lexical neologisms is from 1989 to 2009, and for the post-inclusion data from 2010 to the present.

The cumulative release of new words from the OED online will be the source of English lexical neologisms, while the entry difference between the two editions of the Cihai dictionary over the time period will be the source of Chinese lexical neologisms. This study will randomly select 100 target candidates from each of the English and Chinese sources among all new words included in the two dictionaries. The definition of a lexical neologism follows a technical criterion as follows: ‘a lexical neologism is usefully often a lexical item which occurs for the first time in a diachronic corpus’ (Renouf 2013:178). Therefore, the 100 target candidates will be checked in diachronic corpora for their unattestedness to determine the legitimate status of being a lexical neologism. If the technical criterion is not satisfied, a new round of collection and checking will be initiated to fill the potential gaps left by the previous unsatisfied candidates.

The two diachronic corpora used in this study are the WebCorp for English and the BCC for Chinese. WebCorp is a web-based English corpus being able to record and report diachronic frequency information (Renouf 2002), while the BCC is a Chinese on-line corpus having the same functions (荀恩东 2016). Both corpora mainly comprised of newspaper data which is the ideal source of lexical neologisms. After the 100 English and Chinese lexical neologisms are selected and determined, their frequency data over the time period will be retrieved from the two corpora respectively, and there will be two types of quantitative analyses.

The time series data of frequencies are collected on an annual basis in two periods: before and after the year 2009 which is the cutting point of dictionary inclusion. There will be two types of regression analyses for the pre-inclusion frequency data and the post-inclusion frequency data respectively. Firstly, a linear least-squares approach will be applied to the time series data of each lexical neologism to estimate the coefficient of the regression equation, and their regression coefficient data will be compared in the same coefficient of determination group. Secondly, a multi-variant statistical approach
will be applied to the frequency data of different lexical neologisms with the same time interval of corpus records to generalize their predictive patterns of frequencies to dictionary inclusions.

The data of time intervals between the first corpus record and the dictionary inclusion will also be analyzed in two ways. Firstly, the mean time intervals will be calculated in three categories: the language type i.e. English versus Chinese, the coefficient of the least squares i.e. negative versus positive and the cumulative frequencies i.e. high versus low. Secondly, the standard deviation will be calculated in the same categories to summarize their changing patterns.

4. Findings

It is calculated that there are more than twelve thousand new entries included in the OED at the end of 2009 since their online release in 2000, and it is reported that, coincidentally, there are more than twelve thousand new entries in the Cihai 6th edition, compared with its 5th edition (巢峰 2010), despite the fact that, interestingly enough to mention, the number of entries deducted from the previous edition is around seven thousand, which makes the net increase of entries only around four thousand in record.

The one hundred neologism candidates are randomly selected in an alphabetic order with an interval of one hundred entries. Their corpus attestedness before 1989 is checked to confirm their lexical neologism status in accordance with the selection criteria of this study. There are five attested candidates in Chinese and one in English, which trigger a second round of selection.

In the pre-inclusion set, the least squares analysis indicates that 64 Chinese lexical neologisms have positive regression coefficients from 0.9 to 6.5 with the coefficients of determination from 0.44 to 0.84, while the rest have negative regression coefficients from -2.8 to -0.43 with a similar range of the coefficients of determination. 81 English lexical neologisms have positive regression coefficients from 0.69 to 5.8 with the coefficients of determination from 0.42 to 0.85 while the rest have negative coefficients from -3.2 to -0.33 with a similar range of the coefficients of determination.

In the post-inclusion set, the least squares analysis indicates that 82 Chinese lexical neologisms have positive regression coefficients from 0.8 to 3.5 with the coefficients of determination from 0.54 to 0.87, while the rest have negative coefficients from -2.8 to -0.43 with a similar range of the coefficients of determination. 87 English lexical neologisms have positive regression coefficients from 1.7 to 2.8 with the coefficients of determination from 0.53 to 0.88 while the rest have negative coefficients from -2.5 to -0.67 with a similar range of the coefficients of determination.

A positive regression coefficient indicates the increasing frequencies of a lexical neologism over the time period of the diachronic corpus data, which suggests that the target lexical neologism has an increasing trend of use and coverage, while a negative
regression coefficient denotes decreasing frequencies, which suggests the opposite trend. A high coefficient of determination indicates a high consistency of frequency development over the time period of the diachronic corpus data, while a low coefficient indicates the opposite. In this study, the coefficient of determination above 0.7 is marked as high consistency while below 0.7 as low consistency abbreviated as Cst. in the table below.

Table 1 A least squares analysis of the diachronic frequencies

<table>
<thead>
<tr>
<th>Least squares analysis</th>
<th>Increasing trend</th>
<th>Decreasing trend</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High Cst.</td>
<td>Low Cst.</td>
</tr>
<tr>
<td>Pre-inclusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>33</td>
<td>48</td>
</tr>
<tr>
<td>Chinese</td>
<td>31</td>
<td>33</td>
</tr>
<tr>
<td>Post-inclusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>51</td>
<td>36</td>
</tr>
<tr>
<td>Chinese</td>
<td>42</td>
<td>40</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>48</td>
<td>41</td>
</tr>
<tr>
<td>Chinese</td>
<td>47</td>
<td>43</td>
</tr>
</tbody>
</table>

When data from the pre-inclusion and the post-inclusion periods are combined together for a least squares analysis, 90 Chinese lexical neologisms and 89 English lexical neologisms have an increasing trend of frequencies. It can be easily noticed that regardless of language difference, the number of lexical neologisms with increasing frequency trends is substantially higher than the number of those with decreasing trends, which means that most lexical neologisms in dictionaries maintain a substantial degree of diffusion. However, individual lexical neologisms vary greatly among each other in terms of their frequency development, indicated by a large proportion of lexical neologisms with low consistencies suggested by the low coefficients of determination, which reflects the volatile nature of lexical neologisms due to their initial occurrences in language use and coverage.

In another dimension of data analysis, when all frequency data of the 100 lexical neologisms are combined together in a single Cartesian coordinate, a multi-variant analysis indicates a general trend of frequency movement, shown in the table below. The mean in the table indicates the annual mean frequency of lexical neologisms per
million and the SD indicates the standard deviation of the annual mean frequency.

**Table 2** A multi-variant analysis of the diachronic frequencies

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>$R^2$</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pre-inclusion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>2.8</td>
<td>0.72</td>
<td>10.3</td>
<td>6.3</td>
</tr>
<tr>
<td>Chinese</td>
<td>2.5</td>
<td>0.75</td>
<td>9.6</td>
<td>4.2</td>
</tr>
<tr>
<td><strong>Post-inclusion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>3.1</td>
<td>0.86</td>
<td>11.2</td>
<td>5.8</td>
</tr>
<tr>
<td>Chinese</td>
<td>2.9</td>
<td>0.84</td>
<td>10.5</td>
<td>5.3</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>2.9</td>
<td>0.82</td>
<td>10.3</td>
<td>5.4</td>
</tr>
<tr>
<td>Chinese</td>
<td>2.7</td>
<td>0.79</td>
<td>9.8</td>
<td>4.8</td>
</tr>
</tbody>
</table>

There is a generally increasing trend of diachronic corpus frequency data over the observation period from the initial corpus record to the present time, which suggests the constant diffusion of neologisms in terms of their use and coverage. A comparatively higher coefficient of determination in the post-inclusion period than the one in the pre-inclusion suggests a gradually consistent trend of frequency development after the dictionary inclusion. That the annual mean frequencies are relatively similar across different categories of time intervals and languages suggests the convergent patterns of lexical neologisms to dictionary inclusions despite that they are not completely statistically significant.

However, the fact that there are similar annual mean frequency data before and after dictionary inclusion implies the developmental patterning of a lexical neologism to be a listed lexeme in the lexicon of a language. This also helps to answer how often a lexical neologism is used with reference to its dictionary inclusion, which suggests a sufficient condition for a frequency-based rejection criterion of lexical neologisms from a dictionary. When the annual mean frequency of a neologism candidate is no more than 5 per million in a corpus, which is derived from the table above by applying the calculation method that the mean of English and Chinese annual mean frequencies is subtracted by their mean standard deviation, it is highly likely that the neologism candidate will be rejected from a dictionary, despite the fact that frequency is not a necessary condition for the dictionary inclusion of a lexical neologism when the time intervals between corpus record and dictionary inclusion is referred to.
Table 3 Mean and Standard Deviation of the time intervals (Years)

<table>
<thead>
<tr>
<th></th>
<th>Positive Coefficient</th>
<th>Negative Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>English</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Freq.</td>
<td>7.7</td>
<td>3.3</td>
</tr>
<tr>
<td>Low Freq.</td>
<td>9.3</td>
<td>3.9</td>
</tr>
<tr>
<td>Unbiased</td>
<td>8.4</td>
<td>3.8</td>
</tr>
<tr>
<td>Chinese</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Freq.</td>
<td>8.4</td>
<td>3.2</td>
</tr>
<tr>
<td>Low Freq.</td>
<td>9.5</td>
<td>3.8</td>
</tr>
<tr>
<td>Unbiased</td>
<td>8.7</td>
<td>3.5</td>
</tr>
</tbody>
</table>

The time intervals are categorized in terms of frequencies and coefficients, and the cumulative frequency above 100 per million is marked as high frequency and the rest as low frequency abbreviated as Frq. in the table above. The table 3 shows that the mean time intervals of increasing frequency trend indicated by positive regression coefficients are slightly shorter than the ones of decreasing frequency trend indicated by negative regression coefficients, which suggests that more diffusive lexical neologisms tend to have earlier dictionary inclusions. Meanwhile, the standard deviations follow the same pattern in terms of frequency development direction, which suggests that less diffusive lexical neologisms tend to vary substantially in terms of their timing of dictionary inclusions.

Noticeably, the time intervals of high frequency are generally shorter than the ones of low frequency in both languages, which suggests that the more frequent lexical neologisms tend to have earlier dictionary inclusions than the less frequent ones. The unbiased mean and standard deviation of the time intervals before the dictionary inclusion is 8.25 years and 3.77 years for English and 8.62 years and 3.43 for Chinese. Therefore, the sufficient condition of dictionary inclusion in terms of time intervals is 5 years by applying the same calculation method above. When a lexical neologism has no more than 5 years corpus presence, it is highly likely that the lexical neologism will be rejected from a dictionary.
Figure 1 The mean annual frequency and the mean time interval of lexical neologisms

The Pearson coefficient of the correlation between annual frequencies and time intervals is 0.87 in English and 0.84 in Chinese, which suggests the connections between diffusiveness and consistency of lexical neologisms. The validity of the observation and the analysis method in this study is partially confirmed by the correlation test which indicates that the more frequently a lexical neologism is used, the less time it will consume before its dictionary inclusion. Therefore, these findings reveal two sufficient conditions for the dictionary inclusion of lexical neologisms: sufficient frequency and sufficient time. The similarities of frequency and time between English and Chinese suggest the convergent patterns of lexical neologisms to their dictionary inclusions.

5. Discussion

There are many debates over different criteria and principles pertaining to the dictionary inclusions of lexical neologisms when different approaches and perspectives to lexicography and neology are considered. However, there are convergent patterns found in this study for different lexical neologisms between English and Chinese for the same type of dictionary, despite the fact that there are substantial variations in terms of diachronic corpus frequencies and time intervals of individual lexical neologism. The convergent minimum requirement that the lexical neologisms investigated here have the mean annual frequency above 5 per million and the mean time interval above 5 years suggests an implicit rejection criterion of frequency and time sufficiency for the dictionary inclusion of lexical neologisms.

It should be acknowledged that the implicit criterion revealed here may not be
universally applicable to other dictionaries but it should be maintained as an internal consistent criterion for neologism selection and inclusion in the same dictionary. Although the register and genre variations of lexical neologisms are not examined in this study, it is believed here that the single sufficient condition of frequencies over substantial length of time constitutes the demand for dictionary explanation even in a single register or genre, which means that the neologism candidate with a frequency above certain level over certain amount of time should be included in the dictionary regardless of other factors. So, it may be problematic for the OED or the Cihai not to include a potential neologism candidate which satisfies the minimum criterion of frequency and time. This opinion echoes the idea from a prestigious Chinese linguistic expert Shuxiang Lu (1984) who once commented on the dictionary inclusion of neologisms. ‘It is more acceptable to commit an error in the dictionary inclusion of neologisms by loose criteria than by stringent criteria’; originally in Chinese, it is written as ‘与其失之于严，无宁失之于宽’ (吕叔湘 1984:14). Therefore, the 5 over 5 rejection criterion of frequency over time should not be regarded as a loose criterion when the decision of lexical neologism inclusion is to be made.

It is interesting to notice a phenomenon that there are negative criticisms over the decisions by the OED to include the heart-shaped symbol ♥ as in the sentence ‘I ♥ NYC’ and the internet jargons such as LOL and OMG (Petri 2011), the reasoning behind the criticism being that the OED should promote standard English language. However, it may be a more reasonable argument if the OED is renamed as a learner’s dictionary. These neologisms, or new linguistic forms, with sufficient communicative functions, should be included in dictionaries not only for the motivation of exhaustiveness from a supply-side perspective of dictionary listing but also the simple validity of search effectiveness from a demand-side perspective of dictionary usefulness. The rejection of these neologisms which satisfy the frequency over time criterion is illegitimate from the technical ground of lexicography and neology and also undemocratic from the user-based ground of information symmetry and knowledge economy.

It is also interesting to notice that the Cihai dictionary’s rejection of ‘超女’, an event-related neologism whose corpus frequency is above the threshold found in this study, translated as ‘Super Girl’, the Chinese version of a singing program equivalent to ‘Eurovision’ or ‘American Idol’, arouses a deeply divided debate over not only the inclusiveness of the Cihai but also the attitude to social establishment and its change. Whether a dictionary should be impartial to social opinions is as unsolvable as whether there is a single criterion for the dictionary inclusion of neologisms, but the initiative of Oxford Global Languages (Pearsall 2016) should be unquestionably applauded when the compiling and editing principles of a dictionary are derived from the idealism to promote people of other languages to use the language from the dictionary. The idealistic principles may attribute to the ‘loosest’ criterion for the dictionary inclusion of neologisms as the ‘error’ to be committed by its editors is the most oblivious for its
users who should be the truest stakeholders of a dictionary.

6. Conclusion

In this study, the quantitative patterns of lexical neologisms from the OED and the Cihai are examined in terms of their diachronic frequencies and time intervals based on the data from the WebCorp and the BCC. It is found that there are convergent patterns of lexical neologisms to their dictionary inclusions, but some of the lexical neologisms are not statistically significant in terms of their frequency and time data, therefore, the hypothesis of this study cannot be fully accepted that there are unwritten rules of dictionary inclusion. This is due to the fact that some lexical neologisms are not treated by a common stringent criterion of a frequency-based rule which is easily applicable but unequally practical.

The lenient criterion proposed here is based on the convergent pattern found in the study that most lexical neologisms have a least annual mean frequency of 5 per million and a least mean time interval of 5 years between the initial corpus occurrence and final dictionary inclusion. As many a lexical neologism rejected by dictionaries satisfies the criterion and indicates similar patterns, it is proposed that an explicit quantitative criterion e.g. a diachronic corpus frequency-time ratio, to accept or reject neologism entries, apart from implicit expert suggestions of specific dictionary design.is a good ‘error’ (Lu 1984) to initiate and promote the idealism of an encyclopedic global dictionary so that the uncritical acceptance of lexical neologisms will be a standard practice.

References

巢峰, 张敏. 2017 《辞海》是怎样炼成的 ——写在《辞海》出版80周年之际. 中华读书报. 2017年01月04日第9版.
巢峰. 2010《辞海》(第六版)编纂出版工作总结. 辞书研究.112-134.
万茹, 曹炜 2009《现代汉语词典》第5版用例修订计数考察——兼论《现代汉语词典》第5版用例修订的特点. 语言研究 (2) :98-103.
荀恩东, 饶高琦, 肖晓悦, 臧娇娇. 2016. 大数据背景下BCC语料库的研制, 语料库语言学(1).


Falk, I., D. Bernhard and C. Gérard. 2014. From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers. LREC.


Stenetorp, P. 2010. Automated Extraction of Swedish Neologisms using a Temporally


Technical Words Boundary for General Purpose Dictionaries: A Grounded Theory *

Ferdi Bozkurt
Anadolu University, Turkey
ferdib@anadolu.edu.tr

Abstract

Recently, users of general purpose dictionary can face with technical words in various texts and medium. There are many automotive-related technical words that a person whose car has broken down may encounter when s/he opens the user guide or when s/he talks on the phone for road assistance. Or, there are many technical words that ordinary computer users may encounter about computers. Where can they learn the definition or explanations of these words? Where must they find these words?

It is very difficult to determine the limit of the technical words to be included in the general dictionaries. In the past, some compilers of general purpose dictionary were limiting the technical words according to their personal thoughts and intuition. What should be the limit of the technical words to be taken in general purpose dictionaries? What should this limit be contained the specialized fields and the general language? These questions and problems are related to many general purpose dictionaries.

This study is a qualitative research and it has been constructed according to the theoretical design pattern from research design of grounded theory. The aim of this study is to describe how the general purpose dictionary compilers in Europe determine the technical words boundary and to consider how technical words can be delimited when taken in general-purpose dictionaries. In this study, a model will be made about how to make the technical words boundary for general purpose dictionaries by going out from the findings obtained with semi-structured qualitative interview questions was asked from compilers of European general-purpose dictionaries.

Keywords: technical words, general purpose dictionaries, qualitative research

* This study derived from Bozkurt, Ferdi (2016). “Lexeme Selection for General Purpose Dictionaries” Ph.D. dissertation.
1. Introduction

Hartmann and James (1998: 61) pointed out the functions of general dictionaries at the beginning of the dictionary article "general dictionary", and the meaning, knowledge, writing, idiomatic usage, etc. He emphasizes knowledge and suggests that it is a reference work aimed at providing a comprehensive picture of the whole language in a way that takes into account the existence of the word.

Hartmann and James (1998: 13) also point out that the general dictionary covers the core words of the language. At this point, he does not give information about what should be the core words of a language that general dictionary will involve. Jackson (2002: 26) also notes that general dictionaries such as Hartmann and James tend to present a list of words in a language to cover the core vocabulary of a language. These dictionaries are separated from each other according to their technical, expertise, daily speech-language, slang, and dialect vocabulary. Editors will be concerned to be up-to-date, especially in socially and culturally significant areas such as computing, medicine, the environment, fashion, and so on.

Lexical items belonging to regional and social varieties of a language, as well as specialized terminology, are thus exempt from the word list of a general dictionary, whereas dialect dictionaries and terminological dictionaries that specifically deal with multi- or single-field terminology are regarded as restricted lexicographic publications.

The boundaries of the word lists of general dictionaries are broader than the specialist dictionaries and it is a more difficult task to determine this boundary than the specialist dictionary Béjoint (2000), Swanepoel (2003), Burkhanov (1998).

What should be the term boundary for general dictionaries? The main question is in this study. Can ordinary dictionary user find a term which is used on specific biological text?

2. Research questions

The present study addresses the following questions:

How general dictionary editors decide the technical words boundary for their dictionaries?

Do they have standard decision model for technical words boundary?

3. Participants

There are three kinds of participants in this study.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxford English Dictionary</td>
<td>English</td>
<td>England</td>
<td>Deputy editor</td>
</tr>
<tr>
<td>II Devoto-Oli Vocabolario Della Lingua Italiana</td>
<td>Italian</td>
<td>Italy</td>
<td>Editor</td>
</tr>
<tr>
<td>Słownik Języka Polskiego</td>
<td>Polish</td>
<td>Poland</td>
<td>Editor</td>
</tr>
<tr>
<td>Kielitoimiston Sanakirja</td>
<td>Finnish</td>
<td>Finland</td>
<td>Editor</td>
</tr>
<tr>
<td>Id-Dizzjunarju Malti u Teżawru ta’ Malti Mhaddem</td>
<td>Maltese</td>
<td>Malta</td>
<td>Editor</td>
</tr>
<tr>
<td>Slovar Slovenskega Knjižnega Jezika</td>
<td>Slovene</td>
<td>Slovenia</td>
<td>Co-editor</td>
</tr>
<tr>
<td>Deutsche Wörterbücher und Grammatik</td>
<td>German</td>
<td>Germany</td>
<td>Editor</td>
</tr>
<tr>
<td>Svenska Akademiens Ordbok</td>
<td>Swedish</td>
<td>Sweden</td>
<td>Editor</td>
</tr>
<tr>
<td>Norsk Ordbok</td>
<td>Norwegian</td>
<td>Norway</td>
<td>Editor in chief</td>
</tr>
<tr>
<td>A Magyar Nyelv Nagyszótára</td>
<td>Hungarian</td>
<td>Hungary</td>
<td>Editor</td>
</tr>
</tbody>
</table>
As you can see in Table 1 there are 13 participants answered questions via e-mail.

**Table 2** Face to face participants

<table>
<thead>
<tr>
<th>The Name of the Dictionary</th>
<th>Language of Dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ötüken Türkçe Sözlük</td>
<td>Turkish</td>
</tr>
<tr>
<td>Misalli Büyük Türkçe Sözlük</td>
<td>Turkish</td>
</tr>
<tr>
<td>Güzel Türkçenin Sözlüğü</td>
<td>Turkish</td>
</tr>
<tr>
<td>Dil Derneği: Türkçe Sözlük</td>
<td>Turkish</td>
</tr>
<tr>
<td>Türk Dil Kurumu: Türkçe Sözlük</td>
<td>Turkish</td>
</tr>
</tbody>
</table>

As you can see in Table 2 there are 5 participants answered questions face-to-face.
### Table 3 Via Skype participants

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Macmillan English Dictionary</td>
<td>English</td>
<td>England</td>
<td>Editor</td>
</tr>
<tr>
<td>The Longman Dictionary of Contemporary English</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As you can see in Table 3 there are there is 1 participant answered questions via skype.

### 4. Data Collecting Instruments

For this study, one semi-structured questions asked to all participants. The question is:

“What do you think the criteria should be in including words based on written language, spoken language, and specialized language, i.e. technical terms? How do you determine the limitation of written language, spoken language and specialized language, i.e. technical terms?”

### 5. Method

This study will be qualitative research. By the term “qualitative research,” we mean any type of research that produces findings not arrived at by statistical procedures or other (Straus ve Corbin, 1998). The strategy of this study will be grounded theory. Grounded theory is a strategy of inquiry in which the researcher derives a general, abstract theory of a process, action, or interaction grounded in the views of participants Creswell, 2009).
6. Results

The eighteen of the nineteen participants reported various opinions about the technical language and the general language boundary. It is necessary to decide which technical vocabulary will be included in the general dictionary and to establish a definite line between the technical language and the general language limit. The eighteen participants were in the same opinion that the necessity of this boundary is relevant. The Participant 9 emphasizes this boundary, as an example:

“Considering the dissemination in the real use.” (Participant 9)

It is a necessity determines this boundary, but it is so difficult task. Participant 6 and participant 13 emphasize this difficulty:

“I am convinced that there are no exact criteria for clear-cut distinctions between these different types of sub-languages. We do not apply such criteria because, unfortunately, we are not able to provide pragmatic information.” (Participant 6)

“... It's really hard to decide because there are so many different areas, there are sub-domains.” (Participant 13)

Participant 1 and Participant 18 expresses that they do not include specialist terms in their dictionaries, but they include those used in the general language.

“We do not include spoken language or specialised language (only if at least one meaning is common in the general language) in the ...” (Participant 1)

However, at this point, both participants do not provide information on how to decide the language boundary limit with the general language.

“...Usual dictionary user should understand a vocabulary of general dictionaries. For example, everybody can understand gear, almost everyone knows headlight, but the radiator is a specialist word. Everyone may not know this word.” (Participant 18)

Participant 16 specifies that the determination of the general language boundary of the technical language is based on the knowledge, ability, and competence of the person who is compiling the dictionary, or in other words, to determine this limit based on lexicographer’s own intuition.

“Technical Terms should be taken in a dictionary according to a scale, but what is the measure? It depends on the lexicographer's ability. Because of this, everyone can’t compile a dictionary. Lexicographer looks at the other dictionaries. Are these
terms in the other dictionaries, or not? For example, a term is often used in philosophy books. I mean, we have to include these high frequent terms. Because these terms are being used in common language... Exactly there is no sharp boundary.” (Participant 16)

Participant 11 and Participant 3 indicate that the usage frequency and dispersion of the candidate lexemes are monitored at this boundary, but they do not give a clue as to how to decide the dispersion and frequency.

“We determine a boundary of common language, technical language, and spoken language by looking at the frequency and dispersion of words.” (Participant 11)
"In many dictionaries, the frequency is the most important criterion, but it is important to get the terms equally." (Participant 11)

Participant 5 also refers to the dispersion of the words searching the web.

“The criterion for the inclusion of the technical terms is the frequency criterion, plus the criterion of access to common usual language, the use in mass-media, on the internet, etc. All the specialized terms are accompanied by a mark that indicates the respective field and the definition must be accessible.” (Participant 5)

Participant 16 and 19 point out that in the selection of technical terms to be taken into the dictionaries the field experts are the decision makers, but they do not provide information on how this decision is made by field experts.

“Field advisers decide when terms are taken into the dictionary.” (Participant 16)
“We apply to field experts about the terms whether we will include or not.”(Participant 16)

Participant 15 states that, technical term which are being used in Daily usage are included in the dictionary. He also says that he understands a technical term that can be included into the dictionary intuitively.

Participant 2 draws attention to the text types and the ability of the text types to reach the dictionary users in terms boundary.

“As to specialized language, we include only those terms that are more commonly known, i.e. not only used by experts among themselves but e.g. also in newspapers on the specialized topic.” (Participant 2)
Participant 13 expresses how they use their own corpus to determine this boundary, the number of people using this technical word, the situation of a term in other dictionaries:

“Since this a comprehensive dictionary, words from all layers of the language can be included – general language, technical language, dialects, old and new words, spoken and written language – as long as they have been published in print.

The criteria are the following:
- Frequency;
- How many people use it (e.g. specialized words that frequently appear during high school studies);
- Appearance in other dictionaries (e.g. dictionaries of technical terms, dialects, history of the language, spelling etc.);
- Grammatical viewpoints (e.g. mirroring in changes of the language);
- Historical significance.

We rate the words according to the whole of our own corpus (not the internet).
We decide whether a word is technical or general based on the corpus (where do we find data).” (Participant 13)

Participant 14 states that it is important to identify the subject areas and level of knowledge in order to determine this boundary.

“We do not include too many technical words, technical vocabulary, and specialist terms in our dictionary. It’s really hard to decide because there are too many different fields and there are so many sub-fields. For instance, there are many sub-fields of sports, engineering etc....There are specialist dictionaries for these fields. It is difficult to put together all of these words in a general dictionary because it is hard to determine where you will finish the boundary.... It is necessary to take words like badminton, shuttlecock (badminton ball). But you had better not to enter a specialized field in your general dictionary.” (Participant 14)

Participant 10 says that terms which have been used in high school textbooks are important to determine this boundary. The reason for this is explained by the fact that many people face with these and texts and terms.

“However, there are certain exceptions to this, e.g. technical terms can be included when used in sources such as high-school textbooks...Questions regarding limitations are not easily tangible. However, we do rely on text types to determine the relevance in terms of how the use of words is spread among different text types, authors, publishers, time etc.” (Participant 10)
The views of the Participant 8 also support the importance of the level of knowledge for determining technical word boundary.

“We omit technical terms which are sourced only in terminology dictionaries or in literature for tertiary education and professional specialists. We include technical terms which are or have been commonly used in teaching literature for school up to the secondary level. We also look for documented usage in general literature or newspapers. We take a stricter view of compounds that of basic word forms.” (Participant 8)

7. Discussion and Implications

As shown in the answers to the qualitative research questions, all participants emphasize the importance of the technical words boundary. However, most of the participants have an intuitive approach to the decision of the technical word boundary for their general dictionary. It was found that some participants used traditional methods instead of using a corpus when compiling general dictionaries. Some participants who used a corpus as a lexicographical database did not provide detailed information about the method they followed for this boundary. Some participants indicated that the determination of this boundary was far from a database rather than a lemma selection. Some participants are following their personal preferences rather than considering dictionary users when they determine this boundary. Some participants emphasized that the texts in the corpus should be classified in the form of specialized texts and general user texts in order to determine the boundary.

8. Conclusion

This study focuses on revealing the technical word boundary point view of the lexicographers who compiled dictionaries for various languages in Europe. A comprehensive model has been developed in relation to the way this boundary is determined by scientific methods. Since this model will be more comprehensive than this presentation, it will be dealt with in a wide article.
References


Gender Representation in the *Contemporary Chinese Dictionary*

Huilian Hu  
Guangdong University of Foreign Studies  
Jane20080219@126.com

Abstract

Although gender issue in lexicography has been extensively studied, analytical attention is largely limited to the dictionaries in the English speaking world and relied on researcher's introspection. Little attention has been paid to Chinese dictionaries. Drawing on Appraisal theory, this study sets out to investigate the similarities and differences of attitudinal resources in the representation of the male and the female in the examples of the seventh edition of the *Contemporary Chinese Dictionary*(CCD7). The causes and implications of different linguistic resource for male and female representation will be discussed.

**Keywords:** Gender representation, Chinese dictionaries, Appraisal Theory

1. Introduction

Gender study has achieved a certain degree of centrality in lexicography in recent decades. Relevant studies often put their focus on the representation of androcentrism, or the practice of placing males above females, consciously or unconsciously. Moon (1989: 59), among others, argues that it is impossible for lexicographers to avoid displaying their prejudices in the language that they choose to define and explain. Thus dictionary criticism has a crucial role to play in transforming the often unsatisfactory situation (Akasu 2013). Since the beginning of the 21st century, there has been a flurry of studies concerning gender and lexicography. More dictionaries were brought into perspectives and lexicographers started to treat the representation of gender in dictionaries more seriously. Studies tend to manifest deeper insights and more dimensions. Landau (2001), for example, noticed that dictionaries were often accused of biases for their treatment of definitions concerning women and minority groups. “Every established dictionary reflects, however it may strive to be impartial, the prevailing biases of its times, because the biases often inhere in the very manner of expression used in its definitions” (Landau 2001: 420-421). Bejoint (2010: 213) offered a brief description of how “for a long time, women were badly treated in dictionaries”. Investigations of gender representation in specific dictionaries are more inclined to use corpus data, yielding more convincing results. Moon (2014) acknowledges the lexicographical challenge to balance the conflict between equal legal status in theory and disparities in practice for the lives of men and women. “An
interesting tension between language, lexicography, and receiver” was noticed in lexicography (Moon 2014: 98).

Gender terms usually carry explicit or implicit evaluative meaning. Their evaluative meaning can be encoded in the inclusion of headwords, explanations and in examples. The language of evaluation is the focal point of a branch of SFL, the Appraisal theory. In this study we decide to apply the appraisal theory to the analysis of dictionaries. The dictionary we choose for this study is CCD7. A brief look into the research of CCD show that the focus issue in this study has not been adequately addressed in the academic world, though we notice that the issue of gender and lexicography has drawn some academic attention in the Chinese context. Zhao (2015) questions the inconsistency in the entry inclusion of CCD6, which results in its failure to incorporate some very commonly used gender lexicons, notably, 剩男 (literally, leftover men, men at a mature age unable to find an ideal spouse), 剩女 (literally, leftover women, woman at a mature age with a good education and high salary but unable to find an ideal spouse) and 同志 (gay). Zhao suggests that it is pointless to exclude them from the dictionary for personal dislikes, or for social or political reasons. In this study, we will look at how different appraisal resources are employed in the representation of males and females. In particular, we will look at the attitudinal resources employed in sentence examples. Differences will compared, analyzed and explained from a variety of perspectives.

2. Theoretical framework

The present study will be conducted mainly within the framework of Feminist Critical Discourse Analysis (FCDA), while at the same time will rely heavily on systemic functional linguistics (SFL). Within SFL, the appraisal system constitutes the major analytical framework for analyzing evaluative meaning in the gender representation in CCD.

Feminist Critical Discourse Analysis may be described as doing CDA from a feminist perspective (Lazar 2005). As an extension of CDA, FCDA shares some of the same major tenets with CDA. FCDA aims to critique discourses which sustain a patriarchal social order(Lazar2005: 5). It takes a critical stance by exposing the less obvious and implicit meanings for the subtle, complex renderings of ideological assumptions and power relations in contemporary societies(Lazar2005: 13). FCDA focuses on “how gender ideology and gendered relations of power are produced, negotiated and contested in representations of social practices” (Lazar 2005: 11). Besides, FCDA adopts the post-modern feminist idea of “the diversity of gender identities and gendered practices” (Cameron, 2005: 482). Rather than looking at women as a group who is always constructed powerless, as opposed to men who are always constructed powerfully, postmodern feminists acknowledge the diverse ways in
which gender identities are constructed.

Appraisal theory is concerned with the linguistic resources with which a text/speaker comes to express, negotiate and naturalize particular inter-subjective and ideological positions. Located within the framework of Systemic Functional Linguistics, the theory is concerned with the language of evaluation, attitude and emotion (White 2001). Appraisal theory divides evaluative resources into three broad semantic domains: attitude, gradation and engagement. This study mainly concerns the attitude system.

Attitude includes those meanings by which texts/speakers attach an intersubjective value or assessment to participants and processes by reference either to emotional responses or to systems of culturally-determined value systems. Attitude itself divides into three sub-systems.

Affect, which covers the explicit expression of positive or negative feelings by the speaker/writer or someone else.

Judgement, ‘deals with attitudes towards behavior, which we admire or criticize, praise or condemn’ (Martin and White, 2005: 42). Judgement concerns social esteem and ethical evaluations, and applies mostly to persons or institutions.

Appreciation, which ‘involves evaluations of semiotic and natural phenomena, according to the ways in which they are valued or not in a given field’(Martin and White, 2005: 43). The evaluations are aesthetic or functional, and they refer mostly to works of art or literature and to non-human physical objects, rather than to humans.

The attitude system will be used to pinpoint the evaluative meaning in illustrative examples in CCD7. A comparison between the different evaluative patterns concerning males and females will be carried out to demonstrate the evaluative tendencies of the dictionary.

3. A Case Study: the Contemporary Chinese Dictionary and the Construction of Gender

This section explains the choice of CCD7 as the object of study, and describes data collection, the analyzing tools and the analytic procedure in the current study.

3.1 Data and method

The reason for choosing CCD as the study object is CCD’s relative importance compared with other Chinese dictionaries. Contemporary Chinese Dictionary (CCD), the 7th edition of which was published in 2016, was from its beginning endowed with the task of setting standards for the modern Chinese language. CCD is extremely influential on the Chinese market, and due to the lack of qualified Chinese dictionaries on the market, CCD also serves as a learners’ dictionary for foreign learners (Zhao, 2015).

Corpus investigation is going to be employed for the collection and comparison of attitudinal resources in example sentences for gender terms in CCD7. According to
Hunston (2011: 166 -167), the roles corpus investigation techniques can play in the study of evaluative language are as follows:

They allow a researcher to establish that a given word or phrase has a typical evaluative use or polarity. They permit quantification of evaluative meaning in one set of texts over another, by counting the occurrences of given forms. They permit mapping of meaning elements on to form elements where these coincide consistently. They allow a researcher to observe consistency in co-text in meaning as well as in form. In addition, corpus assisted discourse analysis also places importance on providing explanations for findings, taking the contextual information into account (Baker & Ellece 2011: 25).

This study will address the following three questions:

1) What are the similarities and differences of the distribution of attitudinal resources in male representation and female representation in the example in CCD7?
2) What are the reasons to cause the similarities and differences?
3) What are the implications for dictionary criticism and dictionary compilation?

3.2 Analysis and Discussion

This section is intended to report what kinds of attitudinal resources are employed and how they are distributed in dictionary examples. The distribution of attitudinal resources including affect, Judgement and appreciation for male and female gender will be investigated. Differences in the patterns found are then interpreted to reveal different representation of two genders in the dictionary. Finally, implications for dictionary criticism and dictionary compiling are briefly explored.

3.2.1 Attitudinal resources and distributional characteristics

We often expect dictionaries to be objective and scientific to such an extent that we overlook the fact that lexicographers are human beings nothing much different from the rest of us. Their personal inclination, vision of the world, values they hold to be true, cannot be shielded altogether from the job they undertake. It is just as natural as Samuel Johnson’s despise for Scotland that modern lexicographers express their attitude when it comes to different genders, though the fact might escape their attention. Compiling a dictionary is such a long and gigantic project that people who participate in the project might not foresee what kind of influence their sporadic decisions, choosing words, picking examples or making up examples, will have on the dictionary as a whole. However, the cumulative effect of individual decisions sometimes lead to a result that contributors to the project failed to see in advance or to recognize in retrospect. It takes an outsider, the reader of dictionary, to put the dictionary in the big
picture.

The following table shows the distribution of attitudinal resources of male and female related examples in CCD7.

Table 1 Distribution of attitudinal resources of male and female related examples in CCD7

<table>
<thead>
<tr>
<th>Examples</th>
<th>Attitude</th>
<th>Affect</th>
<th>Judgement</th>
<th>Appreciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>677</td>
<td>93</td>
<td>482</td>
<td>102</td>
</tr>
<tr>
<td>Male</td>
<td>2831</td>
<td>97</td>
<td>2701</td>
<td>33</td>
</tr>
</tbody>
</table>

For male, there are 2831 Attitudinal resources, with 97 Affect resources, 2334 Judgement resources and 33 Appreciation resources. For female, there are 677 Attitudinal resources, with 93 Affect resources and 102 Appreciation resources. There are several things we can notice from the table. The first is the surprising contrast of the total number of attitudinal resources between male and female, with the number of attitudinal resources for male about 4 times of that of the female. The simple fact behind the sharp contrast is that males appear far more frequently than females. An attempt to explain the contrast is the grammatical tradition of using the male pronoun to substitute a person whose gender is unknown and to cover both genders when it is required. Be that as it may, we are still surprised to see the huge difference in the example sentences. Not to mention the fact that the grammatical tradition doesn’t lack a female pronoun when it is needed.

The second is that for both genders, Judgement resources constitute the largest portion of attitudinal resources. Males have a higher percentage of Judgement resources than female and a lower percentage of Affect resources than female. Females have a much higher percentage of Appreciation resources than males. The result is more or less out of our expectation because we predict that for females, there will be a higher percentage of Affect and Appreciation resources than Judgement resources. It proves that even for females, Judgement resources is still the paramount attitudinal resources. Though there are some differences compared with males, for example the social domains in which the examples take place, which we will discuss in greater details in later sections.

3.2.2 Comparative analysis of Affect resources

It is human instinct and human nature to feel, to love, to hate, to be happy or angry, to be overjoyed or irritated. A person might go around with his social identities unknown and concealed, but can hardly hide his/her emotions from others. Since emotions display itself through one’s facial expression, one’s posture, one’s manner of talking, or even one’s way of walking. Appraisal theory puts Affect in the first place of the first category of the evaluative resources for a good reason. The expression of emotion is the most direct and most obvious demonstration of the speaker or writer’s attitude. It is evaluation in terms of psychology and emotion. According to Martin & White(2005),
the person who engages him/herself in the emotional experience is called the Emoter and the object which brings out the emotion is termed as the Trigger. Martin & Rose (2003) classified emotion into three categories, with three pairs of variants, happy/unhappy, secure/insecure, satisfied/dissatisfied. The first category concerns one’s psychological feelings, the second concerns interpersonal communication, and the third is related one’s objectives.

As we put together the example sentences related to males and females and read through them in a systemic way, we notice that dictionaries are like story books, with heroes and heroines, with plots and storylines. The men and women in the dictionary (in our case, the CCD7), though appearing in separate sentences, manage to provide us a general and vague picture of their emotional world, with similarities and differences as shown in Table 2, which demonstrates the distribution of Affect resources of male and female dictionary examples.

**Table 2** Comparison of Affect resources between male and female examples

<table>
<thead>
<tr>
<th>Types of Affect resources</th>
<th>Distribution of affect resource types for male examples</th>
<th>Distribution of affect resource types for female examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy—Unhappy</td>
<td>18—33</td>
<td>24—18</td>
</tr>
<tr>
<td>Secure—Insecure</td>
<td>8—9</td>
<td>4—6</td>
</tr>
<tr>
<td>Satisfied—Dissatisfied</td>
<td>1—19</td>
<td>12—17</td>
</tr>
<tr>
<td>Incline—Disincline</td>
<td>7—2</td>
<td>12—0</td>
</tr>
<tr>
<td>Total</td>
<td>34—36</td>
<td>52—41</td>
</tr>
</tbody>
</table>

Table 2 shows that the types of Affect resources employed in describing men and women differ greatly from each other. It is important to mention in the first place that although the total number of Affect resources for women is less than that of men, the ratio of women’s Affect in the attitudinal resources is much higher than that of men, indicating that women are more likely to demonstrate their feelings in an open manner. In the subcategories, more women are depicted as happy than unhappy while the opposite is the case with men. In the secure-insecure category, there isn’t much of a difference between the two genders. In the satisfied-dissatisfied category, only one man is described as satisfied while 12 cases of women are described as satisfied. In the last category, incline-disincline, none of the women appearing in the examples is depicted as disincline.

We notice that among 93 cases of Affect resources for the females, women are mostly projected in a family role. Loving mothers and grandmothers seem to be the favorite picks of lexicographers, as illustrated by the following examples

祖母钟爱小孙子。 (The grandmother especially dotes on her little grandson.)

母亲最疼爱小女儿。 (Mother dotes on her younger daughter more than anyone else.)
The mother showered all her affection on her children.

The old granny has a loving heart for children.

Mothers and grandmothers who miss their children or grandchildren frequently occur in examples, as illustrated by following concordance:

You’ r back just in time! Mother has been missing you.

My mother finally set her mind at rest after I had settled down.

She feels a little better with her daughter keeping her company.

Her letter is full of nostalgia.

The mother is reluctant to leave her child far away from home.

The mother worries herself very much over her daughter who works far away.

The granny blames her children for not visiting her often.

She is happy to see her future daughter-in-law.

The young lady blushes when hearing about her marriage.

Hearing that an old lady is trying to match her with someone, she blushes.

Women are sometimes involved in negative feelings such as sadness, disappointment, regret, etc, and they burst into tears sometimes, as illustrated by speaking of her grief, she could not help but choke on her sobs.

She got so worked up talking about it that her tears flowed freely.
Sadness took over her after the departure. Their happiness is usually not related to their own success, but to the joining of family or the success of their children, as in 儿子总算理解了自己的苦心，她感到很宽慰(Knowing that her son understand her intentions, she felt reassuring); 女儿考上了大学，妈妈喜欢得不得了(The mother is filled with great joy to know that her daughter is enrolled into university); 奶奶笑眯眯地看孙子的立功喜报(The grandmother smiles at her grandson's award.)

Analysis of the Affect resources for females reveal several dimensions of the lexicographical choices of CCD7. First, women are mostly projected in a family and domestic roles. Their emotions are closely related to the reunion or parting of family members, usually an offspring of theirs. Second, women are more inclined to have negative emotions which cause them to cry, rather than positive emotions such as joyful or cheerful. They can be happy sometimes, but their happiness is usually related to the success of a family member, rather than their own success in career or life. Third, women are pictured as having a particular passion towards relationships and marriage. We tend to explain this as the influence of a patriarchal beliefs. One of the founding pillars of the patriarchal society is the marriage system. Women are considered to be of value only as the possible spouse of a man and especially when they are able to produce heirs for the male sex. If we combine the above three points together, the logic will be clearer. Together, they draw a map or line of development for the female kind, the order mixed up though. It situates women into the linear development of finding a man, getting married, having children, fostering children, bearing the pain of parting with your children, enjoy the growing up and success of your children. These steps seem to dominate women’s life and rule the emotional world as well. This finding seems to echo the words of Sunderland (2004: 69), ‘women as domestic’ discourse can be regarded as an overarching discourse under which are more specific discourses such as ‘woman as cook’, ‘woman as nurturer’, ‘woman as cleaner’.

Affect resources employed in the male-related examples tend to place males in a broader social context and a wider historical dimension. It is true that men in the dictionary examples share some of the same emotions with women. Regular emotions such as happiness, sadness, anxiousness are prevalent with men as well as women. What’s different is the ‘trigger’ of their emotions. While women’s emotions are usually caused by domestic affairs, for instance, reunion and parting of the family members, marriage of a daughter, illness of a baby, men’s emotions are often triggered by outside matters. The success of a scientific experiment, for example, can cause sadness or pleasure in a man, as shown in examples 虽然试验失败了，但他并不悲观。(Although the experiment was a failure, he did not lose heart.); 难怪他今天这么高兴，原来新
机器试验成功了。(No wonder he is so glad, experiment of the machine was a success.) Examples often put men in a public venue where social activities take place, a meeting or gathering in 他的不安情绪迅速传染给了在座的人(His anxiety infected all the people present.); a class in school as in 他和蔼地瞥视了一下每个听讲学生.(Attention from the leaders warms his heart.) In the example of 他向来鄙视那些帮闲文人(He has always despised such hack scholars.), the subject is likely to be a prestigious writer and in 领导的关爱, 让他心里暖融融的(, the subject is likely to be a worker or bureaucrat at the lower level. In 他在青少年时期就十分同情被压迫的劳苦大众 (He felt for the oppressed toiling masses while still a boy.), the hero is put into the historical context of Chinese revolution. Loving of the countryside or one’s hometown is also the common theme for men, as in 他对农村产生了深厚的感情(He has developed a deep affection for the countryside);他站在高台上，深情地望着家乡的土地.( Standing on the tower, he is staring at hometown earth.) In 多年的收藏毁于一旦，对他刺激很大(He was terribly upset after his collection of many years was destroyed overnight.), the subject has a hobby of collection. It is noticed that men express their emotions in a different way from women. Women often shed tears when they encounter unhappy events or obstacles in life. Men in dictionary examples often fly into violent anger, especially when they have clashes with others, as illustrated in examples 他整天黑着脸, 准是碰到了什么不痛快的事(He was terribly upset after his collection of many years was destroyed overnight.);他听了这种不三不四的话非常气愤(He was exasperated by such dubious remarks.);他正在气头上，别人的话听不进去(He is so angry that he won't to listen to anybody.);他一听这话就动起火来(He flew into a rage on hearing such words.);气得他差点儿背过气去(He is so angry that he can not breath.);他尽力使自己的怒火平抑下来(He tries to sooth himself.);他气得直哆嗦(He shakes in great anger.) In 他激动得说不出
话来，嘴唇在微微颤动（He was too excited to say anything-- his lips trembled slightly.），the emotion of the character is enhanced by the slight movement of his lips. In one example, 他就是这么婆婆妈妈的，动不动就掉眼泪（He is womanish shedding tears now and then），the man who easily shed tears is described to be womanish, and it is the only case in which a man is described as a crier.

Analysis of the Affect resources of men reveal several features. First, though men have the same emotions as that of women, their emotions take place in a larger social, cultural and historical context, in comparison to the largely domestic environment for women. Second, in general, men are described as able and strong-minded, with a variety of social roles to play. When they express their emotions, they tend to express in a different manner from women. Their emotions are stronger, more intense and sometimes couples with physical movements. Third, men are sometimes depicted against women as their object of comparison. A man who lacks masculinity can be described as woman-like, ending up prioritizing men over women.

### 3.2.3 Comparative analysis of Judgement resources

Table 3 shows that Judgement take up the biggest proportion in male and female examples.

<table>
<thead>
<tr>
<th>Judgement</th>
<th>Judgement for male</th>
<th>Judgement for female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social sanction</td>
<td>417</td>
<td>57</td>
</tr>
<tr>
<td>Social esteem</td>
<td>2284</td>
<td>425</td>
</tr>
<tr>
<td>Total</td>
<td>2701</td>
<td>482</td>
</tr>
</tbody>
</table>

According to Martin & Rose (2003), Judgement belongs to the realm of ethics, against which human behaviors are judged. Judgement resources are further divided into social sanction and social esteem. Social sanction is evaluated though veracity (honest or dishonest) and propriety (moral or immoral), and is related to laws and regulations. Social sanction can be further divided into praise and condemn. Social esteem has to do normality (normal or abnormal), capacity (able or disable), tenacity (reliable or unreliable). Two tendencies can be distinguished, admire and criticize. Social sanction is mainly evaluated through the terms of laws, regulation and social constraints while social esteem is more based on ethical ground or psychological situation.

The realization of Judgement can be divided into two types, namely the overt the covert. Overt Judgement is realized through property adjectives, nouns or verbs. Covert Judgement is realized through inscribed linguistic resources which do not employ evaluative vocabulary. Selected materials show that male and female examples differ greatly in their realization of Judgement resources.

In Appraisal system, Judgement is mainly related to the social, legislative, moral and ethical domains. Judgement can be thought of as the institutionalization of feeling, in
the context of proposals (norms about how people should and shouldn’t behave) (Martin & Rose 2007: 67). Like Affect, there are positive and negative dimension to Judgement. Human characters and human behaviors are judged against certain sets of standards, be it moral or legal. Women’s characters and behaviors are also subject to certain sets of moral and social standards and in our case, the Judgement resources for women more or less reflect the Chinese ideology regarding women. Take the following dictionary information for examples, 克尽妇道(strictly observe the rules of proper female behaviour); 谨守妇道(carefully observe the rules of proper female behaviour); 举止端庄,有大家闺范(graceful and well-bred); 足不出户的大家闺秀(well-bred girl who never steps out of house). 妇道(rules of proper female behaviour) literally means women’s code of conduct. It is a phrase taken from classical Chinese that requires the obedience of women in the patriarchal system. Women who are loose with their sexual life would be described as 风骚的女人(flirtatious woman) or 卖弄风骚(play the coquet). In the sentence 她不愿沦为有钱人的玩物(Shes n’t want to be a doll for the rich people), it seems that it is normal for a beautiful woman to become the sexual object for the rich people.

Marriage and relationship are important background settings for the Judgement of women. In example sentences 我当闺女那工夫,婚姻全凭父母之命,媒妁之言。(When I was a girl, matches were arranged by gobetweens according to the wishes of the parents.) and 她择偶的标准过于理想化,因而至今仍是单身(Shes too ideal about choosing husband to be married.), a successful marriage is described as of paramount importance for the success of a woman. The arranged marriage in the first sentence and the failure in finding a proper spouse indicate that a marriage is a proper way for women to be gain social cognition and to be evaluated positively in the society. Covert Judgement of women’s abilities are found in examples such as 妇道人家(women) and 妇孺皆知(Even women and children know.) without being labeled or alerted as such.

Positive Judgement are mainly found in the social production sector, where women have to prove themselves with their working ability. The following examples put women in social productive sectors, running machines, gaining prizes for working abilities. Such as 人们都夸赞她心灵手巧( Everyone praised her for being clever and adept.); 她能同时管十台机器(Shes can tend ten machines at the same time.); 她是个好
强的姑娘，从来不肯落后 (She has a progressive mind, never wanting to lag behind.)

她为人明快达观，工作起来雷厉风行 (She takes a philosophical attitude towards others, but handles her work vigorously and resolutely); 她们姐妹俩都是先进生产者 (The two sisters are both excellent workers); 她很有灵气，一定能成为出色的服装设计师 (She is a woman of intelligence, and will without doubt become an outstanding fashion designer.)

In some cases, Judgement of a woman’s social esteem relies on her male descendents, for example her son, as in 儿子有出息，当妈的也觉得风光 (A son’s success brings credit to his mother.) In other cases, the positive Judgement of woman is achieved by equaling her to a man, as in 巾帼丈夫 (heroine) and 巾帼英雄 (heroic woman). Woman are also positively judged according to their moral character, as in 她走到哪里, 就把好事做到哪里 (Wherever she goes, she does good deeds.) Since Judgement occurs in social context, some of the covert Judgement is embedded in the description of social phenomenon. In 老婆外出打工，妻子留守在家 (The husband works as migrant worker while the wife stays at home.); the wife is judged to be of less value since she doesn’t have the ability of earning money. Housework requires labor as well, but it is not put on a equal footing with works that earns money. It is because of this that the woman in 她要洗衣、做饭，外带照顾老人，没有时间干别的事了 (She simply has no time to spare for other things, having to do all the washing besides looking after the elderly.) is deprived of the time and effort to make the full potential of her time, since she is occupied with housework.

Some Judgement resources have to do with some stereotyped images of woman. For instance, the connection of woman with talking is exaggerated in examples

她说着说着就走了嘴了 (Words slipped her tongue.)

这小鬼嘴刁，差点儿被她骗了 (What a little trickster she is. I was almost taken in.)

这小姑娘嘴乖，挺逗人喜欢 (What a dear little girl, talking so sweetly!)

老太太嘴碎，遇事总爱唠叨 (The old lady is loquacious and always babbling on something.)
她能言善辩，说出的话滴水不漏。(She has the gift of the gab, and leaves nothing for people to criticize when she speaks.)

她是一个爱说爱逗的姑娘。(She is a girl fond of chatting and joking.)

老秦瞪了她一眼，嫌她多嘴。(Old Qin glared at her, displeased that she was so long-winded.)

Judgement resources for males are much more numerous than that of females. As the following examples demonstrate, men are situated in a wider social context than women with various kinds of jobs.

在生意场上，他黑道白道都走过。(He walked paths both legal and illegal in the business world.)

他当了官好（hào）摆个谱儿。(After becoming an official he likes to put on airs.)

他虽然年纪轻，辈数儿小，但在村里很有威信。(Although young in age and low within the family hierarchy, he enjoys high prestige in the village.)

他原来是医生，还是让他干老本行吧。(Let him take up his original profession as a doctor.)

这两年，他编导了几部新戏。(He wrote and directed several new plays.)

他在厂里多次受到表扬。(His factory commended him on several occasions.)

他做生意童叟无欺，深受群众称许。(He was highly commended for his honesty and fairness in business dealings.)

他曾被朝廷贬到边远地区做官。(He had been degrade to work in remote places by the governent.)

他当了一辈子教师。(He worked as a teacher all his life.)

他的初愿是当个中学教师，没想到后来成了大学教授。(At first he wanted to become a middle school teacher, but ended up a university professor.)

他除了教课，还负责学校里工会的工作。(He is in charge of the trade union besides teaching.)

他曾经当过三年翻译。(He worked as a translator for three years.)
Men appear in roles such as businessmen, government officials, doctors, writers-directors, technicians, farmers, professors, translators etc. In most of these roles, their abilities are positively evaluated. Not to mention the fact that these social roles are in themselves a symbol of one’s abilities.

Apart from occupational roles, abilities are also demonstrated through one’s social activities. Take the following sentences for examples.

他成年累月和牲口打交道，养牲口的经验很丰富。(He has rich experience in raising animals since he has dealt with them year in, year out.)

这个收音机让他摆划好了。(This radio works well after he repaired it.)

他试着用半生不熟的英语跟外宾交谈。(He tries to speak to foreigners in broken English.)

他的藏书大部分都赠给国家图书馆了，自己只保留了一小部分。(He has donated most of his book collection to the National Library, keeping only a small part for himself.)

他在工作中的表现很好。(He is doing very well in his work.)

Abilities such as repairing, speaking, writing and working are connected with the subject. Though some of Judgements are negative in appearance, they still contribute to one’s overall social image. Public sphere, a term invented by Habermas (1984), might help to explain the differences between men and women and their relative importance in society. Public sphere refers to the relationship between social systems and everyday life. People act as citizens, express their opinions, influence public policies in the public sphere. Usually, the more powerful members of a society have more access to the public sphere, taking part in public affairs that impact on public policies and directions of social development. (Baker & Ellece 2011: 107)

3.2.4 Comparative analysis of Appreciation resources

Appreciation refers to the evaluation of man-made or natural objects and abstract structures, consisting of three sub-categories, namely reaction, composition and value. Reaction focuses on the influence and quality of the evaluated objects., Composition evaluates whether they conform to the structural regularities, focusing on the
complexity and details of the evaluated objects. Value is based on social norms, concentrating on the social property of the evaluated objects. (Martin & Rose 2003)

According to the above categorization of Appreciation resources, we compared the results based on the collected materials. See Table 4.

Table 4 Comparison of Appreciation between male and female dictionary examples

<table>
<thead>
<tr>
<th>Types of Appreciation</th>
<th>Types of appreciation of male examples</th>
<th>Types of appreciation of female examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reaction</td>
<td>17</td>
<td>67</td>
</tr>
<tr>
<td>Composition</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>Value</td>
<td>11</td>
<td>21</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>102</td>
</tr>
</tbody>
</table>

As the above table shows, the distribution of different Appreciation subcategories of male examples and female examples doesn’t manifest much difference. We do notice, that the total number of Appreciation resources for women is much higher than that of men, despite the fact that the total number of attitudinal resources employed for women is much lower than that of men. This suggest that for female examples, the frequency of Appreciation resources is much higher than that of male examples. Otherwise, female examples employ more Reaction than male examples, and a smaller percentage of Value resources.

Appreciation resources for females show a strong tendency to focus on their looks and dresses. As in the examples. (The little girl has turned even more beautiful since I saw her half a year ago.);这姑娘出挑得越发标致了(The young lady has grown more and more beautiful.);那姑娘出息得更漂亮了(The young woman has turned even more beautiful.), the subject is evaluated in a positive manner for their looks, as if their looks would contribute to their value as a person. Other examples include 这小姑娘长得真美（This little girl is really beautiful.）; 她长得十分美貌(She is very pretty.);新来的女孩儿长得有模有样的，很讨人喜欢.(The new girl is good-looking and is very popular with everybody.) Some of the appreciation resources reflect the common standard for beauty in the society. Being slim is to be valued in 她身材纤细，很有骨感 (She has a slim figure.) Being young or looking young is also a celebrated quality, as in 妙龄女郎(young lady) and 她已过五十，但看上去也就四十来岁(She is over fifty but looks only forty.) Meanwhile, what women wear is also an element in their
appreciation, as in the negative appreciation in 她这身打扮土不土，洋不洋，怪模怪样的。(It's queer looking; the way she is dressed is neither traditional nor Western.) and the positive appreciation 她穿上这身男装，显得格外俊雅(She looks elegant and handsome in this man's outfit);她长得少相，岁数儿可不小了。(She looks much younger than her age.)

The female voice is another point of interest in the appreciation resources for women. Examples include 她用歌声安慰母亲心灵上的创痛(She heals mother's wounded heart with her songs.); 她的嗓音挺清脆(Her voice is clear.);她的歌声高亢、激越，很有张力(Her voice is powerful and full of tension.);她一夜之间蹿红歌坛(She turned a diva overnight.)

The connection between women and clothes and ornament is demonstrated in the appreciation resources for females as well. In sentences 老太太又在开箱子兜翻她那点儿绣花的活计(The old lady is rummaging through the chest for her embroidery work again.); 姑娘们一个个打扮得花红柳绿(The girls were all colorfully dressed.); 她很有灵气，一定能成为出色的服装设计师。(She is a woman of intelligence, and will without doubt become an outstanding fashion designer.); 她把贴己首饰卖了，贴补家用(She sold her jewelry to help out with the family expenses.). 几种颜色的布她都不中意(She couldn't find one that she liked among the cloth of all the colors.); 这件衣服很中她的意(She likes this clothes very much.), the women in the subject position are all connected with clothes or ornament. Frequent jobs for women in the examples include nanny, singer, designer, or housewife.

The number of Appreciation resources for males is smaller than that of females. Men are not judged by the similar standards with women. In examples like 他的眉毛显得浓黑粗重(He has bushy black eyebrows.); 他人矮，踮着脚才能看见, the men are judged by their masculine features. In 他只是显得有点儿面老，其实还不到四十(He is less than forty though looked much older.), the men is negatively evaluated for
looking older than his age. Otherwise, men are also positively appreciated by their writing skills, performance, etc, as in examples 他的作文被评为优秀 (His composition earned an excellent mark.); 他的表演太做作了 (His acting is overdone.) ; 他的精彩表演使晚会生色不少 (His superb performance made the evening party more enjoyable.); 他们的表演各有特色。 (There's something special about each of their performances.)

4. Conclusion

While the imbalance and incompatibility between Attitudinal resources for male and female examples are easy to notice, the underlying reasons for that imbalance and incompatibility are rather complicated and, to a certain extent, immune to regular reproaches. We may as well blame the lexicographers of the CCD7, or any similar dictionaries, for doing this and not doing that, for including this and excluding that, to emphasize a certain aspect and overlook another aspect. But still, we need to understand their dilemmas. All lexicographical choices are made for a reason. Whether the reason is good enough is subject to dissents and argumentations. As long as finger-pointing won’t do much help, we probe into the reasons behind the representation. And of course, the explanation stage would try to focus on the wider social context, asking why men and women are being represented in this way and ‘what consequences this may have for society and various group in it’ (Baker & Ellece 2011: 62).

4.1 The influence of ideology

Fairclough (2003) defines ideology as the representation of world outlooks which constitute social cognition and world. Ideology is a means to establish, maintain power relationships and a means of controlling and depriving. His definition is echoed by Reisigl & Wodak (2009:88), who argue that ideology is an important way to establish and maintain power inequality. Fairclough (1992) argues that the dominant ideology-discourse structure is capable of naturalized ideology as common sense or standard knowledge accepted by the society. To a certain extent, dictionaries are products of ideology which are presented as standard common sense. Examples which have been taken out of their context, put together, presented in the dictionary through a process of naturalization. The naturalness and commonality of these examples, separated and scattered in the dictionary pages, are thus internalized as daily discourse. This process is also called legitimation, ‘whereby something becomes legitimate according to the values of a particular society.’ (Baker & Ellece 2011:68) Van Leeuwen (2007) identifies four legitimation strategies: authorization, moral evaluation,
rationalization and mythopoesis. Legitimation through dictionaries fits into the authorization strategy, since dictionaries are normally considered authoritative source of knowledge and explanations. Discourse practice, through its representation of the reality and social situating of characters, helps to produce and reproduce the inequality among social classes, genders and nationalities. (Fairclough & Wodak, 1997: 258) ‘Discourses are ways of representing and constructing reality so that power relations are constructed, maintained and contested via discourse’ (Baker & Ellece 2011: 99). Lexicographical practice is also discourse practice, contributing to the construction of power relationship in the society.

Though the attitudinal resources employed for male and female examples, we notice that males are usually related to reason, activeness, instrumental, knowledge, competence, action and culture, while females are usually linked to emotion, passiveness, expressiveness, ignorance, incompetence, speech, nature and family. This observation is by no means definitive, but is able to provide a big picture of the gender ideology behind the lexicographical decisions.

4.2 The influence of perspective

Perspective is a basic conception based on spatial metaphor in cognitive science. We look at things through different angles and the representation of the observed subject is somewhat determined by the perspectives we choose. While spatial metaphor is the starting point for understanding perspective, it also concerns the time and means of cognition. All feelings and cognitions of human beings are decided by time, space and cognitive angle. To put in simple terms, perspective means the angle through which we observe and represent things. As an important linguistic phenomenon, perspectives have physical, cognitive and social properties, which are manifested on all layers of the language.

Dictionary compiling is a complicated process of selecting, arranging, arraying, explaining and representing. Under the influence of the patriarchal traditions, we tend to observe and represent men and women through different perspectives. Men are usually pictured in a much larger social context, doing ability-requiring tasks, while women are usually depicted as revolving around the domestic sphere, raising children, doing housework, looking after the elderly. The spatial and time constraints for women in dictionaries limited the possibilities of the perspectives through which we look at them. As a result, male examples are abundant with Judgement resources, since their sphere of social life chosen in the dictionary concern more about social sanctions and social esteem. Women, on the other hand, are more or less confined in family and domestic life. There is also a tendency to objectify women, as we observe the appreciation resources for females boasts a higher percentage than that of males. Even if when women are put in a wider social context, they are usually judged not by their abilities but by their properness of their behaviors.
5. Implications

Dictionaries are a sort of social representation, and there is no neutral representation of reality (Fowler 1987:67). Language itself is a social practice because it is an intrinsic part of a society. Through social construe, they tend to present, maintain or change specific ideologies. Social discourses not only express one’s own views of the world but also have an impact on others’ ideologies. Halliday (1990/2003: 168) finds out that the construing of gender characters plays a key role in the construction of gender discrimination. Sexist assumptions and biases are reflected, perpetuated and naturalized in language use and critical linguistic analysis can bring these discursive practices into sharper focus. Baker (2010) found that the most successful strategies of non-sexist language initiative seemed to be stop people from using a particular word. This can be achieved if lexicographers choose to exclude certain sexist words from the dictionary, which is commendable in the Chinese context as well. To recapitulate, the above analysis and explanation suggest that CCD7 differs greatly when it comes to the attitudinal resources for males and females. In general, we find that women have been treated unfairly in sentence examples, thus confined to a narrow social context and give less respect than men. The lexicographical treatments toward women, though having its deep-rooted reasons, fail to contribute to gender equality. Lexicographical practice may have to take new angles into perspectives when they choose examples for males and females. For examples, the lexicographical team may discuss the gender issues before the project begins, in order to offset the imbalance which existed in current dictionaries. The examples can be evaluated through their perspectives and ideologies in a systemic way.

References


Moon R 2014 ‘Meaning ideologies and learners’ dictionary’ In AAC Vettori & N Ralli (eds)


A Brief Introduction to *The American Heritage Dictionary of the English Language* (5th)

Li Li  
Shaanxi Normal University  
931965552@qq.com

Abstract

*The American Heritage Dictionary of the English Language* (AHD) is a dictionary with distinctive features among many American college dictionaries. AHD is a very authoritative and influential dictionary in America. It was once appraised by Amazon.com as the reference chosen by editors; it is also regarded as the essential tool for professional English workers. The first edition of the dictionary was published in 1969 and the latest edition was published in 2011. The author tries to make a brief introduction and evaluation of the AHD5, analyze the features of this dictionary from two aspects, which are macrostructure and microstructure. The former focuses on the front matter and back matter, and the latter focuses on the pronunciation, definition, illustrative examples, labels, illustrations, notes and cross-references, especially the notes, which includes four categories: Synonym Paragraphs, Usage Notes, Word History and Our Living Language Notes. Then the author tries to point out the advantages and the disadvantages of AHD5, with the hope that this article will provide some implication and enlightenment to the compilation of college dictionary in China.

**Key words**: AHD5, Macrostructure, Microstructure

1. Introduction

The American Heritage Dictionary of the English Language (AHD) is an American dictionary with distinctive features among many American college dictionaries, which is published by Boston publisher Houghton Mifflin, the first edition of which appeared in 1969. The AHD broke ground among dictionaries by using corpus linguistics for compiling word frequencies and other information. It took the innovative step of combining prescriptive information and descriptive information. It sold more than a million copies in the first year, it was once appraised by Amazon.com as the reference chosen by editors; it is also regarded as the essential tool for professional English workers. The author tries to make a brief introduction and evaluation of the AHD5, analyze the features of this dictionary from two aspects, which are macrostructure and microstructure. Then the author tries to point out the advantages and the disadvantages of AHD5, with the hope that this article will have a certain sense of enlightenment on
2. Macrostructure

Macrostructure refers to the way in which the entries are arranged within the dictionary. It also includes the dictionary’s format, binding and layout, front and back matter, and so on. We will focus on the introduction of the front and back matter of AHD5 in this article.

2.1 Front matter

The front matter of AHD5 is quite rich. The first page is Table of Contents, which may be “considered some kind of survey article summarizing the entire dictionary structure” as “it is probably the only component to contain references to all the other components in the dictionary”. It provides not only the exact page number of the front and back matter, but also the page number of the charts and tables. There are 13 categories: “Tables of Alphabets”, “Development of the Alphabet”, “Books of the Bible”, “Three Principal Calendars” and “Color”, and so on. These charts and tables can be found among the body of the dictionary, they add a lot of information concisely and vividly. For example, “Development of the Alphabet” shows the history of 26 English letters in the form of table, from which we can see different ways of writing these letters in each period clearly, we can have a better understanding of the evolution process of English letters. What follows the Table of Contents is an Introduction written by Joseph P. Pickett, the executive editor of AHD5, he gives us a brief description about the differences between this new edition and the last one. The following is the name list of Editorial and Production Staff, Special Contributors and Consultants and The Usage Panel. It’s necessary to mention The Usage Panel, which is a group of nearly 200 prominent scholars, creative writers, journalists, diplomats, and others in occupations requiring mastery of language. The Panelists are surveyed annually to gauge the acceptability of particular usages and grammatical constructions, they will be polled before interpreting any controversial question of usage. The dictionary contains more than 400 “Usage Notes”, appearing at the end of relevant entries, which are said to be based upon consultations with the panel. In many cases, where there is a divergence of opinion in the panel, the notes report the different views in percentage terms. This is one of the most important features of AHD5, we should learn from it.

Following The Usage Panel are two essays, Variation and Change in Our Living Language and Usage in The American Heritage Dictionary, written by John R. Rickford and Steven Pinker respectively. The former introduces the change and various variation of American English, the latter introduces the usage information of AHD5, and gives a brief description of the usage panel and the usage notes. The following is Guide to the Dictionary, this guide explains the conventions used in presenting the great array of information contained in the dictionary in detail, which can help the readers find and understand the required information quickly and easily. And finally,
there is a Pronunciation Key, which gives a list of pronunciation symbols used in this
dictionary and some examples to illustrate how the symbols are pronounced.

2.2 Back matter

The back matter of AHD5 is very informative, too, which mainly focuses on the
exploration of etymologies. It not only provides dictionary users with a lot of
knowledge, but also has considerable reference value for the etymology researchers.
The first page is an essay Indo-European and the Indo-Europeans, written by Calvert
Watkins. This essay provides some basic information about the structure and grammar
of Proto-Indo-European as an aid to understanding the etymologies of the English
words. Then what follows this essay is a thirty-page Appendix I: Indo-European Roots,
which allows the reader to trace English words derived from Indo-European languages
back to their fundamental components in Proto-Indo-European, the parent language of
all ancient and modern Indo-European languages. Following is an essay Proto-semitic
Language and Culture, written by John Huehnergard, which introduces the structure,
grahram and culture of Semitic language, next is a seven-page Appendix II: Semitic
Roots. Accordingly, there are two explanatory guides to each appendix and a table of
the principal sound correspondences for the Indo-European and Semitic languages.
There is also a chart about The Semitic Family Tree, which displays the genetic
relationships among the principal languages of the Semitic family and loosely suggests
their geographic distribution as well as their ages relative to one another. In the last
page, there is a big pie chart about The Indo-European Family of Languages, this chart
shows the relationships among the languages of the Indo-European family. AHD5 is
highly praised for its treatment of etymologies.

3. Microstructure

Microstructure refers to the arrangement of information within the individual
dictionary entries, so it can also be called lemmata structure. This kind of information
may comprise such categories: spelling, pronunciation, word class, etymology,
definition and sense ordering, idiom, synonym, antonym, etc. The author will introduce
the lemmata structure of AHD5 from eight aspects: pronunciation, definition,
illustrative examples, labels, illustrations, notes and cross-references.

3.1 Pronunciation

The pronunciation of AHD5 is represented by special symbols that are enclosed in
parentheses and appear after the boldface entry word, the AHD and IPA symbols are
not the same, so in the margin of every other page, there is a key to the pronunciation
symbols in a block (see Figure 1). This brings great convenience to the users, so as to
avoid looking up the Pronunciation Key in the front matter each time.
3.2 Definition

The definitions of AHD5 are clear and succinct, the defining vocabulary is easy to understand. In order to save space, there are always plenty of abbreviations in the definition of most dictionaries, which bring much inconvenience to readers. However, AHD5 is very user-friendly, for it almost abandons all the so-called “dictionary shorthand notation”, except for some part-of-speech labels, such as adj. adv. prep. conj. n. etc. which are very easy to understand.

For the polysemous entries, the senses are arranged for the convenience of the reader with the central and often the most commonly sought meaning first. Senses and subsenses are grouped to show their relationships with each other. For example, in the fatal entry, the commonly sought meaning “Causing or capable of causing death” appears first and the now obsolete sense “Having been destined; fated” comes last in the series of five.

Among the definition, there are many boldface letters and numbers, they have different meanings. For example, the boldface letters before senses indicate that two or more subsenses are closely related. In a combined entry (Sometimes an entry word fulfills more than one grammatical function, in such cases the different part of speech are defined within a single entry called a combined entry) the senses are numbered in separate sequences after each part of speech. The shift in grammatical function is indicated by the blue symbol and the appropriate part-of-speech label. The relevant phrasal verbs and idioms are set in boldface type and introduced by the heading --phrasal verbs and--idioms. All of these make the definition very clear at a glance. For instance, the entry bid (see figure 2) shows that.
3.3 Illustrative Examples

Examples are effective means of showing how the lemma behaves in combination with other lexical units, that is in the context of its actual use, demonstrating its morphological, syntactical, collocational, connotational, stylistic and sociocultural features(Yong, 2007). In AHD5, there are tens of thousands of illustrative examples that follow the definitions and show the entry word in typical contexts. These illustrative examples are set in italics, about 5,000 of them are quotations, and these quotations are indicated the source. These examples are quite helpful for the reader to understand the meaning of the words.

3.4 Labels

AHD5 uses various labels to identify entries that are part of the terminology of specific subjects and entries for which usage is limited to certain geographical areas. Other labels provide guidance regarding various levels of formality and usage. The labels of AHD5 can be organized into the following types: Subject labels, Status labels, Temporal labels, English-language labels and Dialect labels; then the Status labels include Nonstandard, Usage Problem, Offensive, Vulgar, Derogatory, Slang and Informal seven types; the Temporal labels include Archaic and Obsolete. All of the
labels are set in italics and complete spelling, instead of abbreviation, which are very convenient for readers.

3.5 Illustrations

The dictionary text is amplified and made more appealing by its page design, supplemented by more than 4,000 illustrations, most in full color, and usefully put close by the entry they support. Anyone who reads AHD5 is likely to be impressed by the sheer beauty and variety of illustration. The first page of each letter section is highlighted by a full-width illustration of that letter as it has been represented in Phoenician, Greek, and Roman texts, plus background color illustration of its applied use (see figure 3). This is a unique and handsome feature, which makes browsing the pages of AHD5 a pleasure. These attractive full-color photographs and illustrations are fun to look at and make the definitions immediately understandable. For example, the illustration (see figure 4) of the entry brain shows the human brain and its contiguous structures vividly, which is impressive.

![Figure 3 Illustration of letter A](image_url)
3.6 Cross-references

Cross-references have two main functions: to avoid needless duplication of information and to indicate where further discussion of a word occurs. There are many cross-references in the dictionary. The entry referred to is printed in boldface type preceded by a brief descriptive or instructional phrase: bade (bād, bād) v. A past tense of bid. The word See is also used to introduce certain cross-references. Cross-references appear not only in the definition, but in the Synonym Paragraphs and Usage Notes. Sometimes a word is discussed in more than one Synonym Paragraph, cross-references are given to all the Synonym Paragraphs that include this word. If an entry that has a note is discussed in a note at another entry, the cross-reference to that entry is given immediately following the Usage Note. For example, there is a cross-reference at the end of the Usage Note for the entry all: See Usage Note at every.

3.7 Notes

The treatment of notes is another important feature of AHD5, there are valuable entries throughout the dictionary supplying additional information on synonyms, usage, or word history, and our living language. The layout is easy on the eyes, with blue box setting the words apart from their definitions, every entry with a note carries a black symbol in the left margin next to the entry word (see figure 5). The numbers of each type of notes are shown in the table 1:
Figure 5 The layout of the notes

Table 1 The numbers of each type of notes

<table>
<thead>
<tr>
<th>Synonym Paragraphs</th>
<th>Usage Notes</th>
<th>Word History</th>
<th>Our Living Language Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>558</td>
<td>425</td>
<td>259</td>
<td>61</td>
</tr>
</tbody>
</table>

The following is the introduction of these four types of notes:

Synonym Paragraphs: Synonyms of special interest are listed after the entry for the central word in the group. Synonym paragraphs are introduced by the heading SYNONYMS. There are two kinds of synonym paragraphs. The first kind of paragraph consists of fully discriminated synonyms ordered in a way that reflects their interrelationships. A brief sentence explaining the initial point of comparison of the words is given, followed by explanations of their connotations and varying shades of meaning, along with illustrative examples. For example, beautiful has five synonyms: lovely, pretty, handsome, comely and fair, these words are easy to confuse the Chinese learners, but after reading the synonym paragraph, we will see their differences. The second consists of a group of undiscriminated, alphabetically ordered words sharing a single, irreducible meaning. These synonyms are presented in illustrative examples following a core definition. Sometimes the antonym of the entry word will appear at the end of the paragraph. For example, the synonyms of the entry ancestor are: forebear, forefather and progenitor, their core definition is “a person from whom one is descended”, the antonym is descendant.
Usage Notes: The Usage Notes present important information and guidance on matters of grammar, diction, pronunciation, and registers and nuances of usage. In many cases, these notes not only report the percentage of panelists who consider a given usage or construction to be acceptable, but will also report the results from balloting of the same question in past decades, to give a clearer sense of how the language changes over time. For instance, the following is a quotation from the Usage Note of the entry hopefully.

People often warm to a usage once its novelty fades and it becomes well established. But not so with hopefully. Opposition continues to run high or even higher to this usage than it did in the 1960s. In our 1968 survey, 44 percent of the Usage Panel approved the usage. This dropped to 27 percent in our 1986 survey. We asked the question again in 1999, and 34 percent accepted the sentence Hopefully, the treaty will be ratified, while only 22 percent accepted the adverb when placed at the end of a sentence in the example The new product will be shipped by Christmas, hopefully. It would seem, then, that it is not the use of hopefully as a sentence adverb per se that bothers the Panel, since the comparable use of mercifully is acceptable to a large majority. Rather, hopefully seems to have taken on a life of its own as a sign that the writer is unaware of the canons of usage.

Word Histories: In addition to etymologies, which necessarily contain information in a compressed form, AHD5 provides Word History paragraphs at entries whose etymologies are of particular interest. In these notes, the bare facts of the etymology are expanded to give a fuller understanding of how important linguistic processes operate, how words move from one language to another, and how the history of an individual word can be related to historical and cultural developments.

Our Living Language Notes: The Our Living Language Notes in AHD5 demonstrate that American English, like all living languages, is not uniform and static but diverse and dynamic. Geography, social class, ethnicity, gender, and age, as well as factors internal to the language, influence the way particular speakers use and shape American English. The notes discuss exemplary linguistic variations resulting from these factors and provide a broader snapshot of our language than is allowed by traditional dictionary practice.

4. Conclusion

AHD5 has many advantages, its treatment of notes, etymologies and illustrations is very unique. However, as Dr. Johnson once said: “Dictionaries are like watches: the worst is better than none, and the best cannot be expected to go quite true.” AHD5 has its shortcomings as well.

First, the distribution of illustrations is unbalanced. Some margins are placed with five or six illustrations, while others are empty; some illustrations are unnecessary, for example, the illustrations for entry ballet, kiwifruit, boxing and banana. These common things need not to be explained with illustrations.

Second, sometimes the antonym of the entry word will appear at the end of the
Synonym Paragraph, but only the word, without any other information, like pronunciation, definition, example, and so on.

Third, there is a greater than usual tendency to define a word with a form of the same word, for example, fuzzy, whose first two definitions are “1. covered with fuzz.” and “2. of or resembling fuzz.”

Finally, there are several notes whose classification are inaccurate, for instance, what follows the entry assure is a Usage Note paragraph, but the content is about the discrimination between the words assure, ensure and insure, so the author thinks it’s appropriate to indicate the paragraph by SYNONYMS, rather than Usage Note.

But on the whole, AHD5 is still a recommendable reference book with distinctive features, we should learn from its strong points to overcome the shortcomings of college dictionaries in China.

References


Yan Zhang
Shaanxi Normal University
1073704556@qq.com

Abstract

“Usage notes” is another approach to show the interpretation of the headword by many dictionaries. It is an important part in a process of dictionary compiling. *The American Heritage Dictionary of the English Language* covers diversifying lexicon information in “usage notes”, such as pronunciation, spelling, discrimination, grammar, pragmatic properties and so on; “Usage notes” can apply contrast, instruction, error alarm to give the right direction to dictionary users’ on its entries and usages. This paper tries to analyze the represent approach and strategies of “usage notes” in the Fifth edition of *the American Heritage Dictionary of English Language* through contrasting and listing examples. It can be concluded that *The American Heritage Dictionary of the English Language* should increase the number of “usage notes” of functional words to enrich the illustration of lexicons, increase description, and decrease the subjectivism of Usage Panel during compiling “usage notes”. Through analyzing, it summarizes several compiling principles of “usage notes” further, expecting to give the right guidance for dictionary users from pragmatic and semantic perspectives and let users pay more attention to the “usage notes” of lexicons in every dictionary.

Key Words: “Usage Notes”; *The American Heritage Dictionary of English Language*; Represent Approach; Strategy; Compiling Principles

1. Introduction

“Usage notes” is an important part in dictionary compiling, and it is an effective way to let users master the lexicon’s culture, history, and right usage of the native speakers. But it is always neglected by dictionary users in dictionaries, and they just use dictionary to find the single meaning of words, but not to know about the real meaning of words. There are many approaches to illustrate lexicons in other dictionaries such as usage information, register marks, usage labels and so on. It can be seen that usage is necessary for every dictionaries. In other words, “usage notes” play an important role in dictionary using, which refers to the illustration of the headword with the lexical properties(as regional feature, time feature, stylistic feature and so on) in a specific context in dictionary, or makes comment on the features of syntax, structure, and
relation in any approaches (Zhang 2010).

The American Heritage Dictionary of the English Language (AHD) is one of the five college dictionary and the representative one on the “usage notes” as well. From the first edition to the fifth edition of AHD, “usage notes” makes a perfect change in its way. However, there is a little illustration on how to compile “usage notes” at abroad. They just mention it when they talk about grammar, pragmatic or lexical information. Meanwhile, the study of “usage notes” is just from its applying in teaching and learning at home. There is less people noticing the importance of “usage notes” when they are learning at present. Therefore, in order to give dictionary users’ a right cognitive direction and guidance on lexical using, this paper tries to analyze the “usage notes” in AHD of the fifth edition through analysis its represent approach and strategy. And reminding users that a large amount of “usage notes” can help learners distinguish the difference of lexicons and how to use lexicons in a proper way. Therefore, through analysis of “usage notes” it puts forward several compiling principles of “usage notes”, expecting to be used in “usage notes” compiling in the future.

2. Types of Notes in AHD5

The AHD5 contains four types of notes. The usages of the new words with multiple note such as usage notes, word history, synonyms paraphrases, and our living language (AHD 2011). That is to say AHD5 illustrates lexicons from four approaches, and through accounting the number of each in AHD5. They are totally different in number. As in table1.

<table>
<thead>
<tr>
<th>Types</th>
<th>Usage notes</th>
<th>Word history</th>
<th>Synonyms</th>
<th>Our living language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>425</td>
<td>259</td>
<td>558</td>
<td>61</td>
</tr>
</tbody>
</table>

From table 1, the statistics show that the number of usage notes and the other three types’ number. The four types of notes explain lexicons in their way and use their own background from its surface meaning apparently. That is to say “usage notes” describes every use of lexicon; word history involves the etymologies of lexicons and their historical and cultural development; synonyms consists of undiscriminated and discriminated synonyms ordered words; our living language shows the variation and change of words.

AHD5 applies these four notes to explain lexicons in detail, and every note provides a supplement for the center lexicon. It makes the headword comprehend easily and provide an overall perspective of the headword for dictionary users. Thus, “usage notes” takes an important place in AHD5 to illustrate the headword.
3. Background of “Usage Notes” in AHD5

In order to explain lexicons more flexible and specific, dictionary editors compile “usage notes” in many aspects. What’s more, “usage notes” in AHD is mainly compiled by its big Usage Panel. From the first edition to the fifth edition of AHD Usage Panel play an important role in its usage notes compiling. Due to its big Usage Panel the first three edition apply prescriptive method to compile “usage notes”, because it mainly depends on the idea or personal illustration about lexicons of Usage Panel. However, the fourth and fifth edition take more descriptive method, and the prescriptivism decreases relatively. There are some problems in framework of “usage notes” in dictionaries, it needs improvement from linguistics and dictionary should provide an authority, scientific, and right usage, but not just depends on Usage Panel’s judgment(Ottenhoff 1996).

AHD5 has about 10,000 new words and senses(new entries and new definitions), which is greater than the fourth and third edition(AHD 2011). Meanwhile, “usage notes” in AHD5 makes a progress than its first four editions no matter in numbers and in other information.

3.1 The Represent Approach of “Usage Notes” in AHD5

Clearly, due to their different function in dictionary, “usage notes” is the most important form to account for lexicons and it gives interpretation of special lexicons not every words. The number of “usage notes” of every parts of speech is completely different, as in table 2.

<table>
<thead>
<tr>
<th></th>
<th>Noun</th>
<th>Verb</th>
<th>Adj</th>
<th>Adv</th>
<th>Con</th>
<th>Pro</th>
<th>Prep</th>
<th>Pre</th>
<th>Abb</th>
<th>Suff</th>
<th>Def.a</th>
<th>Non</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>148</td>
<td>79</td>
<td>112</td>
<td>36</td>
<td>11</td>
<td>16</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

(Adj.: Adjective; Adv.: Adverb; Conj.: Conjunction; Pron.: Pronoun; Prep.: Preposition; Pref.: Prefix; Abbr.: Abbreviate; Suff.: Suffix; Def.art: Definite art; Non: Nonstandard)

From table 2, it can be concluded that the number of the first four part of speech hold a widely distributed in AHD5, and “usage notes” in AHD5 is mainly concentrated in nouns, verbs, adjectives, and adverbs. The other parts of speech just takes a little part of “usage notes” in this dictionary. That is to say AHD5 also neglects the importance of the functional words as other dictionaries. Taking users into consideration, functional words are also as important as content words when users learn a foreign language. Thus, it should contain “usage notes” of many kinds of parts of speech, especially increasing “usage notes” of the functional words. Only in this way can let readers understand the usage of lexicon more thoroughly and effectively.
Many learning dictionaries set a column of “usage notes”, it mainly provides pragmatic information of lexicons. “Usage notes” makes a statement of dictionary and they are part of innovative approach and their substance implies something about overall quality of dictionary (Berner 1981). Represent approach is a way to strength the impression that lexicon makes on readers, which consists of represent in context, represent in association, represent in semantic, and represent in lexical. “Usage notes” follows many entries presenting important information and guidance on matters of grammar, spelling, lexical distinction, pronunciation, and register and nuance of usage (AHD 2011). Through accounting the numbers of the represent approaches is in table 3.

Table 3 Represent approaches

<table>
<thead>
<tr>
<th>Number</th>
<th>Lexical distinction</th>
<th>Supplementary instruction</th>
<th>Pronunciation</th>
<th>Pragmatic</th>
<th>Spelling</th>
<th>Vocabulary expansion</th>
<th>Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>155</td>
<td>42</td>
<td>100</td>
<td>9</td>
<td>10</td>
<td>49</td>
<td></td>
</tr>
</tbody>
</table>

According to table 3, it can be seen clearly that the emphasis of the “usage notes” in AHD5 and the way it shows to learners. In other words, dictionary users can comprehend lexicon not only from its entries, but from its represent approaches. And the represent approaches can provide a clearly picture of lexicon and it is easy for users to understand the deep meaning of lexicons.

Taking the following as examples:

1. The AHD5 has its own phonetic symbol that is different from IPA. It gives a right and particular pronunciation of lexicons, and shows the source of the pronunciation of it.

2. The pragmatic is a big part in dictionary and how to make it easy for learners to remember is a problem. However, AHD5 has its way to show the pragmatic of lexicon in “usage notes” from the word’s history and composition.

Accessory

Although the pronunciation (a-sēs’a-rē), with no (k) sound in the first syllable, is commonly heard, it is not accepted by a majority of the Usage Panel. In the 1997 survey, 87 percent of the Panelists disapproved of it. The 13 percent that accepted the pronunciation were divided on usage: more than half accepted the (k)-less pronunciation for all sense. A few approved of it only in fashion context, and a few others approved of it only in legal contexts.

(AHD, 2011: 10)
3. Lexicon confusing is constantly appearing in learner’s mind, AHD5 use its “usage notes” to provide a clearly distinction about lexicon confusing and its pragmatic.

4. Learners use dictionary just to check the one or two meaning that they want and always neglect the deep meaning of the lexicon. Even how it comes from. The composition of lexicon is a good way to enrich learner’s vocabulary and AHD5 in its “usage notes” increases the column of vocabulary expansion.

The above examples show the represent approaches from pronunciation, grammar, pragmatic, supplementary instruction, lexical distinction and so on. In a word, it shows that the deep illustration of headword can make the lexicon understand correctly and let dictionary users know how to use the words and what is the inner culture of the words. In another word, To a degree, it can let users know about the native speakers’ culture and background about the headword and pass the emotion, register, attitude, and image information to users.
3.2 The Strategy of “Usage Notes”

Almost of “usage notes” applies in descriptive way, so as to let users use lexicon in a right way. Thus, when editors compile it they utilize many different strategies to give learners directions on usage. It can attract learners interest to read and can fortify learner’s impression. The strategies in AHD5 contain contrast, error alarm, and instruction and so on.

1. Contrast is useful in distinguish the confusion words, taking the following as an example:

Chord
USAGE NOTE the words chord and cord are often confused-and with good reason, for they are really three words, not two. There are two words spelled chord(listed as separate entries with homograph numbers in this dictionary). The first comes from the word accord and refers to a harmonious combination of three or more music notes. The second is an alternation of cord, taking spelling from Greek chorda, “string.gut”, by way of Latin. This is the mathematical chord-a line segment that joins two points on a curve. Cord itself means “a string or rope.” it has many extensions, as in an electrical cord and a cord of wood. When referring to anatomical structures, it can be spelled in general usage either as cord or chord ( again by influence of Greek and Latin). Strict medical usage requires cord, however. A doctor may examine a spinal cord or vocal cords not chords.

(AHD, 2011: 329)

2. Instruction gives a complete definition of lexicons, especially in illustrating the usage of lexicons. Example as following:

Adopt
USAGE NOTE Children are adopted by parents, and we normally refer to an adopted child but to adoptive parents, families, and homes. When describing places, one can use either adopted or adoptive:her adopted city; her adoptive city.

(AHD, 2011: 23)

3. Error alarm is a way to warn learners to use lexicons correctly when they use words in a worry way. Example as following:
4. The Development Tendency of AHD

4.1 The Disadvantages and Advantages of “Usage notes” in AHD5

Through the above analysis, “usage notes” compiling in AHD5 have several disadvantages and advantages. Firstly, in terms of its advantages, AHD5 makes a great improvement in increasing the number of “usage notes”, especially the four content words. The frame of the content of every “usage notes” including its connotation and denotation enriches the meaning of the headword, which makes dictionary users have a right cognition towards lexicons.

Secondly, AHD5 also has its disadvantages in compiling “usage notes”. “Usage notes” in AHD mainly edits by its powerful team-Usage Panel. Thus it takes more subjective and little objective. It can easily make dictionary users understand meaning of lexicon from the Usage Panel’s perspective, not the native speaker’s perspective and it is difficult for users to master the real usage and background of every lexicon.

4.2 Compiling Principle

“Usage notes” in AHD5 basically contains grammar, pragmatic, and lexicon and so forth. Looking back to the other dictionaries, they all do not set a serials of compiling principles and they just depends on their own rules to compile “usage notes” so that “usage notes” have different style and represent approach. Through analysis “usage notes” in AHD5, it shows that a good dictionary should have a set of principles of “usage notes” when they compile a dictionary. Therefore, based on this analysis it puts forward several compiling principles as follows:

First, “usage notes” contains many aspects of lexicon use, thus it should be more descriptive not prescriptive.

Secondly, it needs to distinguish center words and difficult words before compiling.

Thirdly, in order to avoid ambiguity, “usage notes” should use simple words to illustrate the headword.

Fifthly, every “usage notes” should avoid repeating in content and addition.

Fourthly, the content of “usage notes” should be comprehend easily and should be

Also

USAGE NOTE some people maintain that it is wrong to begin a sentence with also. They are probably in the minority, since there seems no reason to condemn also and not another conjunctive adverb like nevertheless. In our 1988 survey, 63 percent of Usage Panel found the usage acceptable in this example: the warranty covers all power train components. Also, participating dealers back their work with a free lifetime service guarantee.

(AHD, 2011: 53)
Therefore, to sum up, “usage notes” can become more systematic according to the principles. It can be seen that the development tendency of AHD becomes better in the future. And the rich content of “usage notes” can attract reader’s interest during their using of dictionary.

5. Conclusion

A good dictionary should deliver many kinds of language contrastive information to learners through all kinds of approaches, which can let them grasp a standard English and use English to express their thought and communication information in a right way. In AHD5, no matter describes from its represent approaches such as context represent, lexical represent, semantic represent or from its strategies such as pronunciation, contrast, caution and so on, it shows a clearly interpretation of headword from its use, history, and distinction. Although AHD5 involves some “usage notes” of functional words, it takes “usage notes” of nouns, verbs, adjectives, and adverbs as an important part. However, In terms of dictionary users, they are not just check the meaning of the four lexical words, functional words also take an important place when they learn a foreign language. Learners need to know more about the right and correct usage of lexicons from cognitive and pragmatic aspects. Therefore, the number of “usage notes” of functional words should be increased and pay more attention to the usage of them in the future edition.

Through analyzing of the advantages and disadvantages, the author summarizes five principles when compiling “usage notes”, expecting to give a direction on the development of “usage notes” in AHD in the future.

References

Defining Emotion Words: How to Bring Them down from Top to Bottom

Guiying Jiang
Xiamen University
gyjiang118@126.com

Xie Xi
Xiamen University
xixie@stu.xmu.edu.cn

Abstract

Cognitive Linguistics provides a theoretical basis for the analysis of cognitive process in understanding emotion concepts. Immense research has proved that conceptual metaphor plays a great role in understanding emotion words. In view of this, this paper aims to investigate the semantic presentation of eight emotion words in the five major English learners’ dictionaries, with a focus on the analysis of the metaphorical conception of emotion expressions and how it has been handled in the entries. It has been demonstrated that definitions of emotion words in dictionaries are vague or circular. While some of them do take into consideration the metaphorical nature of emotion words, they are in lack of sufficiency in featuring this. An alternative cognitive approach is provided in this article, which combines Lakoff’s Conceptual Metaphor Theory, Kövecses’ prototypical emotion scenario and Wierzbicka’ semantic primitives to redefine the emotion words.

Keywords: metaphorical concepts, sense presentation, emotion words

1. Introduction

One phenomenon that has long lingered in our mind is that, while there has never been lack of creative minds who work on pure theories of linguistics, they hardly care about how to put their findings into practice, i.e., how to have them “touch the ground” as is generally put. On the other hand, lexicographers, who are supposed to be part of the applied linguistic circle, seldom seem to have matched the linguistic theories with their lexicographical businesses well enough, if any. When we inadvertently flip the pages of some learner’s dictionaries, it occurs to us that we should check how emotion words have been defined in these dictionaries, with all the contributions made by Cognitive Linguistics (CL) to this field in recent decades. The qualm has only increased as a result.

For the past thirty years, there has been a consensus that certain ‘basic emotions’ are found in all human cultures, which commonly consist of fear, anger, disgust, sadness, joy and surprise. These basic emotions have a distinctive facial expression and for most of them there is evidence of distinctive changes in the voice and evidence of
cognitive phenomena like focusing attention on the emotion stimulus (Griffiths 2003: 39-43). It is also suggested that complex emotions derive from basic emotions. In another word, basic emotions are the building blocks that form parts of more complex psychological states (Griffiths 2003: 39-43; Oatley and Johnson-Laird 2013). Therefore, we’re here to make a close examination of eight emotion words: the four basic emotions “happiness, anger, fear, sadness”, which was first regarded by Darwin as “true and original” forms of human emotional behavior, as well as four complex emotions “desire, embarrassment, jealousy, satisfaction” in the five major learners’ English dictionaries, namely *Oxford Advanced Learner’s Dictionary* (2010, *OALD8* henceforth), *Longman Dictionary of Contemporary English for Advanced Learners* (2009, *LDOCE5* henceforth), *Collins COBUILD Advanced Learner’s English Dictionary* (2014, *COBUILD8* henceforth), *Cambridge Advanced Learner’s Dictionary* (2013, *CALD4* henceforth), and *Macmillan English Dictionary for Advanced Learners* (2003, *MED* henceforth), or the “Big Five”, which have well documented their inadequacy and deficiency in both entry description in general and definition in particular (Ye 1997; Griffiths 2003: 39-67; Oatley and Johnson-Laird 2014).

Meanwhile, recent studies in CL have provided unequivocal evidence that emotion and cognition are closely coupled (Lakoff 2016, 1992; Lakoff and Johnson 1980; Kövecses and Csábi 2014; Kövecses 2000; Dolan 2002). Cognition can be the basis of emotion since cognitive processing underlies all emotional processes (Greenwood 1992; Scherer 2005). CL provides a desirable theoretical framework for the analysis of the underlying cognitive process of understanding emotion words. Cognitive linguists argue that most of our everyday language makes use of metaphor and human thought processes are largely metaphorical. And in the process, everyday abstract concepts like time, change, causation, and emotion are understood metaphorically through conceptual mapping (Lakoff and Johnson 1980; Lakoff 1992; Yu 1995).

CL can be employed as a theoretical basis in lexicography (Jiang and Chen 2015; Kövecses and Csábi 2014). Particularly, metaphor can be used in sense ordering, word patterns and example sentences of emotion words in dictionaries to give a comprehensive and iconic sense presentation of each word. Exploring emotion and its embodied metaphors is an important research (Winkielman et al. 2015). For example, studies have found that the association between emotion and its embodied metaphors by empirically linking (emotion) love with journey and anger with fire (Lakoff and Johnson 1980; Lakoff 1987, 1992; Kövecses 2000; Kövecses and Csábi 2014). However, few studies so far have made a close examination of the application of metaphor in defining emotion words. In view of this, this paper is designed to first investigate how the senses of the aforementioned eight emotion words and their derivatives are presented in the ‘Big Five’. We will then focus on whether and how metaphors are integrated in defining emotion words, with an attempt to answer the following questions:

(1) How do “the Big Five” deal with emotion words in general?
(2) Do the senses, word patterns, collocations and example sentences of each headword
document the conceptual structuring of the human mind? If yes, how are they represented? If not, why?
(3) What can be done to systematize or normalize the defining of emotion words in general in learner’s dictionaries?

2. Literature Review

Emotion is the most common and significant human experience and itself rather abstract, so it is often expressed via metaphor (Zhang 2016). Despite the immense body of research on metaphorical understanding of emotion, few studies have centered on metaphorical semantic representation of basic emotion words in learner’s dictionaries. The current review unfolds along two lines: (1) studies on the relationship between metaphor and emotion; (2) features of emotion words treatment in lexicography.

2.1 Metaphorical expression of emotion

The contemporary theory of metaphor emerged from the research of George Lakoff and Mark Johnson in the 1980s as opposed to the deficiency of classical theory of metaphor since Aristotle, which claimed that metaphor was a poetic imagination and rhetorical device—a matter of language rather than thought. As our emotion experience is as basic as our spatial and perceptual experience, metaphors allow us to conceptualize our emotion in more sharply defined terms by relating it to more concrete objects. For example, LOVE IS A JOURNEY and LOVE IS A CONTAINER (Lakoff 1987; Lakoff and Johnson 1980). To a great extent, orientational metaphor and ontological metaphor are used when we talk about emotion, which is grounded in our everyday experience and exhibits systematicity and coherence (Lakoff and Johnson 1980).

Zoltán Kövecses (2000) identifies emotion metaphors, metonymies, and emotion related concepts. Metaphor consists of a source and a target domain and the source domain is, in everyday case, a better understood and more concrete domain than the target domain. He works on emotion concepts such as LOVE, ANGER, FEAR AND PRIDE. For example, LOVE IS FIRE; LOVE IS A CAPTIVE ANIMAL. While conceptual metonymy which involves a single domain is a “stand-for” relation, such as INCREASE IN BODY HEAT STANDS FOR LOVE. There is a wide range of emotion related concepts that can be placed along a gradient of their centrality in the definition of an emotion concept. He argues that the study of emotion concepts shows there exists a single “master metaphor” for emotion: EMOTIONS ARE FORCES. That is, most metaphors in the emotion domain can be characterized as an interaction of forces, given a great deal of overlap among the metaphors that characterize different emotions, such as anger and love, which share same characteristics, like FIRE, INSANITY, NATURAL FORCE, AND OTHERS. In this sense, he suggests there is a hierarchical organization when we make use of metaphorical conceptualization in the form of
Among all the researchers who believe metaphor plays a great role in the process of emotional conceptualization, Citron and Goldberg (2014) have proposed that metaphor may be more emotionally evocative than their literal counterpart in understanding abstract concepts. This is in line with Lakoff’s claims about the prevalence of metaphor in talking about our affective lives. Yu observes that metaphors behind emotion expressions are based on embodied experience and therefore concludes that culture and body are two important factors in understanding emotion (Yu 2002). Sun (2011) holds that emotion metaphors are the perfect crystallization of human body experiences and the surrounding environment. For example, HAPPY IS UP, JOY IS LIGHT, ANGER IS HEAT. Emotion is also considered to be closely linked with space orientation (I am feeling up), light (His eyes are shining), temperature (Her blood runs cold), fluid in a container (He is drowning in sorrow) and weather conditions (He had a sunny smile) (Gong 2016; Zhang 2016). Even different colors can be the source domain for conceptualizing emotion, e.g., FEAR IS BLACK AND SADNESS IS BLUE (Zhang 2016). A central claim of these studies is that human emotions, which are abstract in nature, are largely conceptualized and expressed via metaphor grounded in bodily experiences.

2.2 How emotion words are handled in lexicography

Psycholinguistic studies suggest that emotion words may be distinct from other abstract words and should be treated as a category separate from both concrete and abstract words (Altarriba and Brown 2011; Dewaele 2002). They are less concrete than concrete words but higher in imageability as compared to abstract words (Altarriba and Bauer 2004). Emotion words require longer recognition time due to the fact that they have less “context-availability” than concrete words, since it is typically more difficult to think of a context in which emotion words appear than it is to think of a context in which a given abstract word appears (Altarriba and Brown 2011). Due to the unique feature of emotion words, special attention and treatment should have been paid to the definition of emotion words.

It is claimed that CL can be a good tool in dictionary making by providing a systematic approach to language, relevant information about a specific concept, as well as its motivation (Kövecses and Csóbi 2014). CL provides a solid theoretical basis for the arrangement of semantic representations of lexicography for it is “trying to find the underlying regularity, or rule, in a (sometimes or only seeming) chaos or randomness” (Zgusta 1992: 92). Rundell (2012) advocates that Conceptual Metaphor Theory be integrated in lexicography and he does put the idea into practice. MED (of which Rundell is the chief editor) presents around 60 ‘metaphor boxes’ with contextualized examples of metaphor sets relating to particular concepts, including emotion concepts, as well as explanations of metaphorical mappings, to help language learners gain a better understanding of the abstract concepts as well as raise their awareness of the metaphorical nature of the language.
3. Examination of the semantic presentation in the ‘Big Five’

In this section, in terms of the features of emotion words and the application of metaphor in defining emotion words, the semantic presentation of the eight emotion words will be examined and compared in the ‘Big five’ from two aspects: (1) how the literal meaning of each emotion word is presented in the entry; (2) how metaphor is integrated in the example sentences, lexical phrases, and collocations. First we will have a close examination of the sense presentation of the four basic emotion words “happiness, anger, fear, sadness” in the Big Five, then come to the four complex ones “desire, embarrassment, jealousy, satisfaction”.

3.1 Examination of the four basic emotion words

3.1.1 Entries of the four basic emotion words

In this part, the entries of the four basic emotion words “happiness, anger, fear, sadness” in the ‘Big Five’ are investigated and some key words and phrases are selected and presented in Table1.

<table>
<thead>
<tr>
<th></th>
<th>happiness</th>
<th>anger</th>
<th>fear</th>
<th>sadness</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDOCE5</td>
<td>The state of being happy; Feel pleasant because sth good happened, appose to sadness</td>
<td>A strong feeling of want to hurt someone or shout at him because they have done sth unfair, cruel or offensive</td>
<td>The feeling you got when you are frightened or worried</td>
<td>The state of being sad</td>
</tr>
<tr>
<td>CALD4</td>
<td>Being happy; feeling of pleasure or satisfaction</td>
<td>A strong feeling that makes you want to hurt someone or be unpleasant because sth unfair or unkind happened</td>
<td>Unpleasant emotion or thought; your frightened or worried because sth dangerous, bad or painful</td>
<td>Unhappy or sorry; not satisfactory or unpleasant</td>
</tr>
<tr>
<td>COBUILD8</td>
<td>Feel pleasant because sth nice happened or feel</td>
<td>A strong emotion you feel when you think that someone has behaved in an</td>
<td>Unpleasant feeling you have when you think you are in</td>
<td>Feel unhappy, unfortunate, undesirable</td>
</tr>
</tbody>
</table>
It is evident from the table that the ‘Big Five’ are similar in defining basic emotion words, mainly by providing synonyms of the headword or indicating under what circumstances people may have this kind of emotion and what kind of behavior people may exhibit under certain emotion. For example, all of them define ‘happiness’ as ‘pleasure, satisfaction’. The ‘big five’ unanimously indicate that the emotion of ‘fear’ may appear when someone feels ‘worried’ or ‘frightened’ or when people think they are in danger. However, words like pleasure and satisfaction in the definition of ‘happy’ are still abstract, and they are neither clearer nor simpler than ‘happy’ itself. There is no doubt that pointing out the circumstance and people’s possible behavior will enable users to conceive a picture of the target emotion in their mind, but this may lead to a misunderstanding of the emotion. As a matter of fact, emotion and behavior vary among different people under certain circumstances and people may exhibit different kinds of emotion and behavior or a mixture of emotions. For example, when in danger, some may be fearful while others may feel desperate or both; some may want to cry while others may just sit silently and wait for the consequence. Thus by pointing out the circumstances or people’s likely behavior under certain emotion can only give a vague impression of the target emotion and may cause misunderstanding since it is inaccurate. Another phenomenon is that dictionary compilers tend to describe emotion as a kind of “state”, such as “happiness” is defined as the state of being happy and “sadness” is defined as the state of being sad in LDOCE5. While according to Frijda (2007), emotion is not a state but a process, and cognition can regulate each of its phases.

Therefore, it is safe to say that the definitions of four basic emotions in the ‘Big five’ are insufficient and even misleading since they resort to either intricate abstract words, which seem to be more difficult to understand than the headword, or inaccurate context in which the target emotion might appear to define an emotion headword.
Metaphor is the conceptual mapping between a concrete source domain and abstract target domain. “The essence of metaphor is understanding and experiencing one kind of things in terms of another” (Lakoff and Johnson 1980: 6). Metaphor is rooted in our physical experience such as: HAPPY IS UP; SAD IS DOWN leads to English expressions like “I am feeling up today”, since drooping gestures typically go along with sadness and depression while erect posture with positive emotion state. Emotion can also be fluid in a container with well-defined boundaries giving rise to expressions like “He is full of happiness.” and “He has been immersed in sadness for a long time”, since human body is an independent entity separated from outside world with boundaries. As Kövecses and Csóbi claim, the existence of conceptual metaphor provides motivation for these expressions to exist on one hand, while metaphor helps us to explain systematically why we use them to convey certain meanings on the other (Kövecses and Csóbi 2014). In this part, based on metaphor theory, the word patterns, collocations and example sentences of the four emotion words, namely, ‘happiness’, ‘anger’, ‘fear’, and ‘sadness’ will be examined respectively.

First, for the semantic presentation of ‘happy’, there is no indication of metaphor given by CALD4, and all that users can get are word patterns, example sentences and idioms about happy, such as, happy marriage; happy childhood; She looks so happy; Happy new year; the happy event (the birth of a child); the happy day (HUMOROUS, a marriage). While for ‘happiness’, there is an example sentence employing metaphorical conception without explicitly pointing out the metaphor: It was only later in life that she found happiness and peace of mind. Here happiness is regarded as a concrete object that can be found by someone, which is an ontological metaphor according to Lakoff. In LDOCE5, apart from example sentences, ‘happy’ and ‘happiness’ are treated with no difference from other words, like the concrete ‘desk’ or the abstract ‘idea’. However, it has been proposed that emotion words are distinct from other abstract words and should be treated as a category separate from both concrete and abstract words (Altarriba and Brown 2011; Dewaele 2002). They are less concrete than concrete words but higher in imageability as compared to abstract words (Altarriba and Bauer 2004). There is a Collocation Box where a list of phrases of the headword is presented with their corresponding equivalent phrases and example sentences like great happiness (= a lot of happiness, His grandchildren bring him great happiness). Altogether there are 17 happiness-related phrases listed in Collocation Box, and among them, 11 phrases involve metaphors, but none of the metaphorical phrases are distinguished from the rest of them. For example, in true/real happiness, ‘happiness’ is taken as a concrete entity that has quality, and it can be real or fake; in achieving happiness, ‘happiness’ is regarded as a destination that can be reached or achieved through effort; in be filled with happiness, ‘happiness’ is viewed
as fluid in a container. All these suggest pervasiveness of metaphor in expressing emotion. Similar cases appear in COBUILD8, and OALD8, with a great number of phrases, collocations simply listed without explicate account of the pervasiveness of metaphorical conception integrated in those expressions, nor are they so arranged that their sense relations can be recognized more easily.

Excepting the literal explanation and some example sentences, by contrast, the semantic presentation of ‘happy’ in MED is quite different from that in the other four dictionaries, as the employment of metaphors in the definition is explicitly pointed out. As is stated in the Metaphor Box in MED under the headword ‘happy’, ‘Feeling happy and hopeful is like being high up or like moving upwards. Feeling sad or unhappy is like being low down or like falling.’ Some related examples with the use of metaphor are also explained, e.g., The news lifted her spirits. Cheer up! He’s in low spirit. Such arrangement, with both literal meaning and metaphorical expression of emotion words presented in the definition will give the user a more comprehensive understanding of emotion words and raise their awareness of the metaphorical nature of language.

Second, the semantic presentation of ‘anger’ is investigated in the ‘Big Five’. In CALD4, apart from the examples for ‘anger’, there is a Pattern Box where users can get 17 word partners for the headword ‘anger’, as shown in the following picture (CALD4).

It is evident that most of the word partners for ‘anger’ shown in the excerpt are metaphorical expressions although there is no mention of metaphor at all in the dictionary, like show your anger, arouse/ provoke/spark anger, growing/mounting/widespread anger, in anger, etc. According to Kövecses, the notion anger, as well as other emotion words, has a rich system of conceptualization which includes five stages: offending event (arouse/provoke/spark anger), anger, attempt to control (pent-up/simmering/suppressed anger), loss of control (growing/mounting/widespread anger), and finally act of retribution (express/voice/show your anger) (Kövecses 2000:127-129). According to Lakoff (1980), STATES ARE LOCATIONS, we get the metaphorical word pattern, do sth in anger, in the above box. Compared with CALD4, LDOCE5 lists more collocations of ‘anger’, but it also fails to explain the underlying metaphorical concepts. There are 32 collocations of ‘anger’ listed in the Collocation Box, such as, be filled with anger; be full of anger; be shaking with anger; control your anger, fuel anger, etc., and 27 of them employ metaphors. The same holds true for the treatment of metaphor in COBUILD8 and OALD8, but there are fewer word patterns listed in COBUILD8 and OALD8, compared with those in CALD4 and LDOCE5. Users can only get some more frequently used expressions like anger at sb/sth in COBUILD8 and OALD8.

As for the semantic presentation of ‘anger’ given by MED, it is stated that ‘Being angry is like being hot on fire. Losing your temperature is like an explosion.’ Some
examples involving metaphors are offered, such as: *She burns with indignation; He has a fiery temper; When she told him, he nearly exploded; He blew up at her, etc.* The statement of metaphorical nature combined with related examples will definitely be clear and impressive to users.

Third, for the semantic presentation of ‘anger’ and the treatment of metaphor in expressions related to the headword, all five dictionaries fail to reveal the underlying conceptual metaphors. What users can get are simple lists of word partners or collocations of ‘fear’ combined with example sentences, such as, “overcome fear, I manage to overcome my fear of spiders” (*CALD4*). “without fear, people must be able to express their view without fear” (*LDOCE5*). “The elderly live in fear of assault and murder” (*COBUILD8*) and etc.

Fourth, the metaphorical nature of expressions related to ‘sadness’ is mentioned only by *MED*, while the other four dictionaries only give detailed semantic presentations and example sentences without a touch of metaphorical awareness.

The above analysis of the semantic presentation of four basic emotion words in the ‘Big Five’ shows that only *MED* explicitly points out the metaphorical conception in the expression of three emotion words, namely ‘happiness’, ‘anger’ and ‘sadness’ and leaves ‘fear’ behind, while the other four dictionaries list many word patterns and collocations which are metaphorical in nature but fail to inform users of this. By comparison, *MED* does a better job than the others for its awareness of metaphor, which helps the user to gain a better understanding of emotion concepts, but it is far more than satisfactory for its lack of consistency in defining emotion words.

### 3.2 Examination of the four complex emotion words

#### 3.2.1 Entries of the four complex emotion words

In this part, the entries of the four complex emotion words “desire, embarrassment, jealousy, satisfaction” in the ‘Big Five’ are investigated and some key words and phrases are selected and presented in Table 2.
### Table 2 Entries of each complex emotion word presented in the “Big Five” (key words and phrases)

<table>
<thead>
<tr>
<th></th>
<th>desire</th>
<th>embarrassment</th>
<th>jealousy</th>
<th>satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LDOCE5</strong></td>
<td>A strong hope or wish; want sth very much</td>
<td>The feeling you have when you feel embarrassed. Embarrassed: feeling uncomfortable or nervous and worry about what people think of you</td>
<td>A feeling of being jealous. Jealous: feeling unhappy because someone has sth you wish you have</td>
<td>The feeling of happiness or pleasure because you have achieved sth or got what you wanted</td>
</tr>
<tr>
<td><strong>CALD4</strong></td>
<td>The strong feeling that you want sth</td>
<td>The feeling of being embarrassed. Embarrassed: feeling ashamed or shy</td>
<td>A feeling of unhappiness and anger because someone has sth that you want</td>
<td>The pleasant feeling which you get when you receive sth you want</td>
</tr>
<tr>
<td><strong>COBUILD8</strong></td>
<td>A strong wish to do or to have sth</td>
<td>The feeling you have when you are embarrassed. Embarrassed: feels shy, ashamed or guilty about sth</td>
<td>The feeling of anger or bitterness which someone has when they think that another person is trying to take sth away from them</td>
<td>The pleasure that you feel when you do sth or get sth that you wanted or needed to do or get</td>
</tr>
<tr>
<td><strong>MED</strong></td>
<td>A strong feeling of wanting to have or do sth</td>
<td>A feeling of being nervous or shamed because of what people know or think about you</td>
<td>A feeling of anger and sadness because sb has sth that is better than you have</td>
<td>The feeling of pleasure that you get when you achieve or obtain sth you want</td>
</tr>
<tr>
<td><strong>OALD8</strong></td>
<td>A strong wish to have or do sth</td>
<td>Shy, awkward or guilty feelings; a feeling of being embarrassed. Embarrassed: shy, awkward or ashamed, especially in a social situation.</td>
<td>A feeling of being jealous. Jealous: feeling angry or unhappy because sb you love or like is showing interest in sb else.</td>
<td>The good feeling you have when you have achieved sth or when sth you wanted to happen does happen.</td>
</tr>
</tbody>
</table>
In the above table, it is evident that the way those complex emotion words are defined falls into the same pattern as that of the basic emotion words, which is vague and inaccurate, to some degree. Similarly, neither pointing out the possible circumstances in which the target emotion might appear nor resorting to other emotion words, which is also abstract, can give user a clear understanding of the sense of the target emotion headword. For example, the ‘Big Five’ anomalously use words like “hope”, “strong wish”, “strong feeling” to define “desire”; while words like “unhappiness”, “anger”, “bitterness”, and “sadness” are used to define “jealousy”. Another failure is that dictionary compilers fall into the trap of circularity by using the adjective form of a noun emotion headword to define that noun, such as, “embarrassed” for “embarrassment”, “jealous” for “jealousy”, and resort to other emotion words to define the adjective form of the emotion headword. For example: the definition of “embarrassment” in LDOCE5 is that The feeling you have when you feel embarrassed, and “embarrassment” is defined as feeling uncomfortable or nervous and worry about what people think of you. The same goes with CALD4, COBUILD8 and OALD8, while in MED, “embarrassment” is directly defined by other emotion words as being nervous or shamed.

3.2.2 Word patterns, collocations and example sentences of the four complex emotion word.

As a matter of fact, the treatment of the four complex emotion words in the ‘Big Five’ makes no difference from that of other words (say concrete words). And the ‘Big Five’ are similar in defining emotion words, i.e., simply presenting the meaning of each word with an example sentence; a list of word partners and collocations, although the CMT is sometimes involved in the definition. Take the definition of “jealousy” given in CALD4 as an example. The example sentence she was consumed by/eaten up with jealousy (=She was very jealous) when she heard that he had been given a promise. There is an ontological metaphor in this sentence. ‘Jealousy’ is taken as a concrete object that can be consumed and quantified.

From the above analysis, it is clear that dictionary definitions of emotion words are vague and circular. There is no doubt about the difficulty in defining emotion concepts (Ostermann 2012). A description of language in dictionaries based on cognitive linguistics would therefore be more realistic and more tangible (Geeraerts 2007; Ostermann 2012). The semantic description of emotion words offered by linguistics, however, stands in sharp contrast to that given by dictionary compilers. The semantic description of emotion terms has received a fair amount of cognitive linguistic attention (Ostermann 2007). Kövecses (2000) puts the role of metaphors in the centre of the description of emotion words. In order to render emotion words more accurately and provide users with a more illuminating definition of emotion words, a new defining approach is needed.
4. A cognitive approach to defining emotion words

In order to render emotion terms unambiguously and, so to speak, “cognitively” in dictionary definitions, a new defining approach is needed (Ostermann 2007). Based on the metaphorical nature of emotion expressions, this new cognitive approach is developed to deal with the literal meaning (or meaning description) as well as the arrangement of word patterns and collocations of each emotion.

The literal meaning is designed by combining Kövecses’ (2000: 127-129) prototypical emotion scenario and Wierzbicka’s (1992: 541) semantic primitives to redefine emotion words. Kövecses’ prototypical emotion scenario is a five-stage model consisting of cause of emotion, the emotion, an attempt at control, a loss thereof and a response (Kövecses 2000: 129), which regards emotion as a process integrated with prototypical instances of emotion experience rather than something static, and filled with cognitive linguistic information on emotion words. Wierzbicka describes emotion terms with the help of semantic primitives such as ‘do’, ‘good’, ‘bad’, ‘happen’, ‘know’, which are universal in language since they are non-technical and can be understood intuitively and which provides a framework that all emotion concepts can be clearly and rigorously portrayed and compared (Wierzbicka 1992: 542). This new cognitive approach for literal meaning representation will therefore give the user a panoramic and thorough understanding of the emotion word.

As for the word patterns and collocations, i.e., the metaphorical expressions of each emotion headword randomly distributed currently in dictionaries, they will be arranged systematically since metaphorical language is merely a surface manifestation of conceptual metaphor. Conventional metaphorical expressions are governed by conceptual metaphor in a systematic way, and people use metaphorical expressions systematically (Lakoff and Johnson 1980; Lakoff 1992). Ning Yu also argues that “metaphorical expressions are systematically tied to a conceptual metaphor, with each of the former as a particular linguistic instantiation or manifestation of the latter” (1998:16). For instance, metaphors for the emotion of anger should be understood in terms of the central metaphor that Anger is body heat and internal pressure, and we have a large group of conventionalized expressions related to the emotion of anger, such as, burst with anger, seething with anger, flush with anger, control/contain/suppress one’s anger, and etc. The central metaphor is a really productive one in that it forms the basis of all these linguistic expressions that are metaphorical or metonymical in nature (Lakoff 1987). Therefore, presenting those word patterns and collocations systematically by rearranging them according to their central metaphors will no doubt give the user a clearer and better understanding of emotion concepts.

Following this new defining approach of literal meaning presentation and rearrangement of word patterns and collocations of each emotion word, this definition fits the conceptual structuring of the human mind to a large extent. The following table lists the definitions for the eight emotion words:
<table>
<thead>
<tr>
<th>headword</th>
<th>cognitive definition</th>
</tr>
</thead>
</table>
| happiness | The good feeling you have when sth you want happens; you are in good mood and may smile.  
The major conceptual metaphor for the notion of happiness is orientational: HAPPINESS IS UP (Lakoff & Johnson, 1980; Kövecses, 1991). Under this conceptual metaphor some conventionalized expressions are: *cheer up; high spirit; boost one's spirit; flying high; on cloud nine; soaring with happiness; walking on air; be floating and etc.* Another major metaphorical concept of happiness in English is HAPPINESS IS LIGHT, under which the metaphorical expressions are, for instance: *When she heard the news, she was gleaming; she lit up; sunny smile; shining with joy, and etc.* |
| anger | The bad feeling you have when you think sth or sb’s behaviour is unfair or conflicts with your idea. It might be hard to suppress this bad feeling, so you might act in an unfriendly manner or shout at others.  
The conceptual metaphor for anger: ANGER IS HEAT (Lakoff, 1987). Under this conceptual metaphor, when ANGER IS HEAT is applied to solids, the version of metaphor is ANGER IS FIRE. There is a large group of conventionalized expressions, such as: *fuel anger; widespread anger; outburst of anger; burn with indignation; flare up; lose one's cool, and etc.* In the second, the fire is burning outside the body, the container, heating it and raising the temperature inside. The image metaphor is the ANGER IS HEAT: ANGER IS A HOT FLUID IN A CONTAINER (Yu, 1998). For instance: *seething with anger; reach the boiling point, make sb’s blood boil; He flipped his lid; He blew his top and etc. The destructive force of anger is obvious: when the hot fluid is producing too much steam in the closed container, the steam has to find or force its way out; otherwise it will cause explosion. For instance: contain/ control/ hide/ express/vent one’s anger; burst with anger; He nearly exploded; anger outburst, and etc.* |
| fear | The bad and helpless feeling you have when you think sth bad is going to happen or you are in danger. It makes you tremble and shiver, and you may want to run away or need others’ help.  
The conceptual metaphor for fear: FEAR IS FLUID IN A CONTAINER (Sun, 2011). There are some linguistic expressions, such as, *in fear of; full of fear; another conceptual metaphor: FEAR IS AN ENEMY. For instance: conquer/overcome your fear; shake/tremble with fear; be gripped by fear; be paralyzed with fear; ease/dispel one’s fear, and etc.* |
sadness

The bad feeling that brings you down when you lose sb you like or sth has happened that damages your interest; it throws you in bad mood and might make you cry.

The major conceptual metaphor for the notion of sadness is orientational: SADNESS IS DOWN. For instance: deep sadness; low spirit; fall into depression; immerse in sadness. Another conceptual metaphor for sadness: SADNESS IS DARKNESS. For example, in a dark mood.

desire

You like sth very much but it does not belong to you, then you have the strong feeling to get it; you may give up or work for it.

The major conceptual metaphor for the notion of desire is ontological. People talk about desire as a concrete object that can be shown and quantified. For instance: express/show a desire; have no desire for sth/to do sth and etc.

embarrassment

The uncomfortable feeling you have when you did sth wrong and you care how people think about you. You may regret what you have done and your face may turn red.

The major conceptual metaphor for the notion of embarrassment is ontological. Embarrassment can be taken as a concrete object that can be quantified. For example, The court case is a huge embarrassment to the company.

jealousy

The bad feeling you have when you think others are better than you and they do not deserve that; you may become angry at those people and wish sth bad happened on them.

satisfaction

The good feeling you have when you achieve sth great or when you finally get what you want, and therefore you are proud of yourself and feel life is good.

The major conceptual metaphor for the notion of satisfaction is ontological. People talk about desire as a concrete object that can be shown, obtained and quantified. For instance, get/gain no/great satisfaction from; find satisfaction in sth; take satisfaction in/from sth; sth gives/brings sb satisfaction; a source of satisfaction, and etc.

5. Discussion

Some implications can be drawn from this investigation. It is evident that dictionaries to some extent fail to give a scientific and comprehensive definition of emotion words, both the basic and the complex, from cognitive point of view. On the one hand, emotion words are explained by their synonyms which are still abstract and obscure to users and this may be caught in circular explanation, and the possible context given by dictionary, in which the target emotion may appear, is neither accurate nor clear enough for users. On the other hand, either the neglect of metaphorical conception or
incomplete mention of underlying metaphorical nature in structuring emotion expressions will no doubt result in inefficiency and inconsistency of a dictionary in defining emotion words.

There are mainly two reasons accounting for the problems in the semantic presentation of emotion words in dictionaries. First, emotion words themselves should be to blame. Emotion is subtle and unpredictable, and this leads to the unique features of emotion words which appear hard to define and measure (Oatley and Johnson-Laird 2014). Thus it is significant that metaphorical conception should be made clear in dictionaries and facilitates a better understanding of emotion words. Second, compilers’ lack of consciousness to draw on CL achievements in dictionary making results in the ignorance of the role of metaphor in emotion expressions. Dictionary makers prefer to organize senses according to the frequency of use, which is determined on the basis of corpus analysis rather than human cognitive patterns.

Therefore, compilers need to give CL their due attention, and combine it with statistics from corpus to make dictionaries suit users’ needs better. As Kövecses and Csábi (2014) suggest, learners who know how certain conceptual metaphors structure the meaning of words will acquire the meaning of these words more easily than learners who are not aware of (or unfamiliar with) the mechanism (2014). Therefore, a new cognitive defining approach is proposed, which combines Kövecses’ (2000: 127-129) prototypical emotion scenario and Wierzbicka’s (1992: 541) semantic primitives to redefine the literal meaning of each emotion words, as well as arranges the metaphorical expressions systematically according to their underlying conceptual metaphor, to offer a simpler and more comprehensive sense presentation of emotion words.

6. Conclusion

Based on Conceptual Metaphor Theory (CMT), this paper investigates the semantic presentation of eight emotion words in the ‘Big Five’. The investigation shows that the dictionaries more or less fail to give a clear and comprehensive definition of emotion words, owing to the negligence of their metaphorical features. The literal meanings of emotion words given in dictionaries are largely vague and circular, and the application of CMT in semantic presentation is random and arbitrary. The underlying reasons have been explored. The nature of emotion words is to blame, along with negligence from the compilers. Therefore, a new cognitive defining approach is offered to deal with emotion words and facilitate users to better understand emotion words.
References

A. Dictionaries


B. Other Literature

Cambridge University Press.


Abstract

Metaphor has been the focus of cognitive linguistics, psycholinguistics, applied linguistics, corpus linguistics, and metaphor identification lays a solid foundation for metaphor research. Since Lakoff and Johnson (1980) proposed the Conceptual Metaphor Theory, much attention has been given to the conceptual and cognitive dimensions of metaphor, leaving linguistic dimension secondary. However, when MIP was introduced in 2007, which aims to identify metaphorically used lexical units in natural discourses, metaphor researchers have developed a systematic and reliable methodology for identifying linguistic metaphor instead of working with institution and subjective criteria, which enables them to focus their research on different levels—linguistic forms, conceptual structure and cognitive processing. As MIP requires metaphor analysts to work through 4 steps, in which they depend heavily on dictionaries to determine lexical units and specify the basic and contextual senses, the use of dictionaries becomes the critical element in MIP. The Pagglejaz Group chose Macmillan English Dictionary for Advanced Learners as reference tool, while MIPVU, the elaborated version of MIP, used Longman Dictionary of Contemporary English and Oxford English Dictionary apart from MED. The author, by demonstrating the use of different types of dictionaries in MIP, tries to show that together with learners’ dictionaries, historical dictionaries, collocation dictionaries and specialized dictionaries can also be used for cross reference to guarantee the reliability of linguistic metaphor identification in MIP.

Keywords: metaphor identification; dictionary, linguistic metaphor, sense

1. Introduction

Since Lakoff and Johnson published Metaphors We Live By in 1980, metaphor research has become the focus of cognitive linguistics, psycholinguistics, applied linguistics, and corpus linguistics. With the application of corpora, people begin to emphasise the difference between grammar and specific usage of a language in their research (Steen, 2007), and accordingly, large corpora are used to facilitate metaphor research related to specific contexts, in which metaphor identification becomes a pressing issue (Krennmayr 2013). Metaphor research can be approached from two perspectives: linguistic metaphor and conceptual metaphor, and since MIP was introduced in
2007 (metaphor identification procedure, Pragglejaz Group, 2007), the identification of linguistic metaphors has attracted more attention than ever before. In MIP, the most crucial part is the contrast between the basic meaning and contextual meaning of lexical units, and since “a meaning can not be more basic if it is not included in a contemporary users’ dictionary” (Steen et al., 2010:35) and 99% metaphorical usages from native speakers can be found in dictionaries of contemporary English (Steen 2011), using dictionaries, usually learners’ dictionaries, becomes the key factor in metaphor identification based on MIP. The author, by demonstrating the use of dictionaries in MIP, tries to show that that not only learners’ dictionaries, but also historical dictionaries, collocation dictionaries and specialized will help researchers make relatively consistent and objective judgment in linguistic metaphor identification.

2. MIP: a bottom-up approach in metaphor identification

Early in the metaphor study, metaphor identification usually relied on the researchers’ intuition, such as Lakoff and Johnson’s research (1980). Later on, the development of corpus linguistics enabled researchers to get rid of the dependence and search for a relatively unified and effective standard in metaphor identification. Presently, the metaphor identification mainly includes two kinds of approaches: top-down and bottom-up approaches. The former presents conceptual metaphors, then retrieves corresponding linguistic metaphors from the text, while in the latter no conceptual metaphors are presumed and researchers try to derive mappings from linguistics expressions which they identify as metaphorically used. The identification of metaphor based on the application of the basic principles and methods of corpus linguistics in essence can be categorized as the top-down approach, but in recent years, people come to realize the limitation of this deductive research method: there are no standard procedures to identify conceptual metaphors, and researchers have to rely on their intuitions to a great extent. At present, more and more researchers prefer the bottom-up approach, and MIP and its upgrade MIPVU (Metaphor Identification Procedure at VU University level, Steen et al., 2010) are employed as a typical bottom-up approach (for convenience both are referred to as MIP). In MIP a language unit can be divided into metaphorical and non-metaphorical expressions, and once the semantic consistency is destroyed by introducing the conceptual meaning of a different domain, the language unit can identified as a metaphorical expression.

MIPVU, a revised version of MIP, has made a further improvement in metaphor identification. It extends metaphors to similes and implicit metaphors so there are three types of metaphors in MIPVU: indirect metaphors, direct the metaphors and metaphor indicators, eg, in the sentence the marriage is a trap, “trap” is an indirect metaphor; He eats like a pig, “pig” an direct metaphor, while like, as, compare, etc. are metaphor indicators. Moreover, the lexical unit in MIPVU is refined to its part of speech rather than lemma in MIP. In addition to the Macmillan English Dictionary for Advanced Learners (henceforth MED), the reference in MIP, MIPVU also refers to Longman Dictionary of Contemporary English (henceforth LDOCE) and Oxford English
Dictionary (henceforth OED) for help. Perhaps the biggest difference between MIPVU and MIP lies in the fact that in MIPVU it’s not enough to make contrast between the basic meaning and context meaning to identify metaphors, but the semantic references of the two concepts have to demonstrate similarity in the external or function. To a certain degree, MIPVU provides more comprehensive, objective criteria in metaphor identification than MIP.

The specific steps in MIP are as follows:
1) read the entire text–discourse to establish a general understanding of the meaning.
2) determine the lexical units in the text–discourse
3) (a) for each lexical unit in the text, establish its meaning in context, that is, how it applies to an entity, relation, or attribute in the situation evoked by the text (contextual meaning). Take into account what comes before and after the lexical unit.
   (b) for each lexical unit, determine if it has a more basic contemporary meaning in other contexts than the one in the given context. For our purposes, basic meanings tend to be
   —more concrete [what they evoke is easier to imagine, see, hear, feel, smell, and taste];
   —related to bodily action;
   —more precise (as opposed to vague);
   —historically older;
   (c) if the lexical unit has a more basic current–contemporary meaning in other contexts than the given context, decide whether the contextual meaning contrasts with the basic meaning but can be understood in comparison with it.
4) if yes, mark the lexical unit as metaphorical

---

Figure 1 Process of identifying metaphorically used words (Pragglejaz 2007)
As is shown, MIP procedure can only identify linguistic metaphors, restraining from presuming conceptual metaphors. Unlike top-bottom approach, the five-step method of MIP, which restricts itself to dealing with comparing and contrasting meaning as defined in the dictionaries (Steen 2007), helps researchers to get rid of dependence on their intuitions with comparatively reliable basic meaning and contextual meaning of the lexical unit coded in dictionaries. Moreover, as the comparison between the basic meaning and contextual meaning strictly follows dictionary definitions to determine metaphorically used words, providing the basis from which cross-domain mappings are constructed, MIP, with its focus on linguistic metaphors, prevents researchers from “seeing concrete manifestations of conceptual metaphors everywhere” (Steen 2007:27).

3. The use of dictionaries in MIP

MIP strictly adheres to standard English dictionaries to determine the lexical unit and compare and contrast its basic meaning and contextual meaning, so the importance of dictionaries can never be overestimated. As MIP is targeted on formal, contemporary, standard British English (Steen 2007), dictionaries based on a large, general and contemporary English corpus are preferred, mainly learners’ dictionaries, though not restrained to them.

3.1 Use of learners’ dictionaries in MIP

Pragglejaz Group chose learners’ dictionaries for the following reasons: First of all, most contemporary English learners’ dictionaries, with no exception, are compiled based on large corpora of contemporary English from different discourses: MED was compiled based on a systematic processed corpus of 220 million words, large enough to provide a number of citations for all but the rarest words, and LDOCE, the Longman Corpus Network, a 330 million word database. Hence they are considered adequate for general language analysis and can fully satisfy the need for metaphor research (Pragglejaz Group 2007). Secondly, unlike dictionaries compiled for native speakers, in learners’ dictionaries special consideration is given to high-frequency words with exquisite sense divisions, precise definitions, typical examples and collocations. Words like say, see, light and grasp etc. create few difficulties for natives but for non-native speakers, and it is high-frequency words rather than difficult or rare words that pose serious problems when they try to differentiate the literal meaning and metaphorical meaning, hence the dictionaries are heavily used in MIP, especially in step 3.

1) definition

Sometimes not only sense division, but also definitions will help distinguish the basic meaning and contextual meaning of a lexical unit. In the following example, the first sense of “embrace” is related to body action, which is concrete, and the second sense, with abstract collocates “idea”, “belief” and “opinion” etc. will make it a direct
metaphorical sense.

[1] Community standards may **embrace** moral principles or they may not.\(^1\)

**embrace**
MED: to completely accept something such as a new belief, idea, or a way of life
[sense 2a]
LDOCE: to eagerly accept a new idea, opinion, religion etc. [sense 2]

2) collocation information

Most contemporary learners’ dictionaries allocate considerable space to collocation information as it presents the way a word is used in specific context and with its collocates, we can decide on its meaning, even when the definition is not sufficient to make a judgment.

[2] *He turned round and directed a torrent of abuse at me.*

The word “torrent” in MED has two meanings: The first, related to water flow, can be taken as the basic meaning, and the second referring to “a large amount of something, especially something unpleasant” may or may not be deemed to be abstract as “something” is ambiguous though the word “unpleasant” may, to a certain extent, indicate its abstractness. However, if we turn to the highlighted collocation pattern “of a torrent of abuse/words/criticism” in MED for its second sense, we can be fairly assured that “something” is abstract and “torrent” in the second sense most probably relates to contextual meaning.

3.2 Use of historical dictionaries

The core issue in using MIP to identify metaphor is and above all whether the two senses are listed as two separate, numbered sense descriptions in the dictionary. Though it’s believed “the overwhelming majority of cases can be solved by using the Macmillan dictionary, and the Longman dictionary as a second opinion when it is needed”(Krennmayr 2008:107), it is not rare at all that information provided in learners’ dictionaries is insufficient for researchers to determine the basic meaning and contextual meaning. For pedagogical purposes, in learners’ dictionaries senses are sometimes collapsed and subtle meanings are ignored (Steen 2007; Deignan 2005). And to make things worse, for the target readers, the most frequently used sense of a word would appear as the first sense when its historical development is usually disregarded, which will attribute to the disagreement among researchers concerning the basic sense. Should it occur, information provided in learners’ dictionaries will not be sufficient for researchers to make an objective judgment, especially when two meanings are subsumed into one sense description or one of the senses is missing in

---

\(^1\) All examples are from BNC corpus in this paper.
the dictionaries.

1) sense conflation

Here are some examples:

[3] *It would use new methods to teach traditional academic subjects and equip young people with technical skills.*

Our intuition tells us that “use” in “use a method” is different from the one in “use a tool”. However, if we consult learners’ dictionaries, we will find:

**use** v.
MED: to do something using a machine, tool, skill and method etc in order to do a job or to achieve a result [sense 1]
LDOCE: if you use a particular tool, method, and the service, ability etc, you do something with that tool, by means of that method etc, for a particular purpose [sense 1]

We will fail to make a distinction between the basic sense and contextual sense as the literal sense and abstract sense are conflated in the exampled sentence, so if we adhere to MIP the word is not metaphorically used, which is against our intuition. Nevertheless, if we turn to OED, a historical dictionary, we will see:

**use** v.
OED: II. to put to practical or effective use; to make use of, employ, esp. habitually.
From the 20th cent. some senses in Branches I. and III. (e.g. senses 3c, 6, and 16) have increasingly been understood *instrumentally* as implying particular ends or purposes, even when there is no explicit context of that kind; as a result these uses have converged on the senses in this branch (highlighted by the author)

a. to put (an instrument, implement, etc.) to practical use; *esp.* to make use of (a device designed for the purpose) in accomplishing a task. [sense 8a]
c. to make use or take advantage of (a quality, condition, idea, or other immaterial thing) as a means of accomplishing or achieving something. †Formerly also *intr.* with *of*, (occas.) *with*. [sense 10 ]

OED makes a segment between “use” related to material things and immaterial things, but citations in OED show that the two different usages occurred nearly at the same time in middle English(about c1300,c=circa) so the concrete usage is not historically older, neither can the semantic relationship be found between the two. If we apply the criterion of MIP, “use” in the example [3] is not metaphorically used.
2) sense omission

As mentioned in previous part, two separate sense descriptions for a lexical unit are considered as a precondition for contrast between the basic meaning and contextual meaning, however, due to the restricted space, it’s quite possible that there will be only one sense, usually the most frequently used one listed in learners’ dictionaries, while actually there are more than one. Let’s see “fervent” and “ardent” in the example [4] and [5]:

[4] There were fervent arguments both for and against gun control.
[5] Even his most ardent supporters disagreed with this move.

In MED and LDOCE, both “fervent” and “ardent” have only one sense, which describes emotion, but in OED, besides the one associated with emotion, they both have meanings referring to temperature, with which we can feel confident about their metaphorical usage in the example [4] and [5]:

fervent: hot, burning, glowing, boiling[sense 1]
ardent: burning, on fire, red-hot; fiery, hot, parching[sense 1]

Actually, apart from sense division, the etymological information provided in OED also helps us make judgment: It shows that both “fervent” and “ardent” have Latin origins when Latin “fervent” meant “boil”, “glow” and “ardere” meant “to burn”, which supports their metaphoricity.

As most learners’ dictionaries are based on descriptivism and draw data from corpus, and little consideration is given to etymological information (though in CD-ROM etymology may be provided). On the other hand, in learners’ dictionaries frequency is taken as priority in sense arrangement, and people tend to accept the most frequent sense, usually the first one as the basic one, even though it is not necessarily related to its basic meaning (Pragglejaz Group, 2007). To avoid the misjudgment, a historical dictionary becomes a valuable resource in MIP, especially in finding the basic sense:

[6] I'll just leave the engine running while I go in.

The highest frequency usage of “leave” is “to go away from a place or a person”(LDOCE sense 1), and most probably, it may be taken as the basic sense. But OED tells us that “leave” was originated from the Old English “bequeath”, meaning “allow to remain and leave in place”, and still earlier, from German “bleiben”, meaning “remain”, so, the basic meaning should be “to let something remain in a particular state, position, or condition” (LDOCE sense 5) rather than its first sense, and when we compare and contrast the basic meaning and contextual meaning in example [6], the conclusion can be drawn that “leave” is not metaphorically used.

Moreover etymological information is especially useful for determining the basic
meaning of culture-loaded words:

[7] *The students’ rooms are spartan but clean, with no carpets or central heating.*

**spartan**

MED: very plain and simple, without the things that make life comfortable and pleasant  
LDOCE: spartan conditions or ways of living are simple and without any comfort

Only one sense can be found in MED and LDOCE for “spartan”, so if we use the criterion of MIP, “spartan” in [7] is not metaphorically used. However, as a culture-loaded word, its cultural connotation makes it a direct metaphor, and most researchers take the origin or cultural background information of culture-loaded words as their basic meanings (Dorst & Kaal, 2012; Schmitt 2005). In this case, LDOCE in its CD-ROM provides the etymology information of “spartan” as follows: “of Sparta (16-21 century) from Sparta city in ancient Greece whose people lived simply”, which is more than enough for researchers to decide on it metaphorical nature.

### 3.2 Use of collocation dictionaries

Similar to learner’s dictionaries, most contemporary collocation dictionaries are compiled on the basis of large, contemporary, general corpora, and *Macmillan Collocation Dictionary* (henceforth MCD), makes a good choice for identifying metaphors. As one of the most distinguished collocation dictionaries with its unique structure, MCD chooses the high frequent collocations, often associated with the metaphorical meanings of headwords rather their basic meanings, offering help to identify metaphors from following perspectives:

#### 1) selection of headwords

Unlike learners’ dictionaries, MCD only includes nouns, verbs and adjectives as headwords, among which, nouns account for 55%, verbs 21% and adjectives 24% respectively (Coffery 2010). According to Pragglejaz Group (2007), one of the advantages of using dictionaries for metaphor identification is that dictionaries are especially useful for distinguish metaphorical content words from non-metaphorical ones, and for functional words, researchers, to a great extent, have to rely on their intuition. Compared to other collocation dictionaries, eg, *Oxford Collocations Dictionary for Students of English* (henceforth OCD), which has a larger collection of entry words, especially functional words, the headwords included in MCD make it a more convenient means in metaphor identification.

#### 2) segmentation of senses

One of the most distinguished features of MCD is that it highlights metaphorical
meanings of lexical words, and in some cases, lists only high frequency metaphorical meanings. Take “cultivate” as an example: both MED and LDOCE have four different senses, with the first two related to concrete senses and last two abstract senses. In OCD, a traditional collocation dictionary, the compliers provide “cultivate + adv” collocation patterns related to three semantic fields (1. land; 2. crop; 3. try to develop), while MCD only lists one metaphorical sense for collocations: “develop an attitude, ability, or relationship”. Actually, in the entry list of MCD we can find a large quantity of headwords with only metaphorical sense, including “gulf”, “ignite” and “veil” etc. and the heightened awareness on metaphor in MCD offers a direct help for researchers to determine the contextual meaning in MIP.

3) choice of collocates

MCD, with collocates based on semantic groups, can help researchers make decisions when use of learners’ dictionaries leads to confusion:

[8] *I have to repay $250 every month, and that's a big chunk of my salary.*

**chunk** n.
MED: 1. a large, thick piece of something
2. a large amount of part of something
LDOCE: 1. a large thick piece of something that does not have an even shape
2. a large part or amount of something

Both MED and LDOCE have two meanings, but the infinite pronoun “something”, either describes shape of an object in sense 1 or quantity of something in sense 2, is not sufficient to make a judgment about its abstractness, therefore, provides few clues to its metaphorical feature. However, in MCD, the following collocates are listed:

**chunk** n.
MCD:
a large part or amount of something
- adj+N (omitted)
- N + of food  **beef, bread, cheese, chicken, cucumber, lamb, meat, pineapple**
hard solid substance  **antonym, ice, masonry, metal, rock, wood**
time or money  **budget, day, money, salary, and time**

Although in MCD “chunk” is also defined with “something”, its collocates in different semantic groups clear up the confusion caused by the infinity of possibilities in “something”, which may blur the distinction between its concrete and abstract senses, and consequently, lead to researchers’ frustration in MIP.
3.3 Use of specialized dictionaries

As objectivity and precision are crucial in technical and scientific languages, figurative, vague and ambiguous expressions are, to a great extent, undesirable. What’s more, unlike learners’ dictionaries that follow descriptive principles, specialized dictionaries are in essence prescriptive. What’s more, in contrast to historical dictionaries, they are synchronic rather than chronic, and consequently, specialized dictionaries give little consideration to lexicalization process, in which metaphoricity plays an important part (Temmerman 2000). However, as metaphor is an important vehicle for people to conceptualize the world, not only in daily life, but in all kind of activities, including science, business, and legal activities, the language coded in specialized dictionaries cannot be reduced to literal level. Take business for example, as business language, by its very nature, is metaphorical (Koller 2004; White 2003), figurative expressions will certainly make part of business dictionaries. For instance, data from corpora show that the word “bubble” collocates with words related to business in many cases, however, in MED, and LDOCE neither “bubble” as noun or as a verb relates specifically to business, though we can find its connection to “emotion”, “feeling”, “activity” and “time” in the given definitions. However, in Longman Business Dictionary (henceforth LED), we will find:

**bubble**  n.
LED: 1 when a lot of people buy shares in a company that is financially weak, with the result that the price of the shares becomes much higher than their real value
2 **the bubble bursts** if the bubble bursts in a particular area of business, a period of growth and success ends suddenly

As we shall see, though there is only one sense listed in LED, it’s just the metaphorical sense that helps determine contextual meaning more directly, hence more effectively for metaphor identification, especially when we consult LED for a cross reference.

4. Conclusion

Metaphor research is heavily based on metaphor identification, in which MIP is widely applied as a tool. Though in MIP researchers mainly depend on learners’ dictionaries to support their intuition, this will be complemented with use of historical dictionaries, collocation dictionaries and specialized dictionaries for a cross reference. There is no denying that dictionary use in metaphor identification is time-consuming, especially with large amounts of data for analysis, and it’s less applicable when dealing with functional words, special terms and culture-loaded words, yet compared with introspection and corpus method in metaphor identification, MIP is highly recommended to and universally applied by researchers in metaphor identification for the least dependence on intuition (Zhong & Chen 2013), and it is the use of dictionaries that provides an objective basis for the reliability of the MIP.
References

A. Dictionaries:


B. Other Literature


The Presentation of Metaphor in the 7th Edition of Modern Chinese Dictionary

Nan Geng
College of Chinese language and Culture, Beijing Normal University
gengnan52@163.com

Abstract

This article takes the 7th Edition of Modern Chinese Dictionary as an example to analyse the presentation of metaphor from aspects of headword, sense, definition and example. According to the statistical analysis, we found 2812 words with metaphorical meanings and divided them into three types: fixed metaphorical meaning(independent sense, usually guided by the word“比喻 bǐyù‘figurative’”), semi-fixed metaphorical meaning(non independent sense often attached to literal meaning, usually guided by the word“比喻”or“用于比喻 yòngyúbǐyù ‘used metaphorically’ ”) and contextual metaphorical meaning(no sense, use the symbol“◇”before examples which have contextual metaphorical meanings).

These three types reflect a semantic evolution path of metaphorical words in modern Chinese: First, new metaphorical meanings were less conventionalized and accessible. Then, metaphorical meanings gradually became stable. Finally, metaphorical meanings became highly frequent and conventionalized. So fixed metaphorical meaning, semi-fixed metaphorical meaning and contextual metaphorical meaning are in a continuum.

Then we distinguished literal-dominant words, literal-metaphor dominant words and metaphor-dominant words. Finally, we make some suggestions on the definitions and examples of these different types of words.

Keywords: metaphorical meaning, Modern Chinese Dictionary(7th Edition), continuum, polysement, literal-dominant words, literal-metaphor dominant words, metaphor-dominant words

1. Introduction

Modern Chinese Dictionary(MCD) is an authoritative, widely used medium-sized philological dictionary in China. The 7th edition was published by the Commercial Press in September 2016. This revision was based on “General Standard Chinese
Character Table” published by the State Council, added some new words, modified some definitions and deleted a small amount of old words.  

“metaphorical meaning is the meaning of metaphorical usage of the word which was gradually fixed down.” Metaphorical meaning is the sense of a word extended from the polysements by metaphor. As for a dictionary, whether it can deal with the metaphorical meaning of words properly reflects the quality of the definition. The 7th edition of MCD did a good job in this area. According to the statistical analysis, we found 2812 words with metaphorical meanings (identification of metaphorical senses was based on lexicographic custom to add ‘比喻 bì yù ‘metaphorically referring to’ in the definitions to indicate a sense is metaphorical and“◇” symbol is used before examples which have metaphorical meanings.) This article tends to investigate the presentation of metaphor from aspects of headword, sense, definition and example of MCD(7th edition ).

2. The presentation of metaphor in the microstructure of MCD

2.1 Headword

According to the syllable number of words with metaphorical meanings, there are 210 monosyllabic words, 1344 disyllabic words, 376 tri-syllabic words, 741 four syllabic idioms and 141 polysyllabic words. Among them, the disyllabic words play a dominant role, then are four syllabic idioms and tri-syllabic words. According to the part-of-speech, there are 927 nouns, 754 verbs, 46 adjectives and 1078 idioms with no part-of-speech. Besides, there are one numeral, one onomatopoetic word, one measure word and one adverb. Words with metaphorical meanings are mainly nouns and verbs.

2.2 Sense & definition

Su Xin-chun&Zhao Cui-yang(2001) concluded that there were 6 types of metaphorical definitions in MCD(2th edition ). Based on this, we concluded 5 types of metaphorical definitions:  

1Refer to the Release Notes of the 7th edition.  
2This is a quote from Modern Chinese published by the Commercial Press, P215.  
3“经纶 jīnglín、萌芽 méngrýá、欠债 qiànzhài”these three words each has two senses of metaphorical meanings, one is noun, the other is verb. We counted them seperately.  
4 In Su Xin-chun’s classification there used to be a type which was used bracket-annotations to indicate the metaphorical meaning, but in the 7th edition this method is canceled.
The first type (we call it type A) is that metaphorical meaning is independent and unique sense. There are 695 in total, accounting for about 24.72%. For example,

绊脚石 bànjiǎoshí
‘stumbling block’ 比喻阻碍前进的人或事物；骄傲是进步的～。

The second type (we call it type B) is that metaphorical meaning is independent but non-unique sense. There are 671 in total, accounting for about 23.86%. For example,

主流 zhǔliú
‘main stream’ 名干流。②比喻事情发展的主要方面：我们必须分清～和支流，区别本质和现象。

The third type (we call it type C) is that metaphorical meaning is non-independent but often attached to literal meaning. There are 774 in total, accounting for about 27.52%. For example,

先斩后奏 xiānzhǎnhòuzòu
‘acting first and facing the consequences later’ 封建时代臣子把人杀了再报告皇帝‘the minister killed the man first then reported to the emperor’ (literal meaning)。现在多比喻自行把问题处理了，然后才报告上级或尊长‘to deal with the problem ,and then report to superiors or elders’ (metaphorical meaning)。

The fourth type (we call it type D) is that metaphorical meaning attached to literal meaning, but uses“用于比喻 yòngyù‘used metaphorically’” to indicate a sense is metaphorical. There are 97 in total, accounting for about 3.45%. For example,

战绩 zhànjià
‘combat gains’ 名战争中获得的成绩‘achievements in the war’(literal meaning), 也用于比喻‘used metaphorically’；以全胜～夺冠。

The fifth type (we call it type E) is that symbol“◇”is used before examples to indicate that the example is metaphorical. There are 560 in total, accounting for about 19.91%. For example,
“starting line”赛跑时起点的标志线◇在同一～上展开公平竞争“fair play on the same race line”。

⑤In addition to five types, some words have two types. For example,

落马 luòmǎ

‘fall from the horse’动①骑马驰骋时，从马上掉来(literal meaning)，也比喻打仗或竞赛失利(metaphorical meaning)：中弹～|半决赛中，上届冠军意外～。②比喻官员的劣迹败露而遭到惩处（多指被免职）(metaphorical meaning)：经群众检举揭发，这个贪官终于～。

This word has two types : C and B. There are only 15 in total, accounting for about 0.53%.

The pie chart below shows the types of metaphorical definitions in MCD(7th edition).

Figure 1 The types of metaphorical definitions

It can be seen that except for type D, the proportion of the other four types is approximate, the most is type C, that is explain the literal meaning first and then explain the metaphorical meaning.

These five types reflect that there are three stages of metaphorical development: Fixed metaphorical meaning，Semi-fixed metaphorical meaning and Contextual metaphorical meaning. Fixed metaphorical meaning has become highly conventionalized and accessible in the speech community and can be listed as one independent sense, usually guided by the word“比喻”，such as type A&B; Semi-fixed metaphorical meaning is nonindependent sense attached to literal meaning, usually guided by the word“比喻”or“用于比喻 yòngyùbiyù‘used metaphorically’”,such astype C&D; Contextual metaphorical meaning is often used in a certain context, and this
temporary usage is still unstable, although it has a certain frequency of use. Lexicographers often use the symbol “◇” before examples to indicate that it is a metaphorical meaning, such as type E. The following pie chart shows the proportion of three stages of metaphorical development.

![Pie Chart](image)

**Figure 2** The types of metaphorical meaning

These three stages of metaphorical meaning reflect a semantic evolution path of metaphorical words in modern Chinese: First, new metaphorical meanings emerged from the speech community, they were less conventionalized and accessible. Then, metaphorical meanings gradually became stable. Finally, metaphorical meanings became highly frequent and conventionalized. So fixed metaphorical meaning, semi-fixed metaphorical meaning and contextual metaphorical meaning are in a continuum.

### 2.3 Example

Example is an indispensable part of a dictionary. The function of the example is to indicate the meaning and usage of words and to provide further information. For example, the symbol “◇” is used before examples to indicate that this is a metaphorical usage. Compared with the 7th and the 6th edition, we found there were some new contextual metaphorical meaning words with the symbol “◇”. For example,

**绊马索bàn mǎ suǒ**

‘heel rope’ 设在暗处用来绊倒对方人马的绳索。 (6th edition) 设在暗处用来绊倒对方人马的绳索◇一些过时的规定已成为改革的～，应该尽快修订。 (7th edition)

**高压线gāoyā xiàn**
The definition of “绊马索bàn mǎ suǒ ‘heel rope’”、高压线gāoyāoxiàn ‘high voltage cable’、华诞huá dàn ‘birthday’、快车道kuài chē dào ‘fast lane’” in the 7th edition are literal meanings, but the examples are metaphorical(with the symbol“◇”). While the definition of “脉动mài dòng ‘pulse’” is metaphorical meaning, and the examples are metaphorical too. So the question is why the definition is literal but the example is metaphorical? Is it reasonable? According to the Center for Chinese Linguistics(CCL)PKU Corpus of Modern Chinese(http://ccl.pku.cn:8080), we tend to investigate how literal and metaphorical senses are used in these words. After investigating we find that there are three cases:

The first case is that literal meaning is dominant, while metaphorical meaning is subordinate and usually marked with double quotation marks. We call it “literal-dominant word5”. For example,

The word “高压线gāoyāoxiàn ‘high voltage cable’” is used 82% of the time in its literal meaning, but 18% of the time in its extended metaphorical ‘taboo’ meaning.

5Based on the view of Yuchun Chang, Chien-Jer Lin and Kathleen Ahrens, Language and Linguistics 16(4)587-614. They distinguished two types of lexemes: literal-dominant lexemes and metaphor-dominant lexemes.
When people use the metaphorical ‘taboo’ meaning, we usually add double quotation marks on the word to indicate that it is a metaphorical usage. For example,

①警示名单后果：会计碰了“高压线”被记入“警示名单”，国有单位不得聘用。
②专用发票是整个税制改革工程的“高压线”，谁敢违法触动，就必将要受到法律的严惩。
③他说：“现在一讲减轻农民负担，集资办学成了‘高压线’，谁也不敢碰了。”

The second case is that both the literal and metaphorical meaning are dominant. We call it “ literal-metaphor dominant word ”. For example, the word “绊马索 bànmǎsuǒ ‘heel rope’” is used 57% of the time in its literal meaning, and 43% of the time in its extended metaphorical ‘obstacle’ meaning. Another word “华诞 huádàn ‘birthday’” is used 40% of the time in its literal meaning (apply to people), and 60% of the time in its extended metaphorical meaning (apply to institutions, companies, agencies and so on). However the definition of “华诞” only explains that it is a birthday of a person, so the definition is unsuitable.

The third case is that metaphorical meaning is dominant, but literal meaning is subordinate. We call it “metaphor-dominant word”. For example, the word “快车道 kuàichēdào ‘fast lane’” is used only 1% of the time in its literal meaning, but 99% of the time in its extended metaphorical ‘develop rapidly’ meaning. The word “脉动 làidòng ‘pulse’” is used only 6% of the time in its literal meaning (apply to people), but 94% of the time in its extended metaphorical ‘periodic movement of things’ meaning.

For the literal-dominant words, the definition should explain the literal meaning and use symbol“◇” before examples to indicate the metaphorical meaning. The 7th edition’s revision of the word “高压线 gāoyáxiàn ‘high voltage cable’” is reasonable. For the literal-metaphor dominant words, the definition should explain both literal and metaphorical meanings. The 7th edition’s definitions of words “绊马索 bànmǎsuǒ ‘heel rope’” and “华诞 huádàn ‘birthday’” are only literal meanings, but the examples are metaphorical meanings, so we suggest to explain the literal meaning first and then the metaphorical meaning, like type C. The examples should better be shown in the same way: first example with literal meaning and then example with metaphorical meaning. For the metaphor-dominant words, the definition should explain the metaphorical meaning. The definition of the word “快车道 kuàichēdào ‘fast lane’” has only literal meaning but no metaphorical meaning which is unreasonable. The definition of the word “脉动 làidòng ‘pulse’” is metaphorical meaning and the example is also metaphorical. The 7th edition’s revision of “脉动” is reasonable.
3. Conclusion

Overall, the presentation of metaphor in the 7th Edition of MCD reflects the new achievements of metaphor in linguistics. It comprehensively displays the features and usages of metaphorical words in modern Chinese. For some new words and new senses, especially with metaphorical meanings, the lexicographer should pay more attention to distinguish whether the word is literal-dominant, literal-metaphor dominant or metaphor-dominant. It is hoped that the definitions and examples of such words can be more precise and objective.

References


中国社会科学院语言研究所词典编辑室. 2016. 现代汉语词典（第7版）. 北京：商务印书馆.
Method and Motivation of the Foregrounding Language Style of

Paper English Learners’ Dictionaries----A Case Study of COBUILD 8

Weifu Zhang
Southwest University of Science and Technology
913415691@qq.com

Congmei Chen
Southwest University of Science and Technology
632218086@qq.com

Abstract

Recent development in technology has witnessed a decline in the utilization of paper English learners’ dictionaries. This decline, however, in turn, requests better paper English learners’ dictionaries be released, among which Collins COBUILD Advanced Learners’ English-Chinese Dictionary (the 8th edition, shortened as COBUILD 8), with respect to its unique language style, can meet the demand for a relatively frequent refer
rt), we find that the method of foregrounding is the phenomenon of language deviation and parallelism, with the former pointing to the idiosyncrasy in the language style, and the latter being the repetition of patterns to present the language style. Moreover, the motivation of foregrounding, as this paper penetrates, can be better elucidated by the theory of Figure-Ground by Talmy, where the terms figure and ground have both linguistic and psychological explanation. That is, psychological figure and ground can be reached linguistically.

Keywords: foregrounding, language style, paper English learners’ dictionary, COBUILD 8, deviation, parallelism, Figure-Ground

1. Introduction

A survey by two Chinese scholars (Lu & Wang 2006), suggested that the present day was an era consisting of dictionaries of three mediums, namely, paper dictionaries, e-dictionaries and on-line dictionaries. This trichotomy notwithstanding, there appears a remarkably decline in the consulting of paper dictionaries, especially of English learners’ paper dictionaries.

The reason may be more or less explained, at least based upon some domestic researches, by the fact that paper dictionaries are too cumbersome and are doomed time-consuming for their referring process, a defect that can be said to have resulted from the essence of the language presentation in paper dictionaries. It is not unknown to all that, dictionaries, regardless of their mediums, are complied to reflect language in use through language in practice. It is language itself that gives priority to the realization of dictionaries. Thus, paper dictionary compilers have devoted themselves
compiling dictionaries considered to have promoted in their language nature, among which, Collins COBUILD Advanced Learners’ English-Chinese Dictionary, the 8th edition (shortened as COBUILD 8) is unique in its language. This uniqueness has triggered the concentration upon the language in paper dictionaries, therefore, language in paper dictionaries is tackled as style one of a kind, combined with the example of COBUILD 8, to discover its foregrounding method and motivation that can, to a certain extent, anticipate an increase of paper dictionary reference in the near future.

2. On defining COBUILD 8-style

For in stylistic study, foregrounding is typically connected with the term style. We have but to, at the very beginning, prove that the language in paper dictionaries is style one of a kind, combined with the example of COBUILD 8.

In stylistics, the notion of style is varied. Firstly, as defined by J. Thornborrow & S. Wareing (2000), that ‘style is usually understood with this area of study as the selection of certain linguistic forms or features over other possible ones’. COBUILD 8, known for its full-sentence definitions, it is a certain linguistic selection over whose definitions are synonyms or phrases. This is the statement that considers style as choice. Other versions are shown below, from more condensed to more detailed:

‘Style in language can be defined as distinctive linguistic expression’. (P. Verdonk 2012)
‘Style refers to the way in which language is used in a given context, by a given person, for a given purpose’. (Leech & Shorts 2001)
‘Style is the characteristic use of language in a text. When referring to the speaker, style is more or less the controlled choice of linguistic means, whereas in referring to texts, style is the specific form of language. For the reader or the listener, style is the variation (or conformation) of possible expectations’. (H. Bussmann 2000)

Apparently, style has much to do with the use of language and these definitions have beaten a path to our understanding for the language used in paper dictionaries (and other forms of dictionaries), i. e. the dictionary style. Following Verdonk’s definition, dictionary style can refer to the distinctive linguistic expression in dictionaries. Thus in COBUILD 8, its full-sentence definitions, hyphenation points to show where a word may be split across lines, or colour headwords (like after|shock), these distinctions show the evidence of the dictionary style. For a revised version in accordance with Leech & Shorts’ interpretation, we have this: dictionary style refers to the way in which language is used in a given dictionary, by a given person (the compiler(s)), for a given purpose (For example COBUILD 8 is compiled for English learning). To be more detailed as H. Bussmann’s explanation, therefore, dictionary style is the characteristic use of language in a text (i.e. a given dictionary). When referring to the speaker (here the dictionary-maker to be specific), dictionary style is more or less the controlled choice of linguistic means, whereas in referring to texts, dictionary style is
the specific form of language. For the reader or the listener (here refers to dictionary user), dictionary style is the variation (or conformation) of possible expectations. In COBUILD 8, it is comprehensible that its linguistic means are controlled by its makers. In other words, the style in COBUILD 8 is classifiable to this degree. Moreover, recent dictionary compiling has seen an increasing focus upon the dictionary user, which means, in turn, that users are expecting something in a given dictionary designed for them through the variation (or conformation) of dictionary style. At least for this reason, that’s why we have to study the dictionary style for the users’ sake.

Furthermore, to meet what the term style means does not suffice in that, as claimed by Ding (2014), style should carry at the same time three mechanisms: dependence upon certain register (different from Halliday’s register as field, mode, tuner), without which style cannot come into being; guidance upon certain expression, which means expression in such a register is selective in a way that caters to this register; gradableness in form, which means style can be gradable, through degree of formality, appearance of form or other methods. Dictionary style in COBUILD 8 is dependent upon the very existence of COBUILD 8 regarded as certain register accordingly; then, as an English learners’ dictionary, style of COBUILD 8 is selective in some of its expression that we have discussed; last, the style in COBUILD is gradable by using some techniques such as style in bold, blue color, italicization that are considered more highlighted than that not in.

In a word, the language in a dictionary has met the definition of style in stylistics and takes on also the same three mechanisms as the term style embodies, for which we term it as dictionary style. And as we will analyze COBUILD 8, dictionary style in COBUILD 8 is shortened as COBUILD 8-style.

3. Classification of COBUILD 8-style

Since COBUILD 8-style is already identifiable and has three mechanisms aforementioned, especially its gradableness which has a definite link with the form of style, i.e. gradableness can be reached via the form of COBUILD 8-style, next should be a trichotomy classification of COBUILD 8-style in terms of illustrations, symbols and definitions.

The first should be the illustration style that refers to the style in illustrations in the two columns Picture Dictionary and Word Web in COBUILD 8 where coexists both words and pictures as shown following here:
It can be seen directly that style in Picture Dictionary (the left side) has a clear word-picture correspondence with a bar linking the word-picture set, where the words are of bold font; whereas the style in Word Web (the right side) has a noticeably conciser version where pictures are above the bold words. Both of these two illustration styles, in a word, are simple in form but occupy much more space than what symbol style takes up and is the least out of the other two styles.

Then, the symbol style in COBUILD 8, takes two forms as symbols in words and symbols in signs, with the former referring to abbreviations like ADJ, N-COUNT etc., tags like [+of], 【JOURNALISM】 etc usually included in a certain punctuation; the latter being signs such as squared number, where the base is always blue; diamonds with or without blue base; blue based triangle that has a definite direction toward the right; and shadowed square. These symbols, unlike illustrations, are abundant in numbers. For every page in COBUILD 8, there must be symbols. And, word symbols, together with tags, convey semantic meanings themselves, like ADJ for adjective, whereas sign styles are almost free from semantic interpretations. Rather, they are usually doomed as an initiative to conduct a certain part in the definition text. For example, the shadowed square always lies before authentic examples from the Collins corpus (so termed in the preface of COBUILD 8). Lastly, all the symbol style is in the lines of the definition style.

The last definition style means the style in the definition part of COBUILD 8. Note that the definition here does not narrowly refer to semantic meanings of a given entry, but the part that contains all the elements, except symbol styles, in the text of an entry. By stating this, the conception of definition reaches a broader sense that seems more complicated in that there are varied stylistic features in a definition text, which, however, has something to do with the entry:

1) when the definition itself is actually a cross reference system, see entries like kabob, presented in COBUILD 8 as:

   ka|bob /ˈka.bəb/ (kabobs) → see kebab

   Despite the components in slashes and parentheses (symbol style to be specific), definition style consists here of the entity (kabob) that is bold blue and the rest from the arrow to the bold word kebab. Not all entries have such a pattern, though, definitions for these entries are in fact a way of cross reference, for example, the word humankind:

   human|kind /ˈhju.mənˌkiːnd/ N-UNCOUNT Humankind is the same of as mankind.

2) when the definition is narrowed down to merely semantic interpretation, see entries like mastiff, holly etc. Take holly:

   holly /ˈholi/ (hollies) N-VAR Holly is an evergreen tree or shrub which has hard, shiny leaves with sharp points, and red berries in winter.
3) when the definition is followed by examples from the corpus, entries of this trait are found much more easily. Shown below is the word *fitter*:

*fitter* /ˈfɪtər/ (fitters) N-COUNT A **fitter** is a person whose job is to put together, adjust, or install machinery or equipment.  

*Geo* is a fitter at the shipyard.

We notice that the part led by shadowed square is both bold and italic, quite different in font from that of the narrow semantic interpretation.

To sum up, three kinds of style are observed in COBUILD 8, as one can see, the number of illustration style is far less than the other two styles, but it takes much more space; while the symbol style usually lies in the lines of the definition style that is obviously the most dominant in quantities.

### 4. Foregrounding of COBUILD 8-style: method

The notion of Foregrounding can be traced back to the roots of literary theory—the school of the Russian Formalists (mainly Shklovskij) at the beginning of the 20th century (Hakemulder, 2000). Also, it is further developed by Mukarovsky, Jacobson and Leech of Prague School. Both Shklovskij and Mukarovsky speak highly of foregrounding as language *deviation*; while Jacobson and Leech concentrate upon of foregrounding as language *parallelism*. These two dimensions, according to some stylistics (Liu and Zhu 2000 for example), are respectively viewed as vertical (or paradigmatic) and horizontal (or syntagmatic) research of foregrounded language, which will be explained later.

#### 4.1 Deviation as Foregrounding for COBUILD 8-Style

Deviation refers to the language whose usage is far disparate from daily language. Originally, as analyzed in Shklovskij’s works, Foregrounding is realized through the poetic use of language that can have the effect called *defamiliarization* whose function is to make the language in practice strange to the readers, so as to prolong the process of reading and hence to elevate the aesthetic value of a certain work. Defamiliarization is further promoted by Mukarovsky, who, in his 1964 work *Standard Language and Poetic Language*, put forward initially the concept of *foregrounding*\(^1\) contradictory to *automatization* whereby language is systematically and routinely applied, that is, foregrounding means the violation of automatized language application. It should be noted that defamiliarization or foregrounding has a profound historical development that we cannot elaborate here, but at least one point must be clear that despite defamiliarization and foregrounding, they are both deviation of daily language, a method of Foregrounding.

When it comes to COBUILD 8-style, deviation refers to the idiosyncrasy in the language style whose hint can be traced from the very existence of those three styles

---

\(^1\) Here the small lettered ‘foregrounding’ is an aspect of deviated language, while the capitalized ‘Foregrounding’ contains all the foregrounded methods.
1) Concerning the definition of a (paper) dictionary (see Béjoint 2001; Huang 2001; Landau 2005, etc.), much is confirmed that a (paper) dictionary is about words. This fact, however, has reversely proved to some extent that the use of illustrations in a (paper) dictionary is somewhat deviated, so is true considering the use of the symbol style, especially the sign style, based upon the comparison to the conventional use of words. Thus, at least to this point, the illustration style and the symbol style are idiosyncratic in COBUILD 8 that should be viewed language deviation.

2) Speaking of the definition style in COBUILD 8, though, things are quite different from the other two styles. Firstly, the narrow definition is deviated from two perspectives: one is its full-sentence definition that can be said to be very unique, thus being idiosyncrasy, and has been studied by many scholars (Zhang 2000; Xu et al 2012), since other brands (Oxford series or Longman series) of English learners’ dictionaries are disparate in such full-sentence definition. The other is the way every entry is defined, which is in sharp contrast with the way entries are defined in daily language, thus deviated. To explain their distinctions, Huang (2001:83) has designed the following dialogue (translated from Chinese to English):

-He died of leukemia last year.
-What is leukemia?
-A disease of the hemopoietic system, also called blood cancer (in Chinese).

It can be roughly concluded that, corresponding to Huang’s explanation, daily definitions tend to be circumstantial and can be reached simply by facial expressions, gestures, tones so as to make understood. While (paper English Learners’) dictionary definitions must be written, so are the examples for every definition. This requires a more scientific application of language and hence deviated from daily routines. Compare the word leukemia defined in COBUILD 8:

leu|kea|mia /lu'ki:miə/ in AM, use leukemia N-UNCOUNT Leukeamia is a disease of the blood in which the body produces too many white blood cells.

Much is presented about the word leukemia, and its definition is a declarative sentence triggered by nothing on the paper; whereas the definition in the dialogue is triggered by a given question. Moreover, the same problem is also discussed by Zhang and Yong (2007:241). These two points being synthesized, we have held water that the definition style in COBUILD 8 is deviated (because of its idiosyncrasy).

4.2 Parallelism as Foregrounding for COBUILD 8-Style

Parallelism as a foregrounding method was initiated by Jacobson and further enhanced by Leech and Short. Language deviation, as interpreted by Leech and Short, is something like qualitative foregrounding (and quantitative foregrounding for parallelism). For a quantitative foregrounding, they narrate, that at some point it
(deviation)\(^2\) becomes significant not that a writer has chosen \(x\) rather than \(y\) or \(z\), but that he has chosen \(x\) at all. This is the core of parallelism or quantitative foregrounding, that is, to have some elements of language repeated within the body of the same work. It is worth reminding that the quantity of the repeated elements should surpass to a larger degree the norm requires. In this sense, the reason why deviation and parallelism are separately termed vertical (paradigmatic) or horizontal (syntagmatic) foregrounding is now comprehended. In COBUILD 8-style, parallelism is the repetition of patterns to present the language style and is embodied in every three types of style.

Of the illustration style, the ensuing pinpoints that keep repeating within enable its stylistic parallelism in COBUILD 8:

1) Monochrome. From what we have displaced about the illustration style, either from Picture Dictionary or from Word Web, they share fundamentally the same color set, i.e. monochrome. And incidentally, such a color set is responsible for both that seem to enjoy a different base from what the other two styles do.
2) Location. The illustration style appears almost at the front of a paper that allows the inclusion of illustrations. Chances are that a few illustrations happen to be at the bottom of a certain page for the simple and ultimate purpose of a paper dictionary to safe limit.
3) Word-picture correspondence. This has been talked before that the column Picture Dictionary, where the word-picture correspondence is connected with a bar, is in several contrast with Word Web whose correspondence turns out to be much easier. Altogether, however, words for them are all in bold font.

As for the symbol style of COBUILD 8, be it word-based or sign-based, it does share much in common as a whole:

1) High frequency. This means that symbol style can be found almost everywhere in a given content in COBUILD 8. Whenever there are entries to be defined, except for some really rare cases, abbreviations of their part of speech show up as word-based symbol style. Meanwhile, where there is an authentic example from the Collins corpus, there must be a shadowed square. Other word tokens as symbol style are comprehensible to such extent, too.
2) Specialty. All the symbols in COBUILD 8 are specially created for a special use or reference. Word-formed symbols, like BRIT, AM are referring specially to Britain English, American English. Sign-based symbols, on the same account, have also a specially treatment, such as the sign diamond ( ) is always used to indicate the frequency of the entry, while the squared numbers ( ) within blue base are always applied for the meaning splits of the entity.

In terms of the definition style, its parallelism is realized via what is analyzed below:

---

\(^2\) Leech and Short regard the phenomenon deviation as a whole with two sides: qualitative and quantitative, the formal being the very deviation we are referring to and the latter being the very parallelism we are going to discuss.
1) Syntactical concord. This explains the fact that all the narrow definitions of the entries in COBUILD 8 are full-sentence definitions. More specifically, these full sentences are either single sentences that tend to be longer or complex ones that usually start with the pattern if you... On the other hand, syntactically, it has been claimed that the authentic examples from the Collins corpus are all in bold and italic font.

2) Part-of-speech dependence. This is fairly the case with narrow definitions. When the entry is a noun, the definition usually begins with the entity as the subject of the full-sentence definition. For instance, the entry handbag is patterned as ‘A handbag is a...’; the entry incinerator ‘An incinerator is a...’ etc. Verbs have a different pattern, for the most of time, are initiated by if you.... The verb miniature, as an example, is defined at the beginning that ‘If you miniature something..., you...’. Circumstances of other parts of speech are confirmed in COBUILD 8 that their definitions have much to do with their parts of speech.

3) Bold entities inside. For every full-sentence definition, the entries inside are always bold, as can be seen from all the examples above.

5. Foregrounding of COBUILD 8-style: motivation

We have hitherto never coped with the motivation of the Foregrounding of COBUILD 8-Style. The questions facing us are, as one can imagine, why should COBUILD 8 own such foregrounded styles? How can we say the methods of deviation and parallelism are of foregrounding performance? What is the effect of such foregrounding phenomena? The Figure-Ground theory advanced by Talmy (2000) may give answers to all these questions.

5.1 Figure-Ground Theory by Talmy

This theory was originally introduced into psychology by the Danish psychologist Rubin who enacted the famous face/vase reversal, and, all in all, the figure is perceived as being more prominent than the ground (F. Ungerer & H.-J. Schmid). That is to say, figures are more favored by our perceptions than does the ground. This perspective prominence interested some linguists, like Talmy, who was said to be the first to put this theory into linguistic practice in a cognitive study. According to Talmy (2000:312), Figure and Ground are conceptualized as follow:

‘The Figure is a moving or conceptually movable entity whose path, site, or the orientation is conceived as a variable, the particular value of which is the relevant issue.

The Ground is a reference entity, one that has a stationary setting relative to a reference frame, with respect to which the Figure’s path, site, or orientation is characterized.’

When Talmy put forward the above conceptions, he was using them to study syntax, from single clauses to complex sentences, to discover the Figure and Ground in
a given sentence where three elements existed: the Figure object, the Ground object and a reference frame. The Figure object is prominence in a sentence, and will be the cognitive subject; the Ground is what Figure object is characterized, and will be what the cognitive subject will base upon; and the reference frame is the background where these two objects locate. And we should first keep in mind that the object being the Figure or the Ground is not unalterable, which means in some context, the Figure object can change into the Ground object and vice versa. The capitalized set Figure/Ground, he explained, referred to the linguistic Figure-Ground, while the combination figure/ground referred to the psychological figure/ground. When the reference frame was not taken into consideration, linguistic Figure/Ground was in consistence with the psychological figure/ground. When the reference frame was considered, however, he offered three interpretations that are presented here in a table:

<table>
<thead>
<tr>
<th>Interpretation 1</th>
<th>F→f, G→f, RF→g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpretation 2</td>
<td>F&amp;G→f, RF→g</td>
</tr>
<tr>
<td>Interpretation 3</td>
<td>F→f, G&amp;RF→g</td>
</tr>
</tbody>
</table>

(Note: F/G: linguistic Figure/Ground; f/g: psychological figure/ground; RF: reference frame)

Interpretation 1 means both linguistic Figure and Ground can relate to the psychological figure, when the reference framework is regarded as the psychological ground; Interpretation 2 figures out that the combination of linguistic Figure/Ground as a whole can correspond to the psychological figure, reference framework again seen as psychological ground; Interpretation 3 indicates that the combination of linguistic Ground and reference framework become the psychological ground, with the linguistic Figure alone being the psychological figure. His interpretations provide some grounds to explore the motivation of the foregrounding methods in COBUILD 8-style to meet the perceptual prominence so as to have the psychological prominence aroused, that is, psychological figure and ground can be reached linguistically.

5.2 Figure-Ground in COBUILD 8-Style

Following up his interpretations, we now turn to the foregrounding methods of COBUILD 8-style with such a hypothesis that they have much to do with the Figure-Ground theory. To correlate the Figure/Ground of COBUILD 8-style to the psychological figure/ground, situations are divided into two kinds without and with the reference frame.

1) Without reference frame. This calls for a respective study upon the three sorts of COBUILD 8-style, catering to their parallelism foregrounding in (3.2). Thus, in the illustration style alone, its color setting (monochrome) is the Ground in that it is stationary, or, in other words, it is decided and never changed. As for the location of illustration style, it functions as the medium connecting its color setting and the word-picture correspondence considered to be the Figure of the illustration style for
reasons that the illustrations help us get more information about the entry and it is in fact the word-picture correspondence we are looking up, rather than merely the picture or the word in the illustration. That means, in an illustration, it does not suffice provided that there is just the picture; and it would not be an illustration simply with words. Therefore, the correspondence between the word and the picture is the Figure that actually plays the psychological figure, and becomes the cognitive subject.

Thinking of the symbol style, its specialty overshadows its high frequency. And this specialty is conceptually movable in a way that it ‘moves’ with different symbols, for which the specialty is the Figure of the symbol style, meaning that we are paying attention to what significance all these symbols convey. Relatively, the frequency, together with the symbols themselves are the Ground, because they are too many to count and the reasons to discuss about why the diamonds are used to indicate the frequency of an entity or why an countable noun is abbreviated as word symbol N-COUNT, are not of much importance of our research here.

For the definition style, the syntagtic concord parallelism and the part-of-speech dependence are the Ground and the bold entries inside the Figure. Firstly, syntagtic concord and the part-of-speech dependence are not variable that nearly appear to be a framework to be filled in something in accordance with the entry. Then, upon the same syntactic base, the bold entries thus come into being, finally becoming the Figure for its boldness.

So we have a conclusion here as shown below in a table:

<table>
<thead>
<tr>
<th>COBUILD 8-Style</th>
<th>F=f</th>
<th>G=g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illustration Style</td>
<td>Word-Picture Correspondence</td>
<td>Color Setting, Location</td>
</tr>
<tr>
<td>Symbol Style</td>
<td>Significance of Symbols</td>
<td>Frequency, Shape, Form</td>
</tr>
<tr>
<td>Definition Style</td>
<td>Bold Entries in Definitions</td>
<td>Syntagtic Concord, Part-of-Speech Dependence</td>
</tr>
</tbody>
</table>

2) With reference frame. Since the reference frame is the background is where where these two objects (Figure-Ground) locate, obviously, it points to all the pages (the body part) of COBUILD 8. And this time we combine the deviation foregrounding of COBUILD 8-styles talked in (3.1) with the Figure-Ground theory. It’s discussed in (3.1) that the illustration and symbol styles are deviated to an extent compared to the definition style that comprises just words and has a much larger occupation. Thus, when these three types of styles are simultaneously existent in a given page, at the first glimpse, the definition style serves as the Ground. The illustration and symbol styles are viewed as the Figure. And with the interpretations by Talmy being taken into account, Interpretation 1&2 can be good explanations. Other possibilities nonetheless are feasible when the information in the illustration style is not what is factually demanded, but the information in the definition style, the definition style now turns into the Figure, whereas the illustration style and the symbol style together become the Ground.

When we see the definition style as deviated, we have made a comparison with other definition styles in different paper dictionaries, say, Oxford series or Longman
series. So in this sense, the definition style of COBUILD 8 we are discussing is the Figure, while the other styles of definition not in COBUILD 8 are the Ground. Interpretation 3 mentioned should be the explanation for the Figure/Ground towards the figure/ground of the definition style.

Similarly, this part is summarized in the following table associated with Talmy’s interpretations:

<table>
<thead>
<tr>
<th>Talmy’s Interpretations</th>
<th>f</th>
<th>g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpretation 1</td>
<td>F=Illustration and Symbol Styles; G=Definition Style</td>
<td>RF=Pages Where All the Three Style are Located</td>
</tr>
<tr>
<td>Interpretation 2</td>
<td>F=Combination of Illustration and Symbol style; G=Definition Style</td>
<td>RF=Pages Where All the Three Style are Located</td>
</tr>
<tr>
<td>Interpretation 3</td>
<td>F=Definition Style in COBUILD 8; G=Other Definition Style Not in COBUILD 8</td>
<td>RF=Daily Language</td>
</tr>
</tbody>
</table>

6. Conclusion

In this research, in the first place, we pose the problem that results in the decline of paper English learners’ dictionary use. Then, with the coming out of COBUILD 8, things are expected to get better by the analysis of the language style in COBUILD 8. After some basic analyses, we find that the method of foregrounding is the phenomenon of language deviation and parallelism, with the former pointing to the idiosyncrasy in the language style, and the latter being the repetition of patterns to present the language style. Moreover, the motivation of foregrounding, as this paper penetrates, can be better elucidated by the theory of Figure-Ground, where the terms figure and ground have both linguistic and psychological explanation. That is, psychological figure and ground can be reached linguistically.

Last but not least, such a stylistic approach to study a decided dictionary is not very smooth. Some errors must exist. It’s expected to have these errors figured out and revised in near future.

Reference


Heterosemy of Self-reference Lexemes in Modern Chinese from the Perspective of the Two-level Word Class Categorization Theory

Renqiang Wang  
Sichuan International Studies University  
wangrenqiang@sisu.edu.cn

Changning Huang  
Tsinghua University  
cnhuang0908@126.com

Abstract

Heterosemy of lexemes especially the status of self-reference usages has been the most controversial topic in Modern Chinese grammar and lexicography. The mainstream view is to follow the Principle of Parsimony so as to minimize the number of heterosemous lexemes, and considers multifunctionality of lexemes as peculiar to Modern Chinese, which results in category indeterminacy or underspecification. From the perspective of the Two-level Word Class Categorization Theory, it is found that such a view has left a negative impact on the word class labeling in Modern Chinese dictionaries and on the POS tagging in Modern Chinese corpora: The number of heterosemous lexemes has been minimized, creating a false impression that there is less heterosemy in Modern Chinese than in Modern English; The conventionalized self-reference senses for those lexemes with symmetrical semantic relationship have contradictory representation of their word class information both within one dictionary and between dictionaries. Thus, lifting the ban on self-reference heterosemy is an important way out of the Chinese word class dilemma.

Keywords: Modern Chinese, heterosemy, self-reference lexemes, Principle of Parsimony, the Two-level Word Class Categorization Theory

1. Introduction

Heterosemy refers to the multifunctionality of a polysemous lexeme with two or more word classes in the lexicon of the communal language system, which is represented as a heterosemous lexeme in language dictionaries (Wang, 2014:345). Taking Oxford Advanced Learner’s Dictionary of Current English (9th edition) for example, abandon is a lexeme that belongs to both verbs and nouns, while like is a lexeme belonging to prepositions, verbs, conjunctions, nouns, adjectives and adverbs. In typical analytic languages as Modern Chinese, Modern English and Modern Vietnamese, the multifunctionality of lexemes has been the most knotty and heated issue in the academia (Enfield, 2015; Jespersen, 1924; Lieber, 2005; Schmid, 2011; van Lier & Rijkhoff, 2013). And there are also different views with regard to the multifunctionality
of lexemes in Chinese. Some hold that Chinese word classes of lexemes are underspecified (Ma, 1898; Li, 1924; Vogel, 2005:67; Ježek & Ramat, 2009: 397; Lin and Han, 2011; Bisang, 2011), and some argue that Chinese content words do not have word classes (Gao, 1953), and some propose that parts of speech of Chinese words do not change as long as the word meaning remains unchanged (Lü and Zhu, 1951), and some insist that “nouny verbs” and “nouny adjectives” should be added as different word classes in Chinese (Zhu, 1982), and still others argue that the word class system of Chinese content words belongs to “inclusive model” (Shen, 2009). To be brief, while the mainstream view recognizes the multifunctionality of lexemes as an objective existence, the Principle of Parsimony is followed so as to minimize the number of heterosemous lexemes (i.e. “No more word classes needed if the meaning remains unchanged”) (Zhu, 1982: 39; Guo, 2002: 101; Lu, 2013: 57; Shen, 2009: 4). In order to avoid excessive heterosemy in Chinese, some scholars argue that Chinese lexical categories are multifunctional. For instance, it is considered that Chinese verbs can function as subject and object in Chinese (Zhu, 1982; Shen, 2009, 2012, 2015; Lu, 2013). However, this brings about the dilemma that fixing word classes leads to unfixed syntactic functions. Huang et al. (2009: 60) points out that “heterosemy can only be the minimized” has become a commandment to deal with the heterosemy of Chinese content words in Modern Chinese grammar, but no one wants to make it clear the exact number limit.

In the study of heterosemy, whether self-reference lexemes like chūbǎn should be recognized as heterosemous or not turns out to be a knotty problem that has not been properly settled in Modern Chinese grammar (Shen, 2015: 141). The self-reference senses of lexemes in Chinese dictionaries refer to those headwords with conventionalized self-reference usages, which is closely related to nominalization in the study of Chinese grammar. The term “nominalization” in Chinese was first proposed in A Provisional System of Grammar for Teaching Chinese edited by Zhang Zhigong in 1956. Chinese scholars tend to regard nominalization as reference, including self-reference and transferred reference (Zhu, 1983; Guo, 2002). Transferred reference is equivalent to grammatical metonymy (Shen, 1999; Wang, 2006), whereas self-reference corresponds to grammatical metaphor (Wang, 2006, 2009; Yang, 2008). The deverbal or deadjectival transferred reference usages involve not only the changes of word classes, but also variations in meaning. Therefore, they are usually treated as independent senses in Chinese dictionaries. However, the conventionalized deverbal or deadjectival self-reference usages are normally not recognized as independent senses in Modern Chinese dictionaries (Zhang, 1979; Guo, 1999; Su, 2002; Wang, 2004; Mo, 2004). Nevertheless, as Hunston and Francis (2000: 192) put it, it is fundamentally different to recognize the multifunctionality of a lexeme as the multifunctionality of a certain word class or as one lexeme belonging to more than one word class, because it is related to “the very nature of word class itself”. In view of this, the article is intended to examine the heterosemy of self-reference lexemes in Modern Chinese from the perspective of the Two-level Word Class Categorization Theory in order to settle the dilemma of word class research in Modern Chinese.
2. Heterosemy in English and Chinese from the Perspective of Two-level Word Class Categorization Theory

2.1 A Brief Introduction to the Two-level Word Class Categorization Theory

In order to handle the multifunctionality of lexemes in analytic languages like Modern English and Modern Chinese, Jespersen (1924), Guo (2002) and Dash (2005) propose the idea of discussing word classes on different levels, but they all have their limitations (Wang, 2014: 55).

Based on the perspectives of language as a complex adaptive system and the nature of major parts of speech as propositional speech act functions proposed by Croft (1991, 2000, 2001, 2007) and Croft and van Lier (2012), we argue that just as there are two states of existence of word at the two levels of langue (i.e. word type in lexicon in a communal language) and parole (i.e. word token in syntax), word class categorization also happens at the two levels: the word token categorization in syntax at parole is the speaker’s expression of propositional speech act functions like reference, predication and modification, whereas the word type categorization in the lexicon at langue is the conventionalized propositional speech act functions of a word type resulted from self-organization or collective unconscious. The class membership of a word type or lexeme in the lexicon at langue does not have a priori existence, nor is it precategorial, but is liable to change through recurrent use in various propositional speech act constructions in syntax at parole; The multifunctionality or multiple class membership of word types is closely related to frequency (including token frequency and type frequency ); The class membership (either single or multiple class membership) of a word type in the lexicon of a communal language is its meaning potential(s) at langue, which is to be discovered by descriptive linguists through corpus-based usage pattern surveys, as is done by dictionary compilers in word class labeling.

The multifunctionality of lexemes is a common feature of typical analytic languages like Chinese and English (Enfield, 2015: 47; Haselow, 2011: 1-2; Robins, 1989: 214), but few studies have been conducted from the perspective of language dictionaries (i.e., the lexicon in communal language). As Hanks (2013: 348) puts it, “Between the ancients and the moderns, many philosophers speculated about meaning, but as far as understanding the nature of the lexicon is concerned, nothing much of any interest happened for over a thousand years.” Obviously, only by studying word classes based on dictionaries can we really touch upon the nature of word classes. As Munro (2005:307) points out, “Making dictionaries is a vital aid to completing a full grammatical analysis of a language, particularly if the dictionary requires the specification of the part of speech for each entry.” Therefore, only on the basis of the dictionaries in Modern Chinese and Modern English can we make a systematical investigation of heterosemy in these two languages.
2.2 Heterosemy in Modern Chinese from the Perspective of Dictionaries

2.2.1 A Survey of Heterosemy in Modern English Dictionaries

Based on *Oxford Advanced Learner's Dictionary* (7th edition) (hereinafter referred to as OALD-7), Wang (2014) conducts a survey of heterosemy in Modern English. It is found that heterosemous lexemes account for 10.48% of the 46380 entries in the dictionary, and that the percentage of heterosemy in the key words of the Oxford 3000 in OALD-7 is as high as 48.24%. Moreover, the dictionary has 125 kinds of heterosemy, among which heterosemy of two categories accounts for the absolute majority (91.15%), followed by heterosemous lexemes of three categories, while heterosemous lexemes of four categories, five categories or even six categories are very rare. There are 27 kinds of heterosemous lexemes of two categories, among them noun-verb heterosemy occupies the largest percentage (59.65%), followed by noun-adjective heterosemy (28.12%), adjective-adverb heterosemy (5.55%) and noun-adverb heterosemy (1.94%); There are 29 kinds of heterosemy for 358 lexemes belonging to three categories, among which noun-verb-adjective heterosemy accounts for 54.19%. In other words, the heterosemy of lexemes in Modern English is not only very common, but also has a wide variety, and the heterosemy among nouns, verbs, adjectives and adverbs is the most common. The study has obviously shown that the multifunctionality of lexemes is a common feature of analytic languages, and that it is closely related to the frequency of use.

2.2.2 A Survey of Heterosemy in Modern Chinese Dictionaries

Based on *Contemporary Chinese Dictionary* (5th edition) (CCD5 for short), Wang (2013) surveys the heterosemy in Modern Chinese. It is found that the heterosemous lexemes only account for 5.40% in the 51469 entries of the whole dictionary, and the proportion of the heterosemous lexemes of two categories is the largest (95.36%), followed by heterosemous lexemes of three categories, while heterosemous lexemes of four categories and five categories are very rare; Among the heterosemous lexemes of two categories, there are 37 different cases of heterosemy, in which there are 2427 heterosemous lexemes of Chinese content words (nouns, verbs, adjectives and adverbs), accounting for 91.62%. However, compared with the survey of heterosemy in OALD-7, we have found a striking difference: the proportion of heterosemous lexemes in Modern Chinese in CCD5 is only about half of that in Modern English (5.40% / 10.48%). According to the survey in CCD5, the number of heterosemous lexemes has been artificially minimized in order to stick to the Principle of Parsimony. However, the heterosemy presented in CCD5 is partially consistent with the reality of language use in the Chinese community, which reveals the invalid theoretical basis and conflicting practice in the studies of Chinese word classes.
2.3 Negative Impacts of the Ban on Heterosemy in Modern Chinese Dictionaries and Corpora

As will be shown below, the view of “fewest possible heterosemous lexemes” based on the Principle of Parsimony has left a negative impact on the word class labeling in Modern Chinese dictionaries and on the POS tagging in Modern Chinese corpora, with the heterosemy of self-reference lexemes the most conspicuous.

2.3.1 Misrepresentations of Heterosemy of Self-reference Lexemes in Modern Chinese Dictionaries

Based on the studies of some Modern Chinese dictionaries, it is found that the conventionalized self-reference lexemes with symmetrical semantic relations (i.e. antonymy, synonymy and hyponymy) are contradictory within the same dictionary, and the same entries are also contradictory in different dictionaries.

1) The Inconsistencies of Similar Entries in the Same Chinese Dictionary

Symmetry turn out to be a universal phenomenon existing in the microscopic and macroscopic worlds, and language is no exception. Considering that the extension of word meanings generally follows the mechanism of analogy, Wang (2005) points out that those entries with symmetrical semantic relations like antonyms, synonyms and co-hyponyms are more likely to be symmetrical in word class labeling and definition, and that symmetry is undoubtedly a convenient and efficient operating principle and problem finding method, although asymmetric cases do exist sometimes. Nevertheless, we have found that some Chinese dictionaries do not hold the principle of symmetry well in the process of handling the similar entries.

As shown in Table 1, biànhuà is treated as a noun-verb heterosemous lexeme in CCD6, and both zhìbiàn and liàngbiàn are labeled as nouns, corresponding to the nominal sense of biànhuà; The two senses of tūbiàn are both labeled as verbs, but the second sense with the same meaning as zhìbiàn obviously is a self-reference usage, which should be labeled as nouns; jiànbìàn is only labeled as a verb in CCD6, and it seems that its self-reference nominal sense is omitted in the dictionary.

Table 1 The Asymmetrical Treatment of Homogeneous Entries with biànhuà in CCD6

<table>
<thead>
<tr>
<th>单词</th>
<th>词性</th>
<th>含义</th>
</tr>
</thead>
<tbody>
<tr>
<td>biànhuà</td>
<td>动</td>
<td>事物在形态上或本质上产生新的状况：形势～得很快。</td>
</tr>
<tr>
<td></td>
<td>名</td>
<td>事物在形态上或本质上产生新的状况：这几年家乡的～特别大。</td>
</tr>
<tr>
<td>zhìbiàn</td>
<td>名</td>
<td>事物的根本性质的变化。是由一种性质向另一种性质的突变。参看 814 页【量变】</td>
</tr>
</tbody>
</table>
量变

liàngbiàn 名事物数量上、程度上的变化。是一种逐渐的不显著的变化，是质变的准备。参看1679页《质变》

突变

tūbiàn 突然急剧地变化：时局～| 神色～| 基因～。②哲学上指飞跃。也叫质变。

渐变

jiànbìan 逐渐变化。

2) Inconsistencies of the Same Entries in the different Modern Chinese Dictionaries

As shown in Table 2, the same five lexemes are treated quite differently in the Standard Dictionary of Modern Chinese (3rd edition)(SDMC3 for short): No sense is labeled as nouns, even if for philosophical terms like zhibiàn and liàngbiàn, whose definitions are very close to those in CCD6,and none of these lexemes is considered heterosemous.

Table 2 The Treatment of the Homogeneous Entries with biànhuà in SDMC3

<table>
<thead>
<tr>
<th>变化</th>
<th>biànhuà</th>
<th>无常</th>
<th>突变① ▷气～</th>
<th>语言～的过程</th>
<th>由～到突变。②动</th>
<th>量变</th>
</tr>
</thead>
<tbody>
<tr>
<td>质变</td>
<td>zhibiàn 动</td>
<td>哲学上指事物根本性质变化，是事物由旧质向新质飞跃（跟“量变”相区别）。</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>量变</td>
<td>liàngbiàn 动</td>
<td>哲学上指事物逐渐的不显著的变化，包括数量的增减、场所的更换、构成成分排列次序和结构形式的变化等（跟“质变”相区别）。也说渐变。</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>突变</td>
<td>tūbiàn 突然发生变化（跟“渐变”相区别）▷ 形势～。②动</td>
<td>质变。</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>渐变</td>
<td>jiànbìan 动</td>
<td>缓慢地变化（跟“突变”相区别）▷ 语言～的过程</td>
<td>由～到突变。②动</td>
<td>量变。</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the case of sports lexemes, among the 16 lexemes in CCD6 like báhé, dēngshān, huáxiáng, huáxué, jūzhòng, pányàn, púyōng, qiánshuí, sàitíng, sǒndō, shējī,shējiàn,
13 of them include the self-reference sense of “a kind of sports”, among which 10 are labeled as verbs and 3 as nouns, but none of them is treated as verb-noun heterosemy. By contrast, among the above 16 sports lexemes in SDMC3, 13 of them include the self-reference sense of “a kind of sports”, among which 12 are labeled as nouns and 1 as verbs, and thus 12 of them are treated as verb-noun heterosemy. See Table 3 for details:

**Table 3 A Contrastive Representation of Sports Lexemes in CCD6 and SDMC3**

<table>
<thead>
<tr>
<th>体育项目</th>
<th>《现汉》（第6版）</th>
<th>《规范》（第3版）</th>
</tr>
</thead>
<tbody>
<tr>
<td>拔河</td>
<td>bá/he 一种体育运动，人数相等的两队队员，分别握住长绳两端，向相反方向用力拉绳，把绳上系着标志的一点拉过规定界线为胜。</td>
<td>báhe 一种体育运动。人数相等的两队各执长绳的一端，同时向相反的方向用力拉，把长绳中间的标志拉过规定界线的一方获胜。</td>
</tr>
<tr>
<td>登山</td>
<td>dēng/shān ①上山：～临水～越岭。②一种体育运动，攀登高山：～服～协会。</td>
<td>dēngshān ①上山：攀登山峰～运动。</td>
</tr>
<tr>
<td>滑翔</td>
<td>huáxiáng 某些物体不依靠动力，而利用空气的浮力和本身重力的相互作用在空中飘行。</td>
<td>huáxiáng 不依靠动力，只借助空气的浮力在空中飘行。</td>
</tr>
<tr>
<td>滑雪</td>
<td>huá/xué ①脚蹬滑雪板，手撑滑雪杖在雪地上滑行。②体育运动项目，两脚固定在滑雪板上，项目之一，比赛分为高山滑雪、跳台滑雪和花样滑雪等。</td>
<td>huáxué ①在雪地上滑行。②体育运动项目，两脚固定在滑雪板上，手撑两支滑雪杖，在雪地上滑行～高山～。</td>
</tr>
<tr>
<td>举重</td>
<td>jǔzhòng 体育运动项目之一，运动。</td>
<td>jǔzhòng①两手举起杠铃。②名指</td>
</tr>
<tr>
<td>运动项目</td>
<td>定义</td>
<td>备注</td>
</tr>
<tr>
<td>----------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>挺举、抓举</td>
<td>举重运动，使用杠铃。</td>
<td>分挺举、抓举两种。</td>
</tr>
<tr>
<td>攀岩</td>
<td>一种体育运动，只使用少量器具，攀登陡峭的岩壁。</td>
<td>分攀岩和速降。</td>
</tr>
</tbody>
</table>
| 跳水 | 潜水运动项目之一，比赛时运动员脚穿脚蹼，戴泳镜，口衔呼吸管等，以潜水游泳速度快慢决定胜负。 | 指该项目。
| 潜水 | 在水面以下活动。 | 指潜水。
| 赛艇 | 水上运动项目之一，在比赛航道内用人力划桨使艇前进。分为单人、双人、四人和八人。 | 按照有无舵手、划船人数、单桨还是双桨又可分为不同小项。
| 散打 | 体育运动项目之一，双方按照规则，利用踢、打、摔等方式徒手搏击、对抗。 | 指这种竞技体育项目。
| 射击 | 用枪炮等火器向目标发射子弹。 | 指射击。

以上运动项目介绍了举重、攀岩、跳水、潜水、赛艇、散打和射击等运动项目的基本定义和特点。
<table>
<thead>
<tr>
<th>运动项目</th>
<th>发射弹头。②体育运动项目之一，按照比赛时所用枪支、射击距离、射击目标和射击姿势，分为不同项目。</th>
<th>shè jiàn①动(-/-)利用弓的弹力把箭射出去。②名竞技体育项目，按照规定把子弹射向目标，以命中环数或靶数计算成绩。</th>
</tr>
</thead>
<tbody>
<tr>
<td>箭</td>
<td>shè jiàn</td>
<td>shè jiàn</td>
</tr>
<tr>
<td>竞技体育项目之一，按照规定把箭射向目标，以命中环数或靶数计算成绩。</td>
<td></td>
<td></td>
</tr>
<tr>
<td>摔跤</td>
<td>shuāi jiāo ①动摔倒在地上；摔了一跤。路太滑，一不小心就要摔一跤。②体育运动项目之一，两人相抱运用力气和技巧，以摔倒对方为胜。</td>
<td></td>
</tr>
<tr>
<td>运</td>
<td>shuāi jiāo</td>
<td></td>
</tr>
<tr>
<td>摔跤</td>
<td>shuāi jiāo</td>
<td></td>
</tr>
<tr>
<td>①动两人徒手相搏，按照一定规则，摔倒对方者为胜 ②名指这种体育运动项目。</td>
<td></td>
<td></td>
</tr>
<tr>
<td>跳伞</td>
<td>tiào sān ①动人携带降落伞从飞机上或跳伞塔上跳下，借助空气浮力，缓缓降落到地面。②名军事体育项目，包括飞机跳伞、氷气球跳伞和伞塔跳伞。</td>
<td></td>
</tr>
<tr>
<td>跳水</td>
<td>tiào shuǐ ①动水上体育运动项目，从跳台或跳板上跳入水中，身体在空中做出各种优美的动作，跳板跳水和跳台跳水两种。②名水上体育运动项目，分为跳台跳水、跳板跳水和跳台跳水三种。</td>
<td></td>
</tr>
<tr>
<td>跳水</td>
<td>tiào shuǐ</td>
<td></td>
</tr>
<tr>
<td>①动运动员从跳板或跳台上起跳后，在空中做各种规定的和自选的动作，然后落入水中。 ②动比喻证券价格、指数等急剧上升。</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
后入水。以两种动作得分的总和决定名次。

动跳入江河湖海自杀。

动比喻证券价格、指数等暴跌。

大盘~|~股值~。

In other words, CCD6 and SDMC3 employ fundamentally different ways in treating the heterosemy of self-reference lexemes, although the two dictionaries include the same number of sports lexemes (13 in both dictionaries): Most of the lexemes are labeled as verb-noun heterosemy in SDMC3 (accounting for 92.31%), while none of them is treated as verb-noun heterosemy in CCD6 which obviously follows the mainstream view in Modern Chinese grammar (i.e., the so-called Principle of Parsimony). What’s more, as shown in Table 4, the word class membership of yóuyǒng changes from a verb-noun heterosemous lexeme in CCD5 into a verb only in CCD6. However, it is found that while the self-reference senses of most sports lexemes in CCD6 are labeled as verbs, similar usages of āi zhòng, pù yǒng and sǎndào are labeled as nouns rather than verbs. It is evident that the lexicographers of CCD hold conflicting views on the treatment of self-reference usages. Although in SDMC3 most of the usages of sports lexemes to express the names of sports events are labeled as nouns, báhé is labeled as a verb. However, its typical usage as a verb, as in Tāmen zhèngzài cāochǎng shǒng báhé, has not been recorded in the dictionary.

Table 4 Discrepancies in Representations of yóuyǒng in CCD5 and CCD6

<table>
<thead>
<tr>
<th>Lexeme</th>
<th>CCD5</th>
<th>CCD6</th>
</tr>
</thead>
<tbody>
<tr>
<td>yóu/yǒng</td>
<td>人或动物在水里游动。体育运动项目之一，人在水里用各种不同的姿势划水前进或进行表演。</td>
<td>人或动物在水里游动。运动项目，运动员在水中按既定姿势划水前进，比赛速度，或进行动作表演。</td>
</tr>
</tbody>
</table>
2.3.2 POS Tagging of Self-reference Usages in Modern Chinese Corpora

Based on the annotated data of 20 million words in Chinese National Corpus, we have retrieved the POS tagging of the two groups of lexemes mentioned above and made further statistical analyses. As shown in Table 5, the POS tagging of the above five lexemes related to biànhuò in Chinese National Corpus is consistent with their word class labeling in CCD6 (but it diverges considerably from that in SDMC3): biànhuò is treated as verb-noun heterosemy, zhìbiàn and liàngbiàn are tagged as nouns except one usage as a verb respectively, while tūbiàn and jiānbiàn both tagged as verbs.

| verbs | 8415 | 1 | 1 | 523 | 50 |
| nouns | 1549 | 159 | 106 | 0 | 0 |

In addition, among the above 16 sports lexemes, 15 of them (except sǒndō) are recorded in Chinese National Corpus. As shown in Table 6, 12 of them have the verb tags, 7 have noun tags, and 4 lexemes are treated as verb-noun heterosemy. Similarly, the POS tagging of these lexemes in Chinese National Corpus is much the same as word class labeling in CCD6, but it is widely divergent in SDMC3.

| verbs | 15 | 106 | 23 | 31 | 2 | 0 | 0 | 232 |
| nouns | 0 | 0 | 0 | 0 | 52 | 1 | 30 | 0 |
According to the Two-level Word Class Categorization Theory, word class labeling in language dictionaries deals with the class membership of word types, whereas the POS tagging in corpora deals with the class membership of word tokens, the latter of which should be tagged on the basis of their propositional speech act functions in sentences at parole. But it is regrettable that the POS tagging of the above two groups of words in Chinese National Corpus has the following two types of confusion:

① No matter which propositional speech act function the word token usage is represented in the sentence, it is tagged as verbs without exception. For example, tiàoyuǎn in example (1), which is used to denote a sports event, is tagged as a noun, but bódé, whose propositional speech act function is apparently reference (i.e., a noun in syntax), is still labeled as a verb; similarly huáxué, shuāijīào, shèjiàn and tiàosànm in examples from (2) to (5), which are used as the names of sports events, are all tagged as verbs. Tūbiàng in the following concordances are tagged as verbs in all cases, but it is obvious that the self-reference usages in examples (6) and (7) should be labeled as nouns. So is the case with jiànbìàn in examples (8) and (9), as both liángbìàn and zhībìàn, which are in coordinate relation in example (9), are tagged as nouns. It seems that the POS tagging of word tokens in Chinese National Corpus tends to follow the Principle of Parsimony like a Chinese dictionary, without taking context into account.

(1) /w 跳远/n 、/w 赛跑/v 、/w 拔河/v 等/u 二十/m 多项/mq 比赛/v 。/w
(2) 雪上/nl 表演/v 项目/n 自由/a 式/k 滑雪/v 今天/nt 进行/v 了/u 第二/m 轮/n
(3) 冠军/n 。/w 柔道/n 即/v 日本式/n 摔跤/v, /w 在/p 国际上/n 开展/v 比较/d
(4) 在/p 男子/n 棒球/n 、/w 女子/n 射箭/v 和/c 跆拳道/n 3/m 个/q 项目/n
(5) 形式/n , /w 从事/v 射击/v , /w 跳伞/v, /w 摩托车/n , /w 滑翔/v 、/w
(6) 基础/n , /w 水资源/n 价值/n 的/u 突变/v 会/vu 对/p 社会/n 经济/n 产生
（7）引起生物遗传基因的突变，并能真实遗传。
（8）这个过程可能存在渐变和突变两种情形。
（9）量变到质变或者说由渐变到突变的转化规律。

② The POS tagging of word tokens in Chinese National Corpus turns out to be casually done for their typical self-reference usages. Besides the asymmetrical treatment between jǐzhòng and tūbiàn as well as between zhìbiàn and liàngbiàn in example (9), such word tokens which express the names of sports events as jǔzhòng and pǔyǒng mentioned above are all tagged as nouns in the corpus, while the treatment of the rest of the sports lexemes is rather confusing. For example, tǐdiào as a self-reference usage to sports events is labeled as a noun in example (10), but it is labeled as verbs in examples from (11) to (13); Similarly, yóuyǒng is labeled as a noun in example (14) while as verbs in examples from (15) to (17).

（10）体操、游泳、跳水、羽毛球、男篮等。
（11）排球、体操、跳水、乒乓球、羽毛球等。
（12）不，跳水和音乐关系很密切。
（13）十九世纪末尾，跳水传到美国，落了。
（14）分别增加了解自行车和游泳等项目。
（15）的？门票大多集中游泳、体操和跳水比赛。
（16），射击、田径、游泳等项目？的？比赛。
（17）大型代表团参赛，游泳、跳水、田径，
3. Lifting the Ban on Self-reference Heterosemy is the Key to Solving the Problem of Category Indeterminacy in Modern Chinese

Hu (1996a: 7) has clearly stated that a halfway standard of syntactic functions has been adopted in the classification of word classes in the prevailing grammar system, namely, the criterion of syntactic functions is used in establishing the principle of classification on the one hand, but the semantic criterion is adopted in the classification of lexemes on the other. The reason why double criteria are adopted is that some scholars worry that the multifunctionality of lexemes without inflections will result in the underspecification or category indeterminacy of word classes in Modern Chinese, and thus they attempt to fix multifunctional lexemes into single categories. As a result, though multifunctional lexemes have fixed single categories in Chinese, the word classes themselves have become multifunctional. Shen (2009) is also in support of the above view.

From the perspective of the Two-level Word Class Categorization Theory, heterosemy refers to the multifunctionality of lexemes in lexicon in a communal language (i.e., language dictionaries) and there is a positive correlation between the heterosemy of lexemes and frequency (i.e. only high frequency lexemes have the possibility of heterosemy); heterosemy does not occur in word tokens in sentences at parole. It is found that previous studies do not get the whole picture of the heterosemy in Modern Indo-European languages, and most of them are confined to contrastive studies on class membership of word tokens between Chinese and some Indo-European languages in syntax at parole. Wang (2015: 23-24) finds that Modern Italian, which belongs to the Romance languages in the Indo-European family, also shows some analyticity and heterosemy in lexicon in a communal language: the usages to express color of 7 basic color lexemes are all treated as the heterosemous lexemes of an adjective and a noun.Wang and Zhou (2015) makes a corpus-based study on the correlation between the heterosemy in Modern Chinese and frequency, and it shows the correlation between the heterosemy of lexemes and frequency results from the competing motivations of economy and iconicity in communication. The Principle of Parsimony, which holds that there should be fewest possible heterosemous lexemes in Modern Chinese, violates the basic definition of heterosemy and applies the law of contradiction improperly, and runs counter to the correlation between heterosemy and frequency; The word class labeling in CCD tends to follow the Principle of Parsimony which is closely related to first-order logic; The reason why word class categorization in such analytic languages as Chinese has remained unsettled as the “Goldbach Conjecture” for centuries is that previous studies take languages as natural objects which has nothing to do with language use rather than culturally inherited artifacts.

From the macroscopic contrast in Section 2.2.2 and the microscopic analyses in Section 2.2.3, it is shown that following the Principle of Parsimony so as to minimize the number of heterosemous lexemes has left a negative impact on the word class labeling in Modern Chinese dictionaries and on the POS tagging in Modern Chinese corpora: The number of heterosemous lexemes has been minimized, creating a false impression that there is less heterosemy in Modern Chinese than in Modern English;
The conventionalized self-reference senses for those lexemes with symmetrical semantic relationship have contradictory representation of their word class information both within one dictionary and between dictionaries; For the derived self-reference senses of primary verbs, the POS tagging in Modern Chinese corpora is also in chaos. Hu (1996b: 215) claims that the heterosemy, especially for verb-noun heterosemy and adjective-noun heterosemy, is a long-standing problem which has troubled Chinese grammar researchers a lot. Based on the DIY database of word class labeling information, Wang and Wang (2016) makes a systematic study of the representation strategy of verb-noun heterosemy in OALD8. It is found that verb-noun heterosemy is very common in Modern English, which accounts for 5.8% of the lexemes in the dictionary and approaches the percentage of verb-only lexemes in it; Among the verb-noun heterosemous lexemes in its strict sense, 73.79% of them include self-reference senses, which are represented in a diversified manner and is conducive to satisfying English learners’ encoding needs. The representation strategies of verb-noun heterosemous lexemes in OALD8 (especially self-reference lexemes) not only have some implications for theoretical studies in linguistics, but also can shed lights on the word class research in Modern Chinese and word class labeling in Modern Chinese dictionaries. Thus, lifting the ban on self-reference heterosemy is the key to the word class underspecification in Modern Chinese and a way out of the dilemma of word class labeling in Chinese dictionaries.

4. Conclusion

In conclusion, heterosemy refers to the multifunctionality of lexemes in such analytic languages as Modern Chinese and Modern English. Those scholars who hold the Principle of Parsimony in dealing with multifunctionality of lexemes take language as natural objects having nothing to do with its use instead of a cultural artifact, and they do not make a distinction between the word class categorization of word types/lexemes and that of word tokens, among which the treatment of self-reference lexemes has been the most controversial topic in Modern Chinese grammar and lexicography. The study shows that the handling of self-reference lexemes in both word class labeling in Modern Chinese dictionaries and POS tagging in Modern Chinese corpora has a lot of contradictions under the guidance of the Principle of Parsimony. Therefore, lifting the ban on self-reference heterosemy is an important way out of the Chinese word class dilemma.
References


Lin, Q. L. and J. Q. Han. 2011. Parts of speech in Chinese from the perspective of


Wang, R. Q. 2013. A study of multiple class membership in modern Chinese with a comment on the significance of linguistic theories of Ferdinand de Saussure.


On Multicategorization of Céngjīng from the Perspective of the Two-level Word Class Categorization Theory

Yongqiang Ma
Graduate School, Sichuan International Studies University
yongqiangma2008@126.com

Abstract

In recent years more and more studies turn their attention to the multicategorization of céngjīng, and the focus of the existing arguments falls on the multifunctionality of céngjīng and the determination and delimitation of its word class membership. From the perspective of Two-level Word Class Categorization Theory, this paper is intended to conduct a corpus-based survey of usage patterns of céngjīng, and to investigate the conventionalization of the adjective usage and the noun usage of céngjīng. It is found that céngjīng has relatively high token frequency and very high type frequency in its adjective and noun usage, the lexeme of céngjīng belongs to adverb, adjective and noun instead of adverb only in the lexicon of the communal language of Modern Chinese, which should be labeled as a heterosemous lexeme in large and medium-sized dictionaries of contemporary Chinese.

Keywords: céngjīng, multicategorization, multifunctionality, lexeme, Conventionalization

1. Introduction

Word class categorization is of paramount importance in linguistic studies, just as Arts (2007: 10) argues, “in present-day linguistics, it is safe to say, no grammatical framework can do without categories”. However, it is not consistent in the use of the term “category” in grammatical theories (Lyons 1968: 270), and linguists define what they regard as linguistic categories in different ways. The majority holds that the categories of language are the word class. Word class categorization in such typical analytic languages as Modern Chinese and Modern English is compared to “Goldbach Conjecture” (Wang and Zhou 2015; Wang and Huang 2017), which has always been the knotty and heated issue in grammar research and lexicography, especially the the multifunctionality of lexemes (Rauh 2010; van Lier 2012, 2016; van Lier and Rijnhoff 2013; Enfield 2015; Guo 2002; Shen 2009, 2012, 2016; Lu 2013; Wang 2006, 2009, 2013, 2014b). The multifunctionality of lexemes is a widespread phenomenon, and it is a property of all human lexicons (Lefebvre 2001: 107), and it refers to the phenomenon of a polysemous lexeme has two or more word classes in the lexicon of
the communal language system, which is represented as a heterosemous lexeme in language dictionaries (Wang 2013: 12). The previous studies of multifunctionality of lexemes invoke a variety of problems, mostly which are closely related to the delimitation and determination of word class membership, that is, there exist many disputes about whether polysemous lexemes, which usually have two or more word classes in the lexicon of the communal language system, should be treated as homographs, heterosemy of lexemes or multifunctionality of certain types of word class, and even some new terms are given to those lexemes, such as “nouny verbs” (Zhu 2010) and “flexibles” (Luuk 2010), etc.

The multicategorization of cèngjīng has aroused the concern of many scholars. It is generally accepted that cèngjīng is only an adverb, functioning as adverbial in Chinese grammar. However, cèngjīng began to appear some non-adverbial usages in recent years, such as attributive, subject and object. The previous studies of the multicategorization of cèngjīng mainly focus on the appropriateness and acceptance of those structures like “cèngjīng + de + NP (X)” in Modern Chinese (Lei 2003; Wen 2003; Zhang 2003a; Xu 2008) and the emergence of non-adverbial usages of cèngjīng (Zhang 2003b; Xu 2015). Among them, Zhang (2003b) and Xu (2015) are highly correlated with the present study. The former argues that the non-adverbial usages of cèngjīng has extended from its adverbial usage, and it is concluded that Chinese adverbs are multifunctional. The latter claims that cèngjīng should be treated as a temporal noun according to the syntactic distribution and collocational competence, and its non-adverbial usages result from the nominalization caused by metonymy. If we take a second look at the both studies, it is obvious that their conclusions cannot hold water. The above two studies have some shortcomings. First, the two studies are not well-supported by data from large corpora, and the conclusions obtained by way of giving examples are not convincing because they cannot reflect the reality of language use in the Chinese community; Second, the logical fallacy of overgeneralization has been made by claiming that Chinese adverbs are multifunctional rather than the lexeme of cèngjīng, because it can be easily proved that not all Chinese adverbs have non-adverbial usages; Third, The wrong standard, level and procedure have been adopted to judge word class membership. The determination and delimitation of word class membership based on its syntactic distribution can contribute to the dilemma of “one word, one class” or the mess of lumping word classes together. Therefore, further discussions on the following questions are badly needed: What are the actual usage patterns of cèngjīng in Chinese language community? Whether its usages of adjective and noun are conventionalized? How to determine and delimitate the word class membership of cèngjīng?

Thus, from the perspective of Two-level Word Class Categorization Theory, this article intends to conduct a corpus-based survey of the usage patterns of cèngjīng and investigate the conventionalization of its usages of adjective and noun so as to truly recognize the process of multicategorization of cèngjīng and hopefully provide some guidance for word class labeling in Modern Chinese dictionaries and POS tagging in Modern Chinese corpora.
2. Research Design

Based on the perspectives of language as a complex adaptive system and the nature of major parts of speech as propositional speech act functions proposed by Croft (1991) and Croft & van Lier (2012) on the basis of Searle (1969), the Two-level Word Class Categorization Theory is proposed to solve the multifunctionality of lexemes in analytic languages, such as Modern Chinese and Modern English (Wang 2014a). It holds that there are two states of existence of word both in lexicon in a communal language at the level of langue (hereinafter referred to as word type or lexeme) and in syntax at the level of parole (hereinafter referred to as word token), and word class categorization happens at the two levels, namely word token categorization and word type categorization; the former is the speaker’s expression of propositional speech act functions like reference, predication and modification, whereas the latter refers to a collective process of self-organization of a speech community, its main focus lies in conventionalization/phase transition; The class membership of a word type changes through recurrent uses in syntax at parole, and the multifunctionality of a word type is closely related to frequency of use (including token frequency and type frequency); The class membership (either single or multiple class membership) of a word type is its meaning potential(s) at langue and the conventionalized propositional speech act functions of word tokens in syntax at parole, which is to be discovered by descriptive linguists through corpus-based surveys of usage patterns, and it is also a process of categorization based on the sampling analysis of propositional speech act functions in syntax at parole and the degrees of conventionalization of a word type, which is typically represented as word class labeling in language dictionaries.

A number of empirical studies have been done based on the Two-level Word Class Categorization Theory, and they are mainly concerned with such topics as the heterosemy of lexemes in Modern Chinese (Wang 2013) and in Modern English (Wang 2014b; Wang and Wang 2016), the relationship between the verbs and constructions (Wang and Chen 2014), the correlation between heterosemy and frequency in Modern Chinese (Wang and Zhou 2015) and the heterosemy of self-reference lexemes in Modern Chinese (Wang and Huang 2017). The studies above show that the Two-level Word Class Categorization Theory bear great values both in academic research and in practice.

After considering the representativeness and applicability of different corpora, we choose the Chinese National Corpus (CN Corpus), Corpus of Center for Chinese Linguistics at Peking University (CCL Corpus), National Language Resources Dynamic Circulation Corpus (DCC Corpus) and Title Search from Chinese National Knowledge Infrastructure (CNKI Title Search) to make an investigation on the usage patterns of céngjīng. According to the standards of conventionalization (Wang and Chen 2014), we are going to analyze the degrees of conventionalization of the modification and reference usages of céngjīng on the basis of the statistics of its token frequency and type frequency, in the hope that the multicategorization of céngjīng can be well recognized and its word class membership can be clearly defined. As far as the chosen corpora are concerned, CN Corpus is a large-scale balanced corpus built in
1998 with widespread materials and long time span, containing one hundred million characters; CCL Corpus is established by the Center for Chinese Linguistics at Peking University at the beginning of 21st Century, which covers 581,794,456 characters; DCC Corpus, containing ten billion characters, is developed by National Language Resource Monitoring and Research Center at Beijing Language and Culture University, whose data are mainly from 17 mainstream newspapers media such as People’s Daily; CNKI, developed by Tsinghua University, is the most advanced and widely-used publishing platform of Chinese academic resources, collecting the full text data from domestic academic journals, doctors’ dissertations and masters’theses, newspapers, conference papers, reference books, yearbooks and so on.

In terms of the data processing, we firstly collect all the usage patterns of céngjīng in different corpora, and then remove such noises as “zēng jīnglǐ”, “céng jīnglì”, “céng jīngcháng (cháng)” and “céngjīngcānghǎi”. Finally the usage patterns of céngjīng collected from corpora are analyzed as follows: 1) The token frequency of céngjīng functioning as adverbs, adjectives and nouns are counted respectively; 2) Given that there is no dispute about its adverb usage because of its long existence, the type frequency of the adjective and noun usages of céngjīng, mainly focusing on the types of the headwords modified by céngjīng as well as the types of its modifiers when it is used as a noun; It should be noted that the central headwords modified by céngjīng are extracted without breaking the integrality of the meaning of the original constructions, and the proper nouns and idioms are taken as a whole, such as “céngjīng de diyī fùrén”, “céngjīng de ‘jiàohǎo bù jiàozuò’”, “céngjīng de niánshào qǐngkuáng” and “céngjīng de ‘fùmǔzhīmìng, méishuòzhīyán’”, and so forth; Moreover, the type frequency of such coordinate constructions as “céngjīng de huīhuáng hé cāngsāng” (two cases) “céngjīng de fánhuá, xuānnào, xīngwàng” (three cases) are counted respectively; 3) On the basis of the statistics from CNKI Title Search, the synchronic distribution and diachronic development of the usages as adjectives and nouns of céngjīng are investigated.

3. Research Results

3.1 Token Frequency

As shown in Table 1, the adverb usage of céngjīng in four corpora all occupies an overwhelming majority, accounting for 76.74%, 85.55% and 98.93% in CNKI Title Search, DCC Corpus and CCL Corpus respectively, and it even reaches 100% in CN Corpus. Since it was established earlier (1998), there are no adjective and noun usages of céngjīng found in the corpus. Data retrieved in DCC Corpus are mainly from 17 mainstream newspaper media in 2012, and data are collected from CNKI Title Search by the end of 2016. It is found that the adjective and noun usages of céngjīng begin to appear and then increase gradually with the progress of time, from zero to one and from one to many. The growing of the adjective usage is the most typical, the proportion accounts for 14.29% in DCC Corpus, and 20.69% in CNKI Title Search.
Search which has latest data. Although the noun usage of céngjīng occupies a small proportion, it is really a great leap growing from 0 to 2.57%.

**Table 1** The Statistics of Adverb, Adjective and Noun Usages of Céngjīng in Corpora

<table>
<thead>
<tr>
<th></th>
<th>adverb usage</th>
<th></th>
<th>adjective usage</th>
<th></th>
<th>noun usage</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>frequenc y</td>
<td>proportion</td>
<td>frequenc y</td>
<td>proportion</td>
<td>frequenc y</td>
<td>proportion</td>
<td></td>
</tr>
<tr>
<td>CN Corpus</td>
<td>1317</td>
<td>100.00</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>1317</td>
</tr>
<tr>
<td>CCL Corpus</td>
<td>31876</td>
<td>98.93</td>
<td>331</td>
<td>1.03</td>
<td>14</td>
<td>0.04</td>
<td>3222 1</td>
</tr>
<tr>
<td>DCC corpus</td>
<td>32843</td>
<td>85.55</td>
<td>5485</td>
<td>14.29</td>
<td>64</td>
<td>0.16</td>
<td>3839 2</td>
</tr>
<tr>
<td>CNKI Title Search</td>
<td>4124</td>
<td>76.74</td>
<td>1112</td>
<td>20.69</td>
<td>138</td>
<td>2.57</td>
<td>5374</td>
</tr>
</tbody>
</table>

**3.2 Type Frequency**

Based on the corpus-based surveys, it is found that that the adjective usage of céngjīng has very high type frequency. There are 247, 677 and 2482 different headwords modified by céngjīng respectively in CCL Corpus, CNKI Title Search and DCC Corpus, and Table 2 shows the top 20 modified words of céngjīng in different corpora. It should be noted that there are a large number of words used just once, 208 in CCL Corpus, 556 in CNKI Title Search and 1696 in DCC Corpus, and they have great potentials to be copied in language uses. Therefore, it can be concluded that the adjective usage of céngjīng have very high productivity.

**Table 2** Type Frequency of Céngjīng as Adjectives in Different Corpora

<table>
<thead>
<tr>
<th></th>
<th>CCL Corpus</th>
<th>CNKI Title Search</th>
<th>DCC Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modified Word</td>
<td>Frequenc y</td>
<td>Modified Word</td>
<td>Frequenc y</td>
</tr>
<tr>
<td>1</td>
<td>huīhuáng</td>
<td>17</td>
<td>huīhuáng</td>
</tr>
<tr>
<td>2</td>
<td>... zhīlū</td>
<td>6</td>
<td>nī</td>
</tr>
<tr>
<td>3</td>
<td>líchéng</td>
<td>6</td>
<td>suìyuè</td>
</tr>
<tr>
<td>4</td>
<td>mèngxiàng</td>
<td>6</td>
<td>mèngxiàng</td>
</tr>
<tr>
<td>5</td>
<td>líxiàng</td>
<td>6</td>
<td>jǐyǐ</td>
</tr>
<tr>
<td>6</td>
<td>kānkē</td>
<td>4</td>
<td>zǐjǐ</td>
</tr>
<tr>
<td>7</td>
<td>cāngsāng</td>
<td>4</td>
<td>ài</td>
</tr>
<tr>
<td>8</td>
<td>shàonián</td>
<td>4</td>
<td>měi lǐ</td>
</tr>
<tr>
<td>9</td>
<td>tóngxué</td>
<td>4</td>
<td>měihào</td>
</tr>
<tr>
<td>10</td>
<td>nǎrén</td>
<td>3</td>
<td>shēnghuó</td>
</tr>
<tr>
<td>11</td>
<td>xīnbīng</td>
<td>3</td>
<td>zújǐ</td>
</tr>
</tbody>
</table>
Although the token frequency of the noun usage of céngjīng is not very high, it has high type frequency based on the corpus-based surveys of usage patterns. As for the noun usage of céngjīng, we mainly focus on situations when it is used as subject and object, that is, the types of its modifiers when céngjīng is used as a noun and the types of verbs which take céngjīng as their own objects. In CCL Corpus, there is only one case used as subject, which is modified by “měihǎo de”; there are 13 cases used as object, which are governed by 11 different verbs, such as “gǎnwù”, “xúnzhào”, “xiāngxìn”, etc. In DCC Corpus, there 16 cases that céngjīng is used as subject, in which céngjīng is modified by 13 different words, such as “měihǎo de”, “wǒmen de”, “búkān de” and “zùiài de”; Other 48 noun usages of céngjīng are used as object which are governed by 29 different verbs, such as “jìlù”, “jiēshòu”, “jìnglì”, and “shōucáng”, etc. In contrast, there are more noun usage patterns of céngjīng collected fromCNKI Title Search, 138 cases in total. 80 of them are used as subject, and the rest 58 used as object. As far as 80 cases used as subject are concerned, 42 appear in the form of proper nouns, which are used as the titles of poems, proses, plays, songs and movies; 27 are modified by 24 different words (“céngjīng de céngjīng” appears 4 times ), such as “wǒmen de” “měihǎo de”, “níngméngwèi de” and “niànniànbwàng de” etc; 11 of them don’t have modifiers, which function as subject directly, such as “céngjīng shiyizhòng měilì”, “céngjīng yíchéngwéi yǒngyuán”, etc. With regard to 58 cases used as object, they are governed by 41 types of verbs (zhuìyí, gàobié, huáiniàn and jìngli, etc.) and 4 prepositions (dùiyú, guānyu, wèile and xiàng).
3.3 Diachronic Development

Since there is no large-scale Chinese diachronic corpus available, we cannot make a corpus-based investigation on the time span of the adjective and noun usages of céngjīng. Considering that the data in CNKI have the information of publication time, we investigate the usage patterns of céngjīng from the beginning to the present by means of CNKI Title Search. As shown in Figure 1, based on the survey of usage patterns we find that the adjective usage of céngjīng began from 1994, and there are only two cases: “céngjīng de guòchéng” and “céngjīng de měili”; From then on, the adjective usage of céngjīng has increased gradually, and it has increased to 108 cases in 2010, the trend is raised except for a little fluctuations. Similarly, the result shows that the earliest appearance of the noun usage of céngjīng also began from 1994 with only one case: “búxiǎng nítóngyang shuō céngjīng”, in which céngjīng is used as the object of verb “shuō”. Except 1995 and 1998, the non usage of céngjīng in other years has appeared and increased gradually, and there are 17 cases in 2011 and 12 cases in 2016. Although there are some fluctuations in some individual years, the frequency of use has become stable and shows a rising tendency.

**Figure 1** The Diachronic Development Of Adjective and Noun Usages of Céngjīng in CNKI Title Search

4. Discussion

4.1 The Multicategorization of the Lexeme of Céngjīng

Based on the statistics of usage patterns from corpora, it is shown that the adjective usages of céngjīng can be found in CCL Corpus, DCC Corpus and CNKI Title Search except CN Corpus which was built earlier. Compared with its adverb usages, the token frequency of the adjective usages of céngjīng is not high. However, as time goes on and sample size increases, it is found that its frequency of use is getting higher and higher. Take such three lexemes as “rìzi”, “shídài”, “suìyuè” for example, they all appear just once in CCL Corpus which was built in early 2000, and 4, 6 and 20 times respectively in CNKI Title Search. Correspondingly, their frequency increases to 7, 16 and 32 times respectively in DCC Corpus with newer and richer data. In fact, the frequency of many
lexemes increases with the constant use. In addition, we find that the type frequency of the adjective usages of céngjīng is quite high in every corpus. In other words, the headwords that can be modified by céngjīng are various. More specifically, it can modify 247, 677 and 2482 different headwords respectively in CCL Corpus, CNKI Title Search and DCC Corpus. Usage-based Theory holds that linguistic knowledge is the knowledge of language use, language structure is emerged from its usage, and the frequency of use plays an important role in fixation and conventionalization of language structure (Bybee 2010; Wang and Chen 2014). Therefore, we conclude that the adjective usage of céngjīng with high token and type frequency as well as increasing use has highly conventionalized.

Compared with the adjective usage of céngjīng, the token frequency of its noun usage is lower. The proportions in CCL Corpus, DCC Corpus and CNKI Title Search are 0.04%, 0.16% and 2.57% respectively. Hanks (1990: 32) points out that language dictionaries demonstrate the conventionalized usages of the lexemes. Does it mean that the noun usage of céngjīng cannot be represented in dictionaries since it has not conventionalized yet? We believe that this is not the case, because token frequency is not the only criterion for judging the level of conventionalization (Wang and Chen 2014: 26). Type frequency is closely linked to the productivity of language structure (Bybee 2010: 95). As mentioned in Part 3.2, the noun usage of céngjīng, either as subject or object, bears higher type frequency in CCL Corpus, DCC Corpus and CNKI Title Search. To be specific, it can be modified by various words and serve as the object of many verbs and even several prepositions. From the perspective of teleology, we hold that the noun usage of céngjīng has conventionalized on the basis of its token frequency and higher type frequency as well as its increasing use.

In addition, the balance and standard ability of the chosen corpora are well guaranteed. Both CN Corpus and CCL Corpus are large-scale balanced Modern Chinese corpora with widespread materials, long time span, large capacity and good representativeness. The data in DCC Corpus mainly come from such mainstream newspapers media as People’s Daily, Guangming Daily, Southern Weekly and so on. CNKI is currently the most widely-used Chinese full text database, which covers full text data collected from domestic academic journals, doctors’ dissertations and masters’ theses, newspapers, conference papers, reference books, yearbooks and so forth. DCC Corpus and CNKI are the most representative, whose data are mainly from newspaper language. Newspaper language is the written language with characters as its carrier, and it is the advocate and pioneer of language standardization (Wu 2015: 235). As a mass media, newspapers deliver the latest information to the people and exert a leavening influence on language life of the public at the same time, that is, people will consciously or unconsciously imitate the newspaper language and use them. The data chosen for the survey of usage patterns have good representativeness, high standardization and widespread propagation, which further provide sufficient evidence for the conventionalization of the lexeme’s adjective and noun usages.
4.2 The Problems of the Previous Studies and Countermeasures

As mentioned earlier, this article aims to make further discussions based on Zhang (2003b). He thinks that the multicategorization of céngjīng is the manifestation of language complexity and the internal diversity of adverbs, and Chinese adverbs are multifunctional. Zhang (2003b) views the multifunctionality of céngjīng as the multifunctionality of Chinese adverbs instead of the multifunctionality of the lexeme of céngjīng, and it is overgeneralized to assume Chinese word classes are multifunctional rather than a small number of frequently used lexemes are multifunctional. Just as Hunston & Francis (2000: 192) put, it is fundamentally different to recognize the multifunctionality of a lexeme as the multifunctionality of a certain word class or the lexeme with more than one word classes, which involves “the very nature of word class itself”. According to the Two-level Word Class Categorization Theory, the word exists in two forms, that is, the word type in lexicon in communal language at langue and the word token in syntax at parole. With the repeated uses of word token, the class membership of the word type will change, and it is closely related to its frequency of use. The multifunctionality of word type is referred to the phenomenon that a polysemous lexeme bears two or more word classes in lexicon in communal language, which is represented as a heterosemous lexeme in language dictionaries (Wang 2013:12). According to the Lexicon of Common Words in Contemporary Chinese published by National Language Committee in 2008, céngjīng is a lexeme with very high frequency (the frequency rank is 859). Wang and Zhou (2015) find that there is a positive correlation between heterosemy and frequency in Modern Chinese, and lexemes with high frequency of use tend to be polysemous or heterosemous. Therefore, the multifunctionality of céngjīng only belongs to the lexeme itself which has a high frequency rank, not all Chinese adverbs.

Based on the corpus-based survey, it is found that the adjective and noun usages of céngjīng have highly conventionalized because of high token frequency and very high type frequency as well as the increasing frequency of use, which should be represented in large and medium-sized Modern Chinese dictionaries. However, céngjīng is labeled only as an adverb in two commonly used medium-sized Modern Chinese dictionaries, Contemporary Chinese Dictionary (from 5th to 7th editions) and The Standard Dictionary of Modern Chinese (from 1st to 3rd editions), both of which have a full-scale word class labeling. The reason why the multifunctionality of lexemes in Modern Chinese is not well solved is that the Principle of Parsimony is regarded as the golden rule in Modern Chinese grammar research so as to minimize the number of heterosemous lexemes. Wang and Zhou (2015: 67) hold that the correlation between heterosemy and frequency is resulted from competing motivations of economy and iconicity in communication, and the Principle of Parsimony is problematic in dealing with heterosemy in Chinese dictionaries. In spite of this, Contemporary Chinese Dictionary (7th edition) published in 2016 still abides by the Principle of Parsimony blindly to label the lexeme of céngjīng as an adverb without reflecting the reality of language use in the Chinese community. Therefore, we hold that the constraint of the Principle of Parsimony should be removed in Modern Chinese grammar research in
order that the number of heterosemous lexemes is not to be minimized. Céngjīng should be labeled as a heterosemous lexeme of adverb, adjective and noun in Modern Chinese dictionaries, not solely as an adverb. In Modern Chinese grammar studies, the mainstream research method of enumeration is also a major factor to prevent the progression of word class studies in Modern Chinese. The research conducted by means of enumeration cannot demonstrate the whole picture of language use and fails to abstract normal usages of a lexeme in lexicon in communal language at langue from the idiolects at parole. In the era of big data, we can make full use of corpora to conduct real-time and panoramic investigation on language usage patterns. Only on the basis of big data corpora and dictionaries can we make an accurate judgment for the word class membership of lexemes, find out their conventionalized usages, and differentiate word class membership of a word type in lexicon in communal language at langue and word class membership of a word token in syntax at parole.

What language dictionaries record is the conventionalized usages of lexemes (Hanks 1990: 32). They represent the language knowledge in communal language at langue, which is the product of cultural inheritance. Considering that token frequency contributes to the conventionalization of word tokens and type frequency contributes to the conventionalization of abstract schemas (Evans and Green 2006: 188; Wang and Chen 2014: 26), and type frequency is closely related to the productivity of language structures (Bybee 2010: 95), we hold that the adjective and noun usages of céngjīng with high token and type frequency have conventionalized. From the perspective of the Two-level Word Class Categorization Theory, the lexeme of céngjīng belongs to adverbs, adjectives and nouns in the lexicon of the communal language of Modern Chinese. Therefore, céngjīng should be labeled as a heterosemous lexeme in large and medium-sized Modern Chinese dictionaries, which can be treated as follows:

【曾经】

副 过去；从前；。形 过去（的）；从前（的）； ~的恋人 | ~的笑容。名 过去的经历或以前发生的事情；过去：回忆中的~ | 每一个讨厌的现在，都有一个不够努力的~。

5. Conclusion

In Modern Chinese grammar research, the multifunctionality of lexemes has been a controversial issue. Most scholars follow the Principle of Parsimony and assume the multifunctionality of a small number of frequently used lexemes as the characteristics of a certain word class, which exerts negative effect on the word class labeling in Modern Chinese dictionaries. Therefore, we should not blindly adhere to the Principle
of Parsimony and come to a wrong conclusion that does not conform to the fact of language use. Instead, the studies of Modern Chinese word classes should make quantitative analysis based on the big data incorpora and dictionaries to obtain accurate qualitative judgment, and generalize the norms in lexicon in communal language at langue from a mass of data in syntax at parole. Under the guidance of the Two-level Word Class Categorization Theory, the paper is aimed at exploring the conventionalization of the adjective and noun usages of cēngjīng in communal language at langue with the help of CCL Corpus, DCC Corpus and CNKI Title Search. It is found that the adjective and noun usages of cēngjīng have conventionalized owing to their relatively high token frequency and very high type frequency as well as the increasing uses, and the lexeme of cēngjīng belongs to adverbs, adjectives and nouns instead of adverbs only in the lexicon of the communal language of Modern Chinese, which should be labeled as a heterosemous lexeme in large and medium size Modern Chinese dictionaries.

References

Rauh, G. 2010. Syntactic categories: Their identification and description in linguistic
A Study of Multiple Class Membership in The Chinese-English Dictionary (3rd Ed.)

Na Li, Renqiang Wang, Chao Wang, Jiangfan Yu & Huifang Tang
Graduate School, Sichuan International Studies University
270258673@qq.com | wangrenqiang@sisu.edu.cn
867385795@qq.com | 547512493@qq.com | 514000942@qq.com

Abstract

Handling multifunctionality or multiple class membership is the most knotty and controversial issue in Modern Chinese grammar and in the word class labeling of Modern Chinese or Chinese-English dictionaries. Based on the Database of the Word Class Labeling Information in The Chinese-English Dictionary (3rd Ed.) and the Modern Chinese Corpus developed by the China National Language and Character Working Committee, an in-depth survey of the multiple class membership in the dictionary was conducted from the perspectives of lexicography, language typology and cognitive linguistics. It is found that achievements and problems coexist in the treatment of multiple class membership in The Chinese-English Dictionary (3rd Ed.) especially after comparing with the multiple class membership in The Contemporary Chinese Dictionary (5th Ed.) and Oxford Advanced Learner's English Dictionary (7th Ed.). It is argued that probably influenced by word class labeling of The Contemporary Chinese Dictionary (5th Ed.), The Chinese-English Dictionary (3rd Ed.) also minimized the number of multi-category lexemes by following the “Principle of Parsimony”, which however has led to some misrepresentation of the microstructure of some common lexemes, although it has slightly expanded the scope of multi-category lexemes. Thus the findings of this study can shed some insights on improving the word class labeling in Chinese-English lexicography.

Keywords: The Chinese-English Dictionary (3rd Ed.), multiple class membership, defining strategy, “Principle of Parsimony”

1. Introduction

The study of word classes has been considered as a hot and difficult issue in Modern Chinese Grammar, while multifunctionality or multiple class membership has always been the most controversial and knotty issue of linguistics for many years (Hu Mingyang 1996; Lieber 2005; Schmid 2011; van Lier and Rijkhoff 2013; Enfield 2015). Multiple class membership (hereinafter called MCM) refers to the phenomenon that a polysemous lexeme or word type belongs to more than one word class in the lexicon of a communal language, represented as a multi-category lexeme in a language dictionary.
(Wang Renqiang 2013: 12). Seen as a case that a slight move in one part may affect the situation as a whole, the accuracy of word class labeling has direct influence on the quality of Chinese/Chinese-English dictionaries (Wang Renqiang and Zhang Yihua 2006), while how to handle MCM is even more crucial.

The Chinese-English Dictionary, published by Foreign Language Teaching and Research Press, is a prestigious dictionary in the academic field. The Chinese-English Dictionary (3rd Ed.) (hereinafter called CED3), edited by Yao Xiaopin, has not only realized a full and comprehensive word class labeling, but also a thorough labeling on monosyllabic entries. Wang Renqiang (2012) comprehensively and systematically evaluates the word class labeling in CED3 in many respects, including the word class labeling system, the criterion and procedure of word class identification, the object of word class labeling, the “Principle of Translating by Word Classes” and so on. He claims that with regard to the criterion and procedure of Modern Chinese word class identification, some progress has been made in CED3, but there are also some inadequacies of the principle of handling MCM and of its interpretation on Guide to the Use of the Dictionary. Li Xiang (2013) contrasts the multi-category entries (except monosyllabic entries) whose initials are A, N and W in CED3 with their counterparts in The Contemporary Chinese Dictionary (5th Ed.) (hereinafter called CCD5) and Modern Chinese Standard Dictionary (2nd Ed.) (hereinafter called MCSD2), then brings forth a conclusion that the scope of MCM has been expanded in CED3 on the basis of the existing Chinese dictionaries. However, no full-scale survey of MCM in CED3 have been done.

Thus, based on the DIY Word Class Labeling Database of CED3, combined with the Modern Chinese Corpus developed by the China National Language and Character Working Committee (hereinafter called CN CORPUS), this paper will conduct a big-data empirical study on MCM in CED3 from the perspective of the Two-level Word Class Categorization Theory. It is intended to make an in-depth study of Modern Chinese word classes and word class labeling of modern Chinese-English dictionaries, then to propose defining strategies on solving the problem about the criterion and procedure of word class identification, which has been a dilemma in Chinese grammatical study, and consequently, to improve the quality of the Chinese-English dictionaries.

2. Research Design

2.1 The Two-level Word Class Categorization Theory

The Two-level Word Class Categorization Theory is proposed by Wang Renqiang in the 36th Annual Conference of the German Linguistic Society in 2014. The main view of this theoretical model is that word class categorization of analytic languages, like Modern Chinese and Modern English, happens at two levels: the word token categorization in syntax at parole and the word type categorization in the lexicon at communal level of langue. The former is the expression of the speaker’s propositional
speech act functions, including reference, predication, modification and so on, whereas the latter is the conventionalized propositional speech act function(s) of a word type resulted from self-organization or collective unconscious. The class membership of a word type does not have a priori existence, nor is it precategorial(underspecified), but is liable to change along with the frequent usage of word token in syntax at parole, then to be conventionalized as its meaning potential(s) in the lexicon of a communal language. In view of this, the class membership of a word type at langue is supposed to be discovered through corpus-based usage pattern surveys to find its conventionalized usage. The word class information of a lexeme in the dictionary belongs to the communal level of langue, therefore, only with corpus-based usage pattern surveys can the word class membership be identified precisely and well labeled in the dictionary.

2.2 Research Method

In order to achieve the objectives mentioned above, all the entries and their word class information in CED3 have been input into computer to establish the Word Class Labeling Database of CED3.

Every entry labeled with word class sis represented in the CED3 Database, including its specific type. As for those few entries without adequate labeling information or related to cross-references, we complement their labeling information on the ground of Guide to the Use of the Dictionary and the interpretation on the cross-references in the dictionary. In the meantime, we have consulted the labeling methods of Wang Renqiang (2011, 2013) used in Word Class Labeling Database of CCD5 so that some entries(including their senses) with no word class labeling are regarded as non-words. In other words, the very few monosyllabic and disyllabic entries(including their senses) are treated as monosyllabic morphemes or nonsense syllables, while polysyllabic entries with no word class labeling are regarded as phrases.

For the reason that CED3 does not have all the entries labeled word classes, the study consults the subdivision of CCD5 Database, dividing all the entries into 3 Groups: Group I includes entries provided with complete word class labeling information; Group II contains entries provided with partial word class labeling information; Group III consists entries with no word class labeling.

3. Results

The word class labeling information of Chinese entries is provided by CED3 on the distinction between words and non-words. From Table 1, we can see clearly that there are 89041 entries in total, among which Group I takes a percentage of 90.11% with

---

1For ease of statistics, the monosyllabic morphemes or nonsense syllables in CED3 Database include meaningful and nonsense one single syllables respectively which cannot be used independently, while phrases refer to polysyllabic entries, like word group, idiom and so on. In view of the fact that this study is mainly concerned with words rather than monosyllabic morphemes, nonsense syllables and phrases, therefore this treatment can not affect the result essentially.
80231 entries, Group II includes 47 entries which account for 0.05% and Group III includes 8763 entries, taking up 9.84%.

**Table 1** Objects of Word Class Labeling in CED3

<table>
<thead>
<tr>
<th>Types</th>
<th>Group I</th>
<th>Group II</th>
<th>Group III</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>80231</td>
<td>47</td>
<td>8763</td>
<td>89041</td>
</tr>
<tr>
<td>Percentage</td>
<td>90.11%</td>
<td>0.05%</td>
<td>9.84%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

In view of the fact that the number and percentage of entries in Group II are relatively small and the entries in Group III belong to non-word grammatical units which have nothing to do with MCM and due to limited space, this study concentrates on Group I to conduct an in-depth survey of the MCM in CED3.

### 3.1 The Overview of the MCM

As Table 2 shows, single-category entries takes the majority of the total with 94.15%(83832 in number), whereas the number of multi-category entries is only 5209, taking up 5.85%.

**Table 2** Statistics of Word Class Labeling Information in CED3

<table>
<thead>
<tr>
<th>Types</th>
<th>Single-category entries</th>
<th>Multi-category entries</th>
<th>Sub-total</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>With two word classes</td>
<td>With three word classes</td>
<td>With four word classes</td>
</tr>
<tr>
<td>Number</td>
<td>83832</td>
<td>4695</td>
<td>442</td>
<td>62</td>
</tr>
<tr>
<td>Percentage</td>
<td>94.15%</td>
<td>5.27%</td>
<td>0.50%</td>
<td>0.07%</td>
</tr>
</tbody>
</table>

The number of entries that belong to more than one word class in the whole dictionary is 5209, among which entries belong to two word classes take the majority of the total with a percentage of 90.13% and entries with three word classes rank in the second place, which account for 8.49%. While entries with four, five and six word classes only take a small percentage of 1.38% in total. For details, see Table 3 given below:
### Table 3 Statistics of MCM in CED3

<table>
<thead>
<tr>
<th>Types</th>
<th>Two word classes</th>
<th>Three word classes</th>
<th>Four word classes</th>
<th>Five word classes</th>
<th>Six word classes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>4695</td>
<td>442</td>
<td>62</td>
<td>8</td>
<td>2</td>
<td>5209</td>
</tr>
<tr>
<td>Percentage</td>
<td>90.13%</td>
<td>8.49%</td>
<td>1.19%</td>
<td>0.15%</td>
<td>0.04%</td>
<td>100%</td>
</tr>
</tbody>
</table>

#### 3.2 Statistics and Analysis of the MCM of Group I

As Table 4 shows, Group I has 80231 entries in CED3, among which there are 75035 single-category entries, accounting for 93.52% of the total and 5196 multi-category entries with a percentage of 6.48%, which is relatively similar to the percentage of the whole dictionary (5.85%). Therefore, the study of entries in Group I in the case of MCM can largely represent the whole dictionary.

### Table 4 Statistics of Word Class Labeling Information of Group I in CED3

<table>
<thead>
<tr>
<th>Types</th>
<th>Single-category entries</th>
<th>Multi-category entries</th>
<th>With two word classes</th>
<th>With three word classes</th>
<th>With four word classes</th>
<th>With five word classes</th>
<th>With six word classes</th>
<th>Sub-total</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>75035</td>
<td>80231</td>
<td>4691</td>
<td>434</td>
<td>61</td>
<td>8</td>
<td>2</td>
<td>5196</td>
<td>100%</td>
</tr>
<tr>
<td>Percentage</td>
<td>93.52%</td>
<td>6.48%</td>
<td>5.85%</td>
<td>0.54%</td>
<td>0.08%</td>
<td>0.01%</td>
<td>0.00%</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

Among the single-category entries of Group I, the percentages of nouns, verbs and adjectives are 58.07%, 30.43% and 8.77% respectively, which together add up to 97.27%. For details, see Table 5 given below:

### Table 5 Internal Structure of Single-category Entries of Group I in CED3

<table>
<thead>
<tr>
<th>Word class</th>
<th>N</th>
<th>V</th>
<th>ADJ</th>
<th>AD</th>
<th>PRO</th>
<th>NUM</th>
<th>QUANT</th>
<th>PRE</th>
<th>CO</th>
<th>AUX</th>
<th>INTR</th>
<th>ON</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>4357</td>
<td>228</td>
<td>6579</td>
<td>964</td>
<td>156</td>
<td>266</td>
<td>38</td>
<td>74</td>
<td>78</td>
<td>255</td>
<td>750</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>Percentage</td>
<td>58.0%</td>
<td>30.4%</td>
<td>8.77%</td>
<td>1.28%</td>
<td>0.21%</td>
<td>0.07%</td>
<td>0.35%</td>
<td>0.05%</td>
<td>0.23%</td>
<td>0.10%</td>
<td>0.34%</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

Among the multi-category entries of Group I, entries with two word classes take the majority of the total with a percentage of 90.28% and entries with three word classes rank the second place with a percentage of 8.35%. While entries with four, five and six
word classes only take a small percentage of 1.36% in total. For details, see Table 6 given below:

**Table 6 Statistics of MCM of Group I in CED3**

<table>
<thead>
<tr>
<th>Types</th>
<th>Two word classes</th>
<th>Three word classes</th>
<th>Four word classes</th>
<th>Five word classes</th>
<th>Six word classes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>4691</td>
<td>434</td>
<td>61</td>
<td>8</td>
<td>2</td>
<td>5196</td>
</tr>
<tr>
<td>Percentage</td>
<td>90.28%</td>
<td>8.35%</td>
<td>1.17%</td>
<td>0.15%</td>
<td>0.04%</td>
<td>100%</td>
</tr>
</tbody>
</table>

There are a total of 42 types of MCM among the 4691 entries with two word classes of Group I, among which N-V lexemes rank the first place with a percentage of 53.12%, followed by N-ADJ lexemes (16.22%) and V-ADJ lexemes (15.26%). The number of the three types is 3969, taking a percentage of 84.61%. For details, see Table 7 given below:

**Table 7 Entries with TwoWord Classes of Group I in CED3**

<table>
<thead>
<tr>
<th>No.</th>
<th>Types</th>
<th>Quantity</th>
<th>Percentage</th>
<th>No.</th>
<th>Types</th>
<th>Quantity</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N-V</td>
<td>2492</td>
<td>53.12%</td>
<td>22</td>
<td>ADJ-QUANT</td>
<td>5</td>
<td>0.11%</td>
</tr>
<tr>
<td>2</td>
<td>N-ADJ</td>
<td>761</td>
<td>16.22%</td>
<td>23</td>
<td>N-PREP</td>
<td>4</td>
<td>0.09%</td>
</tr>
<tr>
<td>3</td>
<td>V-ADJ</td>
<td>716</td>
<td>15.26%</td>
<td>24</td>
<td>V-AUX</td>
<td>4</td>
<td>0.09%</td>
</tr>
<tr>
<td>4</td>
<td>ADJ-ADV</td>
<td>127</td>
<td>2.71%</td>
<td>25</td>
<td>ADJ-CONJ</td>
<td>4</td>
<td>0.09%</td>
</tr>
<tr>
<td>5</td>
<td>N-QUANT</td>
<td>124</td>
<td>2.64%</td>
<td>26</td>
<td>PRON-CONJ</td>
<td>4</td>
<td>0.09%</td>
</tr>
<tr>
<td>6</td>
<td>V-ADV</td>
<td>123</td>
<td>2.62%</td>
<td>27</td>
<td>PREP-CONJ</td>
<td>4</td>
<td>0.09%</td>
</tr>
<tr>
<td>7</td>
<td>N-ADV</td>
<td>110</td>
<td>2.34%</td>
<td>28</td>
<td>N-INTERJ</td>
<td>3</td>
<td>0.06%</td>
</tr>
<tr>
<td>8</td>
<td>V-QUANT</td>
<td>33</td>
<td>0.70%</td>
<td>29</td>
<td>ADV-NUM</td>
<td>3</td>
<td>0.06%</td>
</tr>
<tr>
<td>9</td>
<td>V-PREP</td>
<td>27</td>
<td>0.58%</td>
<td>30</td>
<td>ADV-AUX</td>
<td>3</td>
<td>0.06%</td>
</tr>
<tr>
<td>10</td>
<td>ADV-CONJ</td>
<td>26</td>
<td>0.55%</td>
<td>31</td>
<td>AUX-INTERJ</td>
<td>3</td>
<td>0.06%</td>
</tr>
<tr>
<td>11</td>
<td>V-CONJ</td>
<td>17</td>
<td>0.36%</td>
<td>32</td>
<td>ADV-QUANT</td>
<td>2</td>
<td>0.04%</td>
</tr>
<tr>
<td>12</td>
<td>N-PRON</td>
<td>12</td>
<td>0.26%</td>
<td>33</td>
<td>N-AUX</td>
<td>2</td>
<td>0.04%</td>
</tr>
<tr>
<td>13</td>
<td>ADJ-ONOM</td>
<td>12</td>
<td>0.26%</td>
<td>34</td>
<td>V-NUM</td>
<td>1</td>
<td>0.02%</td>
</tr>
<tr>
<td>14</td>
<td>ADV-PRON</td>
<td>11</td>
<td>0.23%</td>
<td>35</td>
<td>V-PRON</td>
<td>1</td>
<td>0.02%</td>
</tr>
<tr>
<td>15</td>
<td>V-INTERJ</td>
<td>10</td>
<td>0.21%</td>
<td>36</td>
<td>ADJ-INTERJ</td>
<td>1</td>
<td>0.02%</td>
</tr>
<tr>
<td>16</td>
<td>V-ONOM</td>
<td>9</td>
<td>0.19%</td>
<td>37</td>
<td>ADJ-PRON</td>
<td>1</td>
<td>0.02%</td>
</tr>
<tr>
<td>17</td>
<td>ADJ-NUM</td>
<td>8</td>
<td>0.17%</td>
<td>38</td>
<td>PRON-AUX</td>
<td>1</td>
<td>0.02%</td>
</tr>
<tr>
<td>18</td>
<td>N-NUM</td>
<td>6</td>
<td>0.13%</td>
<td>39</td>
<td>PRON-INTERJ</td>
<td>1</td>
<td>0.02%</td>
</tr>
<tr>
<td>19</td>
<td>N-ONOM</td>
<td>6</td>
<td>0.13%</td>
<td>40</td>
<td>PRON-QUANT</td>
<td>1</td>
<td>0.02%</td>
</tr>
<tr>
<td>20</td>
<td>ONOM-INTERJ</td>
<td>6</td>
<td>0.13%</td>
<td>41</td>
<td>PREP-ADV</td>
<td>1</td>
<td>0.02%</td>
</tr>
<tr>
<td>21</td>
<td>N-CONJ</td>
<td>5</td>
<td>0.11%</td>
<td>42</td>
<td>ONOM-QUANT</td>
<td>1</td>
<td>0.02%</td>
</tr>
</tbody>
</table>
As shown in Table 8, 38 types of MCM exist in the 434 entries with three word classes of Group I. Among them, the percentages of N-V-ADJ lexemes, N-V-QUANT lexemes, V-ADJ-ADV lexemes, N-ADJ-ADV lexemes, N-ADV lexemes and N-V-PREP lexemes are 39.40%, 16.36%, 10.14%, 9.45%, 7.60% and 3.23% respectively. The total number of these six types is 374, accounting for 86.18% of entries with three word classes of Group I.

### Table 8 Entries with Three Word Classes of Group I in CED3

<table>
<thead>
<tr>
<th>No.</th>
<th>Types</th>
<th>Quantity</th>
<th>Percentage</th>
<th>No.</th>
<th>Types</th>
<th>Quantity</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N-V-ADJ</td>
<td>171</td>
<td>39.40%</td>
<td>20</td>
<td>V-ADJ-PREP</td>
<td>4</td>
<td>0.92%</td>
</tr>
<tr>
<td>2</td>
<td>N-V-QUANT</td>
<td>71</td>
<td>16.36%</td>
<td>21</td>
<td>V-ADJ-CONJ</td>
<td>1</td>
<td>0.23%</td>
</tr>
<tr>
<td>3</td>
<td>V-ADJ-ADV</td>
<td>44</td>
<td>10.14%</td>
<td>22</td>
<td>V-ADJ-AUX</td>
<td>1</td>
<td>0.23%</td>
</tr>
<tr>
<td>4</td>
<td>N-ADJ-ADV</td>
<td>41</td>
<td>9.45%</td>
<td>23</td>
<td>V-ADJ-INTERJ</td>
<td>1</td>
<td>0.23%</td>
</tr>
<tr>
<td>5</td>
<td>N-V-ADV</td>
<td>33</td>
<td>7.60%</td>
<td>24</td>
<td>V-ADJ-ONOM</td>
<td>2</td>
<td>0.46%</td>
</tr>
<tr>
<td>6</td>
<td>N-V-PREP</td>
<td>14</td>
<td>3.23%</td>
<td>25</td>
<td>V-INTERJ-ONOM</td>
<td>1</td>
<td>0.23%</td>
</tr>
<tr>
<td>7</td>
<td>N-V-CONJ</td>
<td>1</td>
<td>0.23%</td>
<td>26</td>
<td>V-ADV-PRON</td>
<td>3</td>
<td>0.69%</td>
</tr>
<tr>
<td>8</td>
<td>N-ADJ-PRON</td>
<td>1</td>
<td>0.23%</td>
<td>27</td>
<td>V-ADV-NUM</td>
<td>2</td>
<td>0.46%</td>
</tr>
<tr>
<td>9</td>
<td>N-ADJ-NUM</td>
<td>2</td>
<td>0.46%</td>
<td>28</td>
<td>V-ADV-PREP</td>
<td>7</td>
<td>1.61%</td>
</tr>
<tr>
<td>10</td>
<td>N-ADJ-QUANT</td>
<td>4</td>
<td>0.92%</td>
<td>29</td>
<td>V-ADV-CONJ</td>
<td>5</td>
<td>1.15%</td>
</tr>
<tr>
<td>11</td>
<td>N-ADV-PRON</td>
<td>1</td>
<td>0.23%</td>
<td>30</td>
<td>V-PRON-CONJ</td>
<td>1</td>
<td>0.23%</td>
</tr>
<tr>
<td>12</td>
<td>N-ADV-NUM</td>
<td>1</td>
<td>0.23%</td>
<td>31</td>
<td>V-PREP-CONJ</td>
<td>1</td>
<td>0.23%</td>
</tr>
<tr>
<td>13</td>
<td>N-ADV-QUANT</td>
<td>4</td>
<td>0.92%</td>
<td>32</td>
<td>V-PREP-AUX</td>
<td>1</td>
<td>0.23%</td>
</tr>
<tr>
<td>14</td>
<td>N-ADV-CONJ</td>
<td>1</td>
<td>0.23%</td>
<td>33</td>
<td>V-CONJ-AUX</td>
<td>1</td>
<td>0.23%</td>
</tr>
<tr>
<td>15</td>
<td>N-PRON-CONJ</td>
<td>1</td>
<td>0.23%</td>
<td>34</td>
<td>ADJ-ADV-PRON</td>
<td>1</td>
<td>0.23%</td>
</tr>
<tr>
<td>16</td>
<td>N-V-ONOM</td>
<td>2</td>
<td>0.46%</td>
<td>35</td>
<td>ADJ-ADV-CONJ</td>
<td>1</td>
<td>0.23%</td>
</tr>
<tr>
<td>17</td>
<td>V-ADJ-PRON</td>
<td>1</td>
<td>0.23%</td>
<td>36</td>
<td>ADV-PRON-CONJ</td>
<td>2</td>
<td>0.46%</td>
</tr>
<tr>
<td>18</td>
<td>V-ADJ-NUM</td>
<td>1</td>
<td>0.23%</td>
<td>37</td>
<td>ADV-PRON-AUX</td>
<td>1</td>
<td>0.23%</td>
</tr>
<tr>
<td>19</td>
<td>V-ADJ-QUANT</td>
<td>3</td>
<td>0.69%</td>
<td>38</td>
<td>ADV-CONJ-AUX</td>
<td>1</td>
<td>0.23%</td>
</tr>
</tbody>
</table>

As Table 9 shows, there are 16 types of MCM in the entries with four word classes of Group I, among which N-V-ADJ-ADV lexemes rank the first place with a percentage of 42.62%, followed by N-V-ADJ-QUANT lexemes with a percentage of 24.59%.
Table 9 Entries with Four Word Classes of Group I in CED3

<table>
<thead>
<tr>
<th>No.</th>
<th>Types</th>
<th>Quantity</th>
<th>Percentage</th>
<th>No.</th>
<th>Types</th>
<th>Quantity</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N-V-ADJ-ADV</td>
<td>26</td>
<td>42.62%</td>
<td>9</td>
<td>N-V-PRON-QUANT</td>
<td>1</td>
<td>1.64%</td>
</tr>
<tr>
<td>2</td>
<td>N-V-ADJ-QUANT</td>
<td>15</td>
<td>24.59%</td>
<td>10</td>
<td>N-V-ADV-PRON</td>
<td>1</td>
<td>1.64%</td>
</tr>
<tr>
<td>3</td>
<td>N-V-PREP-CONJ</td>
<td>4</td>
<td>6.56%</td>
<td>11</td>
<td>N-V-ADJ-CONJ</td>
<td>1</td>
<td>1.64%</td>
</tr>
<tr>
<td>4</td>
<td>N-V-ADV-PREP</td>
<td>3</td>
<td>4.92%</td>
<td>12</td>
<td>N-V-ADJ-PREP</td>
<td>1</td>
<td>1.64%</td>
</tr>
<tr>
<td>5</td>
<td>N-V-ADJ-AUX</td>
<td>2</td>
<td>3.28%</td>
<td>13</td>
<td>NUM-ADV-PREP-AUX</td>
<td>1</td>
<td>1.64%</td>
</tr>
<tr>
<td>6</td>
<td>V-ADJ-ADV-AUX</td>
<td>1</td>
<td>1.64%</td>
<td>14</td>
<td>N-ADJ-QUANT-PREP</td>
<td>1</td>
<td>1.64%</td>
</tr>
<tr>
<td>7</td>
<td>V-ADV-PREP-AUX</td>
<td>1</td>
<td>1.64%</td>
<td>15</td>
<td>N-ADJ-NUM-QUANT</td>
<td>1</td>
<td>1.64%</td>
</tr>
<tr>
<td>8</td>
<td>N-ADJ-ADV-QUANT</td>
<td>1</td>
<td>1.64%</td>
<td>16</td>
<td>N-ADJ-ADV-PREP</td>
<td>1</td>
<td>1.64%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>61</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In Group I, there are only 8 entries with five word classes, which are approximately divided into 6 types. For details, see Table 7 given below:

Table 10 Entries with Five Word Classes of Group I in CED3

<table>
<thead>
<tr>
<th>No.</th>
<th>Types</th>
<th>Quantity</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N-V-ADJ-PREP-ONOM</td>
<td>1</td>
<td>12.50%</td>
</tr>
<tr>
<td>2</td>
<td>N-V-ADJ-PREP-AUX</td>
<td>1</td>
<td>12.50%</td>
</tr>
<tr>
<td>3</td>
<td>N-V-ADV-QUANT-PREP</td>
<td>1</td>
<td>12.50%</td>
</tr>
<tr>
<td>4</td>
<td>N-V-ADJ-ADV-CONJ</td>
<td>1</td>
<td>12.50%</td>
</tr>
<tr>
<td>5</td>
<td>N-V-ADJ-ADV-QUANT</td>
<td>2</td>
<td>25.00%</td>
</tr>
<tr>
<td>6</td>
<td>V-ADJ-ADV-PREP-CONJ</td>
<td>2</td>
<td>25.00%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>8</td>
<td>100%</td>
</tr>
</tbody>
</table>

Only two entries in Group I are labeled with six word classes, which are 本 1 and 对:

Through the comparison of word length between single-category entries and multi-category entries in Group I, it is found that the percentage of multi-category entries with one syllable ranks the highest (22.75%), followed by disyllabic entries (5.23%), three-syllables entries (1.44%), and four-syllables entries (1.06%). From Table 11, it can be seen that the longer the lexemes are, the higher the percentage of single-category entries and the lower the percentage of multi-category entries.
Table 11 Relationship between Word Length and MCM in Group I in CED3

<table>
<thead>
<tr>
<th>Word length</th>
<th>Single syllable</th>
<th>Double syllables</th>
<th>Three syllables</th>
<th>Four syllables</th>
<th>Over four syllables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Types</td>
<td>Number</td>
<td>Percentage</td>
<td>Number</td>
<td>Percentage</td>
</tr>
<tr>
<td></td>
<td>Single-category entries</td>
<td>6721</td>
<td>77.25%</td>
<td>55080</td>
<td>94.77%</td>
</tr>
<tr>
<td></td>
<td>Multi-category entries</td>
<td>1979</td>
<td>22.75%</td>
<td>3038</td>
<td>5.23%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>8700</td>
<td>100.00%</td>
<td>58118</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

4. Discussion

The statistics above show that CED3 contains 89041 entries, among which there are only 5209 multi-category entries, accounting for 5.85% of the total. Within the 80231 entries of Group I, there are 5196 multi-category entries, taking a percentage of 6.48%. Entries with two word classes take the majority of multi-category ones with a percentage of 90.28%, among which N-V, N-ADJ and V-ADJ lexemes are the three main types accounting for 84.61% in total.

Compared with the statistics of MCM in CCD5 analyzed by Wang Renqiang (2013), it is found that the number of entries of Group I in CED3 (80231) is greater than that in CCD5 (51469). However, the percentage of MCM of Group I in CED3 (6.48%) is similar to and slightly higher than that in CCD5 (5.40%). The statistics above show that the absolute quantity of multi-category lexemes in CED3 has been increased. In the Preface of CED3, Yao Xiaoping (2010) points out that CED3 has consulted CCD5 when labeling word classes, and appropriately made some improvements if necessary. CED3 is a Chinese-English dictionary taking CCD5 as the blueprint in its word class labeling, thus it is understandable that the percentages of multi-category entries in these two dictionaries are similar. In the meantime, the percentage of multi-category entries in CED3 is slightly higher than that in CCD5, which demonstrates that CED3 has expanded the scope of multi-category entries on the basis of consulting Chinese dictionaries. As shown in Table 12, the following entries are marked as single-category lexemes in CED1 and CED2, but in CED3, they are labeled as multi-category lexemes.
Table 12 Examples of Entries with Additional Word Classes in CED3

<table>
<thead>
<tr>
<th>No</th>
<th>Entries</th>
<th>CED1/CED2</th>
<th>CED3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>省事</td>
<td>V</td>
<td>V-ADJ</td>
</tr>
<tr>
<td>2</td>
<td>始终</td>
<td>ADV</td>
<td>N-ADV</td>
</tr>
<tr>
<td>3</td>
<td>调剂</td>
<td>V</td>
<td>V-ADJ</td>
</tr>
<tr>
<td>4</td>
<td>形象</td>
<td>N</td>
<td>N-ADJ</td>
</tr>
<tr>
<td>5</td>
<td>虚拟</td>
<td>ADJ</td>
<td>V-ADJ</td>
</tr>
<tr>
<td>6</td>
<td>严密</td>
<td>ADJ</td>
<td>V-ADJ</td>
</tr>
<tr>
<td>7</td>
<td>一旦</td>
<td>ADV</td>
<td>N-ADV</td>
</tr>
<tr>
<td>8</td>
<td>隐蔽</td>
<td>V</td>
<td>V-ADJ</td>
</tr>
<tr>
<td>9</td>
<td>拥挤</td>
<td>V</td>
<td>V-ADJ</td>
</tr>
</tbody>
</table>

Compared with the MCM in OALD7 (Wang Renqiang 2014), it is found that there is a great difference between the percentages of multi-category lexemes in CED3 (6.48%) and in OALD7 (10.48%). From the perspective of language typology, Chinese is the typical isolating or analytic language lacking inflections and morphological change. In the 1400 years’ evolution, English has typologically shifted from a predominantly synthetic language to a predominantly analytic language although its analyticity is weaker than Chinese (Haselow 2011; Payne 2011). Crystal (2010: 303) also states that with few inflectional endings like Chinese, English is more like an isolating or analytic language from a typological perspective. In Robins’ opinion, “languages with less morphological word form variation more readily admit MCM” (1989: 214). Wang Renqiang (2014b: 55) also points out that MCM is a common feature shared by analytic or isolating languages. As for the typicality of analytic languages, Modern English is weaker than Modern Chinese, thus theoretically it can be deduced that the percentage of multi-category lexemes in Modern Chinese should be moderately higher than that in Modern English. From the comparison between CED3 and OALD7, it is found that the percentage of multi-category lexemes in CED3 is much lower than that in OALD7. The result of comparison is different from the theoretical assumption and practical use, which may illustrate some problems of CED3 in its handling of MCM.

In dealing with MCM, many Chinese scholars claim to minimize the number of multi-category lexemes so as to stick to the so called “Principle of Parsimony”. This viewpoint can be traced back to the Modern Chinese Grammar (《中国现代语法》)
published by Wang Li in 1943, An Introduction to Grammar and Rhetoric (《语法修辞讲话》) published by Lü Shuxiang and Zhu Dexi in 1951 and “some principle issues on Chinese word classes” (《关于汉语词类的一些原则性问题》) published by Lü Shuxiang in 1954. Lü Shuxiang and Zhu Dexi (1951: 12) put forward the principle of “minimizing the number of multi-category words (lexemes)” to solve the problem of “words with fixed classes”. In order to pursue the simplicity of grammatical analysis, reduce and even avoid MCM, Zhu Dexi (1982, 1985) believes that one important feature of Chinese is no one-to-one correspondence between word classes and syntax. He claims that serving as subject and object is one function of Chinese verbs and adjectives and that Chinese adjectives can be used as adverbial. Later, reducing MCM and advocating the simplicity of grammatical analysis are summarized by Shen Jiaxuan (2009) as “Principle of Parsimony”. However, “Principle of Parsimony” has problems in handling multi-category lexemes. Wang Renqiang and Zhou Yu (2015) point out that “Principle of Parsimony” violates the basic definition of multi-category lexemes and inappropriately uses the Law of Contradiction. The Law of Contradiction in first-order logic requires that two judgments made simultaneously on one object cannot be contradictory. So supporters of the “Principle of Parsimony” state that one lexeme cannot possess multiple classes. Superficially, MCM seems contrary to the Law of Contradiction, but actually it is not. MCM is not the concept in syntax at parole, but the meaning potential(s) of word types in lexicon at langue. The Law of Contradiction can only explain word token categorization in syntax at parole; it cannot deal with MCM of word type at langue. The proposition of the “Principle of Parsimony” is deeply influenced by the lexicon view of structural linguists (including generative linguists), who advocate the simplification of grammatical description. When they are confronted with multifunctionality (MCM), they decrease the information in lexicon and reject to treat as multi-category lexemes in lexicon. This lexicon view is based on a priori view of language, which means that linguistic knowledge exists in people’s mind a priori (congenitally) and has nothing to do with language use. Yet, according to the Usage-Based Theory, language is a complex adaptive system; linguistic structures are not invariable, but emerge from usage and vary through usage (Beckner et al.2009; Kretzschmar Jr.2015).

Studies show that CCD5 has been influenced by the “Principle of Parsimony” and the quantity of multi-category lexemes in CCD5 has been artificially minimized (Wang Renqiang 2013). Compared with CCD5, although the labeling scope of multi-category lexemes in CED3 has been expanded, great differences still remain when compared with the percentage of multi-category lexemes in Modern English. Thus, it can be deduced that CED3 may have still been influenced by the “Principle of Parsimony” and has artificially minimized the number of multi-category lexemes. Therefore, a sample survey will be conducted to examine if CED3 has the above problems. Wang Renqiang (2005) points out that symmetry is both the operational principle of word
class labeling in Chinese-English dictionaries and the method to find the problems of word class labeling and translation in Chinese-English dictionaries. The symmetry of word class and translation in Chinese-English dictionaries is an aesthetic principle for the harmony and balance of translation and a scientific requirement for the accuracy and objectivity of translation. Therefore, the “Principle of Symmetry” for language can be seen as a scientific principle to test the appropriateness of word class labeling for multi-category lexemes in Chinese-English dictionaries, and specifically to test whether the symmetric lexemes have the same word classes so as to find the word class labeling problems in Chinese-English dictionaries. Through randomly drawing synonymous, antonymous and homologous lexemes, it is found that obvious problems exist in CED3. Table 13 shows the results.

Table 13 Contrast of Word Class Labeling between Symmetric Lexemes

<table>
<thead>
<tr>
<th>No.</th>
<th>Multi-category lexemes</th>
<th>Word classes</th>
<th>Single-category lexemes</th>
<th>Word class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>安全</td>
<td>N-ADJ</td>
<td>危险</td>
<td>V</td>
</tr>
<tr>
<td>2</td>
<td>伴唱</td>
<td>N-V</td>
<td>清唱</td>
<td>V</td>
</tr>
<tr>
<td>3</td>
<td>比赛</td>
<td>N-V</td>
<td>竞赛</td>
<td>V</td>
</tr>
<tr>
<td>4</td>
<td>成功</td>
<td>V-ADJ</td>
<td>失败</td>
<td>V</td>
</tr>
<tr>
<td>5</td>
<td>成熟</td>
<td>V-ADJ</td>
<td>幼稚</td>
<td>ADJ</td>
</tr>
<tr>
<td>6</td>
<td>惩罚</td>
<td>N-V</td>
<td>奖励</td>
<td>V</td>
</tr>
<tr>
<td>7</td>
<td>胡说</td>
<td>N-V</td>
<td>瞎说</td>
<td>V</td>
</tr>
<tr>
<td>8</td>
<td>回答</td>
<td>N-V</td>
<td>回应</td>
<td>V</td>
</tr>
<tr>
<td>9</td>
<td>急性子</td>
<td>N-ADJ</td>
<td>慢性子</td>
<td>N</td>
</tr>
<tr>
<td>10</td>
<td>兼职</td>
<td>N-V</td>
<td>专职</td>
<td>N</td>
</tr>
<tr>
<td>11</td>
<td>开心</td>
<td>V-ADJ</td>
<td>伤心</td>
<td>ADJ</td>
</tr>
<tr>
<td>12</td>
<td>口臭</td>
<td>N-V</td>
<td>体臭</td>
<td>N</td>
</tr>
<tr>
<td>13</td>
<td>期望</td>
<td>N-V</td>
<td>期盼、期许、期冀</td>
<td>V</td>
</tr>
<tr>
<td></td>
<td>Meaning</td>
<td>Category</td>
<td>Meaning</td>
<td>Category</td>
</tr>
<tr>
<td>---</td>
<td>---------</td>
<td>----------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>14</td>
<td>齐唱</td>
<td>N-V</td>
<td>齐奏</td>
<td>V</td>
</tr>
<tr>
<td>15</td>
<td>启示</td>
<td>N-V</td>
<td>启发</td>
<td>V</td>
</tr>
<tr>
<td>16</td>
<td>入神</td>
<td>V-ADJ</td>
<td>入迷</td>
<td>V</td>
</tr>
<tr>
<td>17</td>
<td>伤神</td>
<td>V-ADJ</td>
<td>伤心、伤感</td>
<td>ADJ</td>
</tr>
<tr>
<td>18</td>
<td>烧伤</td>
<td>N-V</td>
<td>灼伤</td>
<td>V</td>
</tr>
<tr>
<td>19</td>
<td>失音</td>
<td>N-V</td>
<td>失语</td>
<td>V</td>
</tr>
<tr>
<td>20</td>
<td>私立</td>
<td>V-ADJ</td>
<td>公立</td>
<td>ADJ</td>
</tr>
<tr>
<td>21</td>
<td>塌方</td>
<td>N-V</td>
<td>雪崩</td>
<td>V</td>
</tr>
<tr>
<td>22</td>
<td>跳高</td>
<td>N-V</td>
<td>跳远</td>
<td>V</td>
</tr>
<tr>
<td>23</td>
<td>退潮</td>
<td>N-V</td>
<td>涨潮</td>
<td>V</td>
</tr>
<tr>
<td>24</td>
<td>五彩</td>
<td>N-ADJ</td>
<td>七彩</td>
<td>N</td>
</tr>
<tr>
<td>25</td>
<td>现代</td>
<td>N-ADJ</td>
<td>古代、近代</td>
<td>N</td>
</tr>
<tr>
<td>26</td>
<td>幸福</td>
<td>N-ADJ</td>
<td>痛苦</td>
<td>N</td>
</tr>
<tr>
<td>27</td>
<td>有益</td>
<td>V-ADJ</td>
<td>有利</td>
<td>ADJ</td>
</tr>
<tr>
<td>28</td>
<td>有益</td>
<td>V-ADJ</td>
<td>无益</td>
<td>V</td>
</tr>
<tr>
<td>29</td>
<td>预料</td>
<td>N-V</td>
<td>预测</td>
<td>V</td>
</tr>
<tr>
<td>30</td>
<td>照旧</td>
<td>V-ADV</td>
<td>如常</td>
<td>V</td>
</tr>
<tr>
<td>31</td>
<td>争论</td>
<td>N-V</td>
<td>争辩、争吵</td>
<td>V</td>
</tr>
<tr>
<td>32</td>
<td>中风</td>
<td>N-V</td>
<td>中毒</td>
<td>V</td>
</tr>
<tr>
<td>33</td>
<td>祝愿</td>
<td>N-V</td>
<td>祝福</td>
<td>V</td>
</tr>
</tbody>
</table>

In addition to the phenomena above, another problem has also been found, which refers to the great difference in the word class labeling of symmetric single-category
lexemes. For instance, 全部 is labeled as adjective, while 全体 is labeled as noun; 阴性 and 阳性 are labeled as nouns, while 中性 is labeled as adjective; 远程 is labeled as adjective, while 近程 is labeled as noun; 怯生 is labeled as adjective, while 认生 is labeled as verb; 失常 is labeled as adjective, while 失控 is labeled as verb; but the examples in the two lexemes have the same structure: 精神失常 & 精神失控. 私有 is labeled as verb, but 公有 is labeled as adjective; 无味 is labeled as verb, but 无色 is labeled as adjective and 乏味 (also 无味) is also labeled as adjective; 单程 is labeled as noun, but 双程 is labeled as adjective; 分心 is labeled as verb, but 专心 is labeled as adjective; 早育 is labeled as noun, but 晚育 is labeled as verb; 顾虑 is labeled as verb, but 顾虑 is labeled as noun; 海运 is labeled as noun, but 陆运 is labeled as verb.

It is found that achievements and problems coexist in the treatment of MCM in CED3: the former involves the increasing of the absolute quantity of multi-category lexemes and the expansion of the scope of multi-category lexemes, while the latter refers to the minimization of the number of multi-category lexemes by following the “Principle of Parsimony”. The confusion concerning the word class labeling in Chinese/Chinese-English dictionaries epitomizes exactly the real dilemma of the study on word class in Modern Chinese. From the perspective of the Two-level Word Class Categorization Theory, the word class labeling of dictionary entries is conducted from the communal level at langue. In this case, a corpus-based usage pattern survey is a necessity to make clear the word class labeling conditions of the lexemes, which would further help to correctly clarify the settlement of their class memberships in dictionaries.

A corpus-based case study is to be carried out. The two symmetric lexemes 奖励 and 惩罚, bonded with antonymous relation, are picked out from CED3 to have them testified by the corpus so as to check out whether the number of multi-category lexemes has been deliberately minimized by following the “Principle of Parsimony”.

With the “Principle of Symmetry” in word class labeling as a discovery method, it is found that the word class labeling of the anonymous lexemes 奖励 and 惩罚 fails to be in line with each other. The entry 惩罚 has been labeled as a N-V lexeme, while
奖励 has only been pinned down as a verb. It thus follows that the two lexemes 奖励 and 惩罚 in CED3, bonded with antonymous relation, fail to share a consistent word class labeling fact. That being the case, is it justified to label 奖励 as a N-V lexeme? In other words, we wonder whether or not the usage pattern as a noun of 奖励 in Modern Chinese has been conventionalized.

A full-sample analysis on the POS tagging of 奖励 and 惩罚 has been conducted with the help of the CN CORPUS. Indeed, there are 266 concordances of 奖励, 4 of which are presented to be nouns, while the other 262 are verbs; meanwhile, all the 225 concordances of 惩罚 are tagged as verbs. For details, see Table 14.

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percentage</th>
<th></th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>4</td>
<td>1.50%</td>
<td>0</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Verb</td>
<td>262</td>
<td>98.50%</td>
<td>225</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>Total</td>
<td>266</td>
<td>100.00%</td>
<td>225</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

The results above apparently reflect the inconsistency between the POS tagging in the CN CORPUS and the word class labeling in CED3, which also illustrates the controversy over the genuine word class labeling information of this two lexemes among Chinese scholars. The results above are confusing and make us doubt the accuracy of the POS tagging of 奖励 and 惩罚. Wang Renqiang and Huang Changning (2015) also point out that the CN CORPUS has some problems in POS tagging. Through careful analysis, we do find some POS tagging problems of 奖励 and 惩罚 in this corpus.

There are mainly two kinds of problems of the POS tagging. The first one is the word formation usages of 奖励 and 惩罚 are tagged as verbs. For example, 奖励 in (1), (2) and (3) and 惩罚 in (4) and (5) are word formation usages, but they are tagged as verbs in the CN CORPUS.
（1）奖金与津贴、奖金和奖励是企业为激励员工而付给的。一般按工人工资的一定比例计算。
（2）如企业只受利益机制的作用而无相应机制，行政性、法律性、权力性机制作用，企业会只追求利润，不顾宏观经济效果。企业如果对职工只有奖励权而无用人权，企业就不断面临要求提高工资，性/权/收入的科目/压力而/无/相应/权/机制的抵消，较之两种压力。
（3）作为一/门科学，奖励学/它/向思想家、理论家、党政领导/发出/了/ 呼唤，要求/人们研究它/。发展/它/ 把/学/先进/，树立/标兵，/宣传/典型/的工作/水平/ 提到/一个/新/的高度。
（4）掌握/最高/权力/（/当然/包括/最高/惩罚/权/），/的/人/当然/是/不会/惩罚/自己的/。

Second, reference usages when used as subjects or objects of the sentences are tagged as verbs in the CN CORPUS. For instance, 奖励 in (6), (7), and (8) and 惩罚 in (9) and (10) are usages of reference, but they are tagged as verbs in the CN CORPUS.

（6）学校中种/的/演习/和/过失，不过/是/一/种/演习。
（7）对/揭发/检举/非法/出版物/和/淫秽/书刊/有功/者/，/要/给予/表扬/和/奖励。/
(8) 这个法律文件规定，凡国家行政机关工作人员具备下列表现的，都应该予以奖励：

(9) 如果违背了，就一定受到惩罚。

(10) 惩罚只可禁止儿童破坏学校的规则，而不能建筑良好行为的基础。

The examples above illustrate the problems existing in the POS tagging of 奖励 and 惩罚 in the CN CORPUS, thus the POS tagging of the two lexemes should be revised. With usage pattern as word formation being excluded, the expression of propositional speech act functions of 奖励 and 惩罚 are to be valued in syntax at parole. Besides, both the token frequency and type frequency of this two lexemes have been calculated with an expectation of making clear to what degree has the usage pattern as a noun of 奖励 been conventionalized. Table 15 shows the revised token frequency of 奖励 and 惩罚 respectively. There are 190 concordances in which 奖励 bears the usage pattern as a noun, accounting for 76.00%; and there are 60 concordances in which 奖励 serves as a verb, accounting for 24.00%. Meanwhile, among the 217 concordances of 惩罚，152 of which present the usage pattern as nouns, taking a percentage of 70.05%; the other 65 are identified to be verbs, accounting for 29.95%.

Table 15 Revised POS Tagging of 奖励 and 惩罚 in the CN CORPUS

<table>
<thead>
<tr>
<th></th>
<th>奖励</th>
<th>惩罚</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>Percentage</td>
<td>Frequency</td>
</tr>
<tr>
<td>None</td>
<td>190</td>
<td>76.00%</td>
</tr>
<tr>
<td>Verb</td>
<td>60</td>
<td>24.00%</td>
</tr>
<tr>
<td>Total</td>
<td>250</td>
<td>100%</td>
</tr>
</tbody>
</table>

2 16 concordances of 奖励 are used as word formation; 8 concordances of 惩罚 are functioned as word formation.

3 According to the Two-level Word Class Categorization Theory, in syntax at parole, nouns are used as subjects and objects for the function of reference; verbs are used as predicates for the function of predication.
As Tables 16 and 17 show in detail, the usage patterns of 奖励 and 惩罚 as verbs enjoy a high similarity with respect to their structures, most of which are embodied in the structures of 奖励/惩罚+NP and 对…奖励/惩罚. The structures of 奖励 and 惩罚 functioning as nouns present a relative similarity as well, but the former bears a richer variety of collocation types in the structures. Further analysis reveals that apart from VP+奖励/惩罚 and 奖励/惩罚+NP, the usage patterns as nouns of the two lexemes are also embodied in the structures of 奖励/惩罚+VP, 奖励/惩罚+的+NP, 对…的奖励/惩罚, NP+的+奖励/惩罚 and so on. In addition, 奖励 also appears in structures like NP+奖励, Adj+奖励 and 奖励+似的. Among all the usage patterns of 奖励 as nouns, the number of the structure 奖励+NP ranks the first place with a percentage of 42.11%; noun phrases involved in this structure are various, such as 基金, 制度, 政策, 形式, 工作, 标准, 条例, 办法, 对象, 因素, 问题, 证书 and so on. While the structure VP+奖励 ranks the second with a percentage of 37.37%; the verb phrases related to this structure involve 给予, 受到, 得到, 予以, 获, 作为, 加以, 增加, 制定, 重视, 期望, 带来, 进行, 提供 and so on. To sum up, it is found that the usage patterns of 奖励 as nouns occur in a greater variety of structures, so do collocation types within each structure.

| Verb Usage Patterns of 奖励 and 惩罚 in the CN CORPUS |
|---------------------------------|---------|----------|-------------|---------|----------|
| 结构 | 数量 | 百分比 | 结构 | 数量 | 百分比 |
| 奖励+NP | 51 | 85.00% | 惩罚+NP | 53 | 81.54% |
| 对…奖励 | 3 | 5.00% | 对…惩罚 | 7 | 10.77% |
| NP+奖励（不及物） | 5 | 8.33% | 其他 | 5 | 7.69% |
| 其他 | 1 | 1.67% | —— | —— | —— |

Table 16
<table>
<thead>
<tr>
<th></th>
<th>60</th>
<th>100%</th>
<th></th>
<th>65</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>表17</td>
<td></td>
<td></td>
<td>表17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>奖励</td>
<td></td>
<td></td>
<td>惩罚</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structures</td>
<td>Number</td>
<td>Percentage</td>
<td>Structures</td>
<td>Number</td>
<td>Percentage</td>
</tr>
<tr>
<td>VP + 奖励</td>
<td>71</td>
<td>37.37%</td>
<td>VP + 惩罚</td>
<td>104</td>
<td>68.42%</td>
</tr>
<tr>
<td>奖励+NP</td>
<td>80</td>
<td>42.11%</td>
<td>惩罚+NP</td>
<td>13</td>
<td>8.55%</td>
</tr>
<tr>
<td>NP+奖励</td>
<td>11</td>
<td>5.79%</td>
<td>——</td>
<td>——</td>
<td>——</td>
</tr>
<tr>
<td>奖励+VP</td>
<td>12</td>
<td>6.32%</td>
<td>惩罚+VP</td>
<td>14</td>
<td>9.21%</td>
</tr>
<tr>
<td>奖励+的+NP</td>
<td>8</td>
<td>4.21%</td>
<td>惩罚+的+NP</td>
<td>7</td>
<td>4.61%</td>
</tr>
<tr>
<td>Adj+奖励</td>
<td>4</td>
<td>2.11%</td>
<td>——</td>
<td>——</td>
<td>——</td>
</tr>
<tr>
<td>奖励+似的</td>
<td>1</td>
<td>0.53%</td>
<td>——</td>
<td>——</td>
<td>——</td>
</tr>
<tr>
<td>对…的奖励</td>
<td>2</td>
<td>1.05%</td>
<td>对…的惩罚</td>
<td>11</td>
<td>7.24%</td>
</tr>
<tr>
<td>NP+的+奖励</td>
<td>1</td>
<td>0.53%</td>
<td>NP+的+惩罚</td>
<td>3</td>
<td>1.97%</td>
</tr>
<tr>
<td>Total</td>
<td>190</td>
<td>100%</td>
<td>Total</td>
<td>152</td>
<td>100%</td>
</tr>
</tbody>
</table>

According to the comparative analysis above, both the token frequency and type frequency of 奖励 and 惩罚 enjoy a relatively high similarity with each other. The entry 惩罚 in CED3 has been labeled as a N-V lexeme, 奖励 however has only been pinned down as a verb. In this case, we wonder whether or not the usage pattern as a noun of 奖励 has been conventionalized. The statistics of the revised POS tagging above reveal the fact that the token frequency of 奖励 tagged as nouns is pretty high, accounting for 76% in all. Furthermore, the lexeme 奖励 also presents a relatively high type frequency, and the usage pattern as a noun occurs in 9 types of syntactic structures, in which the words involved exhibit diversity as well. According to the Two-level Word Class Categorization Theory, the usage pattern as a noun of 奖励 has
been the conventionalized. As the conventionalized usages of a language are collected in the dictionaries (Hanks 2013: 91), it is worth considering the inclusion of the usage pattern of “奖励” as a noun in CED3.

5. Conclusion

In brief, it is found that with the help of the DIY Word Class Labeling Database of CED3 and the CN CORPUS, achievements and problems coexist in the treatment of MCM in CED3. In spite of the fact that the absolute quantity of multi-category lexemes has been increased and the scope of multi-category lexemes in CED3 has been expanded to some extent, the number of multi-category lexemes however has been deliberately minimized by following the “Principle of Parsimony” which bears close relations with the first-order logic. In line with the Complex Adaptive System Theory, the Two-level Word Class Categorization Theory would provide an access to the genuine usage patterns of the lexemes when handling MCM of Modern Chinese, which would improve the quality of both the word class labeling and even the whole compilation of Chinese/Chinese-English dictionaries.

Acknowledgments

This study is funded by the Chongqing Social Science Planning Project “A Corpus-based Study of the Compilation of Chinese-English Advanced Learner’s Dictionaries” (2014YBYY083) and the Chongqing Graduate Innovative Research Project “Word Class Labeling in The Chinese-English Dictionary (3rd Ed.): A Study Based on the Two-level Word Class Categorization Theory” (SISUYZ201605).

References

Kretzschmar Jr, William A. 2015. Language and complex systems. Cambridge:
Cambridge University Press.
139-146.


A Study of Wordhood Identification in *The Chinese-English Dictionary* (3rd Ed.)

Dandan Zhou, Renqiang Wang, Ye Xiao, Yuan Zheng, Jiamin Yi & Jianhua Hao
Graduate School, Sichuan International Studies University
569237811@qq.com | wangrenqiang@sisu.edu.cn | 949378636@qq.com;
465714672@qq.com | 2682849120@qq.com | 260362975@qq.com

Abstract

The word is a central notion in descriptive linguistics but has persistently resisted theoretical analysis. The distinction between words and non-words is also a long-standing problem in the study of Modern Chinese grammar and lexicography. In *The Chinese-English Dictionary* (3rd Ed.) published in 2010, the headwords have been labeled with their word classes based on the distinction between words and non-words. In view of the latest developments in the study of wordhood, this paper aims to make an in-depth study of the wordhood identification on the basis of the DIY Database of *The Chinese-English Dictionary* (3rd Ed.). It is found that there are both achievements and problems in the wordhood identification of the dictionary. The problems are most outstanding for monosyllabic and separable headwords: Some of monosyllabic non-words are incorrectly labeled as words; Separable headwords in this dictionary are not clearly distinguished between words and phrases. The deficiencies to a large extent have puzzled dictionary users in their language acquisition and production. From the perspective of language as a complex adaptive system, in order to solve the problems in Chinese-English bilingual dictionaries, the wordhood identification should be based theoretically on the proper definition of word, and operationally on the usage patterns survey in a balanced large-scale Modern Chinese corpus. Only in this way can we improve the quality of dictionary compilation and meet the needs of the learners of Chinese.

Keywords: *The Chinese-English Dictionary* (3rd Ed), wordhood, empirical study

1. Introduction

The distinction between words and non-words has always been considered as a conundrum in the study of Modern Chinese grammar (Lu 2013). Besides Modern Chinese, other languages such as Modern English and Arabic are also faced with such kind of plight (Halliday and Yallop 2007). Preliminary unit of languages as word might be, native speakers can easily tell a word from a non-word by their intuition, while
linguists have always found it difficult to identify and define a word (Robins 1989; Bolinger 1963; Crystal 2000; Halliday and Yallop 2007). Different scholars have tried to define the term “word” from various perspectives. Influenced by structuralism, Bloomfield proposed that “the word is the smallest unit of speech” (Bloomfield 1933), which apparently ignored the fact that some adverbs are words but they cannot construct an utterance independently. From the viewpoint of semantics, Chinese linguist Wang Li (1985) assumed words as “the smallest meaningful units in a language, a word, simply a meaningful unit”. However, such a definition cannot exclude the affixes, which are also meaningful but are non-words. Such a dilemma remains unsolved and has aroused a hot debate among linguists (Guo 2002; Hu 1996, 2004; Lu 2013, 2015; Lü 1979; Shen 2009, 2012, 2015; Wang 2006, 2011, 2014, 2015; Xing 2003; Yuan 1995, 2005, 2010). The nature of the word has been a central topic under discussion for “The Word and the Morpheme” Workshop held in Humboldt Universität zu Berlin, Germany in September 2016. To conclude, the identification and definition of word in linguistics are still unable to be solved and could not reach an agreement.

The distinction between words and non-words is the prerequisite for the comprehensive word class labeling in dictionaries (Wang 2006, 2011; Xu and Tan 2006). The compilation of The Chinese-English Dictionary (3rd Ed) (hereinafter CED3) has referred to The Contemporary Chinese Dictionary (5th Ed) (CCD5 for short) and has made an attempt to achieve the goal of wordhood identification as well as word class labeling for all the headwords. Wang Renqiang (2012) has reviewed CED3 in terms of the wordhood identification and dictionary compilation. In particular, he has affirmed the meticulous manipulation of monosyllabic lexemes. Li Xiang (2013) made an comparison between Modern Chinese monolingual dictionaries and Chinese-English bilingual dictionaries, and commented on the merits and demerits of CED3 based on three levels of linguistic units, i.e morpheme, character and word. However, the previous researches have not dealt with the problem of lexical categorization based on an empirical analysis of the full samples of CED3. On the basis of the Database of the Word Class Labeling Information in CED3, combined with recent theories and approaches at home and abroad, this paper is intended to probe into the following questions: What is the wordhood view of the compilers of CED3? Whether it is the same as that of CCD5? A thorough investigation of the above problems will deepen the study on word class labeling in Chinese-English bilingual dictionaries and improve their quality of compilation, thus severing for the strategy of the internationalization of the Chinese language.

2. Research Design

We set up a database by recording the word class labeling information of each lexeme in CED3 so as to make a comprehensive investigation of the wordhood identification. The Past, Present, and Future of the Dictionary (Preface to the 3rd Edition) and Guide to the Use of the Dictionary have been a main reference for the construction of the database. As for a small amount of lexemes whose parts-of-speech tagging is not fully represented,
their information is completed based on the reference.

All the lexemes in the Database of the Word Class Labeling Information in CED3 have been divided into three categories, namely, Group 1, which consists of lexemes whose word class labeling information is fully represented; Group 2, which includes lexemes whose word classes are partially labeled; and Group 3, which contains lexemes whose word classes are not indicated in CED3.

3. Statistical Findings

The statistics of word class labeling in CED3 show that there are 89041 lexemes in the whole dictionary. And Group 1 occupies 90.11%; Group 2 occupies 0.05%; and Group 3 occupies 9.84%, as shown in Table 1.

<table>
<thead>
<tr>
<th>Group</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>80231</td>
<td>47</td>
<td>8763</td>
<td>89041</td>
</tr>
<tr>
<td>Percentage (%)</td>
<td>90.11</td>
<td>0.05</td>
<td>9.84</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Three groups (Groups 1, 2 and 3) are further divided according to the number of characters. As shown in Table 2, in Group 1, the disyllabic lexeme accounts for the highest percentage (72.44%), and trisyllabic lexeme and monosyllabic lexeme are in the second place, quadrisyllabic lexemes and others account for the lowest percentage (3.27%). In Group 2, the monosyllabic lexeme accounts for the highest percentage (74.47%), and trisyllabic lexeme and disyllabic lexeme are in the second place, and quadrisyllabic lexeme accounts for the lowest percentage (6.38%). In Group 3, the quadrisyllabic lexeme accounts for the highest percentage (77.45%), and monosyllabic word and trisyllabic lexeme are in the second place, and disyllabic lexeme accounts for the lowest percentage (0.38%).

<table>
<thead>
<tr>
<th>Number of Characters</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monosyllabic lexeme</td>
<td>8700</td>
<td>35</td>
<td>35</td>
<td>457</td>
</tr>
<tr>
<td>Disyllabic lexeme</td>
<td>58118</td>
<td>4</td>
<td>4</td>
<td>58</td>
</tr>
<tr>
<td>Trisyllabic lexeme</td>
<td>10795</td>
<td>5</td>
<td>3</td>
<td>0.58</td>
</tr>
<tr>
<td>Quadrisyllabic lexeme</td>
<td>2161</td>
<td>3</td>
<td>3</td>
<td>47</td>
</tr>
<tr>
<td>Others</td>
<td>457</td>
<td>0</td>
<td>0</td>
<td>80231</td>
</tr>
<tr>
<td>Sum</td>
<td>80231</td>
<td>47</td>
<td>89041</td>
<td></td>
</tr>
</tbody>
</table>
From the perspective of the internal constitution of three groups, it seems that the number of characters of a lexeme is closely related to which group the lexeme belongs to. As shown in Table 3, the group belonging of the monosyllabic word is relatively complex. Group 1 accounts for the highest percentage (90.43%), followed by Group 3 (9.21%) and Group 2 (0.36%). The disyllabic lexemes almost belong to the Group 1 (99.94%). And there are only 4 disyllabic lexemes in Group 2, 33 disyllabic lexemes in Group 3.

Table 3 Internal Distribution of Characters with Different Groups in CED3

<table>
<thead>
<tr>
<th>Number of Characters</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monosyllabic lexeme</td>
<td>Number</td>
<td>8700</td>
<td>35</td>
<td>886</td>
</tr>
<tr>
<td></td>
<td>Percentage (%)</td>
<td>90.43</td>
<td>0.36</td>
<td>9.21</td>
</tr>
<tr>
<td>Disyllabic lexeme</td>
<td>Number</td>
<td>58118</td>
<td>4</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Percentage (%)</td>
<td>99.94</td>
<td>0.00</td>
<td>0.06</td>
</tr>
<tr>
<td>Trisyllabic lexeme</td>
<td>Number</td>
<td>10795</td>
<td>5</td>
<td>122</td>
</tr>
<tr>
<td></td>
<td>Percentage (%)</td>
<td>98.84</td>
<td>0.05</td>
<td>1.11</td>
</tr>
<tr>
<td>Quadrisyllabic lexeme</td>
<td>Number</td>
<td>2161</td>
<td>3</td>
<td>6787</td>
</tr>
<tr>
<td></td>
<td>Percentage (%)</td>
<td>24.14</td>
<td>0.04</td>
<td>75.82</td>
</tr>
<tr>
<td>Others</td>
<td>Number</td>
<td>457</td>
<td>0</td>
<td>935</td>
</tr>
<tr>
<td></td>
<td>Percentage (%)</td>
<td>32.83</td>
<td>0</td>
<td>67.17</td>
</tr>
</tbody>
</table>

By analyzing all kinds of grammatical units in CED3, it is found that there are 94899 grammatical units, as shown in Table 4. Among them, there are 86085 words (90.71%) and 8814 non-words (9.28%) of grammatical units respectively.

Table 4 Number and Percentage of Different Grammatical Units in CED3

<table>
<thead>
<tr>
<th>Parts of speech</th>
<th>Number</th>
<th>Percentage (%)</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td></td>
<td></td>
<td>86085 (90.71%)</td>
</tr>
<tr>
<td>Noun</td>
<td>47534</td>
<td>50.09</td>
<td></td>
</tr>
<tr>
<td>Verb</td>
<td>26716</td>
<td>28.15</td>
<td></td>
</tr>
<tr>
<td>Adjective</td>
<td>8560</td>
<td>9.02</td>
<td></td>
</tr>
<tr>
<td>Adverb</td>
<td>1567</td>
<td>1.65</td>
<td></td>
</tr>
</tbody>
</table>
In CED3, there are 5209 multiple class membership lexemes, occupying 5.85% of the whole lexemes. And the majority of the lexemes, occupying 90.71% of the whole lexemes, have single class membership, as shown in Table 5.

### Table 5 Distribution of Multi-category Lexemes in CED3

<table>
<thead>
<tr>
<th></th>
<th>Single class lexeme</th>
<th>Multi-category lexeme</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Percentage</td>
<td>Subtotal</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>89041</td>
</tr>
<tr>
<td>Number</td>
<td>83832</td>
<td>94.15</td>
<td>5209</td>
</tr>
<tr>
<td>Percentage (%)</td>
<td>5.27</td>
<td>0.50</td>
<td>5.45</td>
</tr>
</tbody>
</table>

4. Discussion

The primary principle of word class labeling for dictionaries of Chinese as a second or foreign language is to label all the lexemes with word class information based on the distinction between words and non-words (Wang 2006). The preface of CED3 says that in word class labeling, it refers to CCD5 and makes a supplement or a correction of the word class labeling, if necessary. This paper makes a comparative data analysis of the word class labeling between CCD5 and CED3. By doing so, it is found that the word class labeling in those two dictionaries reflects prototype effects. Firstly, there is no clear boundary between words and non-words. Instead, they form a continuum in CED3. Secondly, the disyllabic lexemes account for the highest percentage (72.44%), trisyllabic lexemes rank only second two the disyllabic (13.45%), and monosyllabic and quadrisyllabic lexemes account for the lowest percentage (10.84% & 2.69%). According
to the result, it is further proved that the prototype of Modern Chinese words is disyllabic lexeme, the trisyllabic lexeme is less prototypical, the monosyllabic lexeme, the quadrisyllabic lexeme and other lexemes are atypical (Wang 2011). Thirdly, the statistical results show that a lexeme in the dictionary can belong to two or more grammatical categories. For one thing, there are 35 single class lexemes which are combination of words and non-words. For another, part of the lexemes belong to two or more word classes. Therefore, this paper analyzes two problems in detail to which we need pay much attention in the compilation of Modern Chinese dictionaries and Chinese foreign dictionaries. Those two problems are the distinction between words and non-words in monosyllabic lexemes, and the grammatical hierarchy of separable words in CED3 (Wang 2015). There is an evident difference in the treatment of monosyllabic lexemes of Group 2 in those two dictionaries: There are only 35 in CED3 while 2278 in CCD5. Based on the Complex Adaptive System(CAS), this paper intends to analyze the difference with the help of Modern Chinese Corpus of National Language Committee and put forward some suggestion for the separable words in the dictionary.

4.1 Enlightenment of CAS on Word Class Labeling of Monosyllabic Headwords

Complex Adaptive System, proposed by John Holland in 1994, introduces the concept of agents with adaptive ability, and guides humans to come to understand and describe behaviors in complex systems from the interaction of agents and environment. Thus, the appearance of CAS provides a new orientation for the research of modern systems science. The successful application of CAS in systems engineering has prompted academic circles to try to use CAS to observe and study the complex systems in various fields. Then the complex adaptive system is applied to the linguistics, forming the theory of language as a Complex Adaptive System.

Language as a complex adaptive system is an umbrella concept, in which Usage-based Theory is its subordinate theory. Usage-based theory (UBT), stemming from the study of function and cognition of language, considers language as a complex adaptive system. Its core idea is that the knowledge of language comes from patterns of language use, and the structure of language emerges in language use. The frequency of use is crucial to the cognitive representation and conventionalization of language structures (Bybee 2010). In other words, the language community creates and inherits new expressions in the process of communication, among which the language facts of high frequency can be conventionalized as knowledge of the communal language. With regard to the specific unit of language, some words gradually evolved into the peripheral member of wordhood, while some morphemes gradually become the central member of wordhood. The main ideas of UBT and large-scale balanced corpus complement each other, which is the perfect combination of theory and methodology.

The preface of CED3 clearly explained that on the basis of the existing labeling, the lexemes have been further labeled, including labeling monosyllabic lexemes. In other words, on the premise of distinguishing between non-words and words, non-words have been further divided into meaningless syllables, monosyllabic morphemes and affixes. After comparing between CED3 and CCD5, it can be found that the lexemes have been
handled in quite different ways, especially for lexemes whose role can be both words and non-words (the lexemes in the Group 2). As shown in Table 6, the monosyllabic lexemes contained in the two dictionaries are roughly equal in number, while lexemes in Group 2 of CED3 account for only 0.21% of the total lexemes, and in CCD5, lexemes in Group 2 accounts for 21.01% of the total lexemes. Compared with CCD5/6/7, the inappropriate labeling of monosyllabic lexemes in CED3 mainly lies on: (1) Some of monosyllabic non-words are incorrectly labeled as words; (2) Some headwords which belong to morphemes and words are improperly labeled as words only; (3) Word classes of some lexemes in CED3 are completely omitted. Factors of the differences between the two dictionaries are as follows:

Table 6: Monosyllabic Lexemes in the Database of CED3 and CCD5

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>CED3</td>
<td>8700</td>
<td>35</td>
<td>886</td>
<td>9621</td>
</tr>
<tr>
<td>Percentage (%)</td>
<td>90.43</td>
<td>0.36</td>
<td>9.21</td>
<td></td>
</tr>
<tr>
<td>CCD5</td>
<td>3005</td>
<td>2278</td>
<td>5562</td>
<td>10845</td>
</tr>
<tr>
<td>Percentage (%)</td>
<td>27.71</td>
<td>21.01</td>
<td>51.28</td>
<td></td>
</tr>
</tbody>
</table>

(1) Some monosyllabic non-words are incorrectly labeled as words. As shown in Table 7, the word “官 2” in CED3 is marked as a noun, but “官 2” is not used independently in Modern Chinese, only as a morpheme in other words, such as “五官”, “感官”, “器官”; similarly, “稽 2” is marked as a verb, but is not used independently as well, only as a morpheme for “稽留”, “稽延” and so on; although “虏” is labeled as both noun and verb, but only used as morpheme for verb “虏获” and noun “敌虏”, “强虏” and so on. Based on the use pattern of Modern Chinese Corpus of National Language Committee, it is found that the word “虏” appeared in the corpus for a total of 190 times, and the frequency of the use as a word was four times while as a non-word 186 times. Therefore, “虏” in Modern Chinese can not be treated as an independent word, but a morpheme for words like “虏获”. After comparing four dictionaries, the principle of labeling in CCD5/6/7 is more in line with the usage mode. For instance, “味” is labeled as both a verb and adjective, but in fact it can not be used independently, but as a morpheme for words, like “拾金不味”, “味良心”, “味死”, “蒙味”, “愚味”, “素味平生”, “幽味”. 
There are many other words like this, such as “复”, “干 1”, “刊”, “瞭”, “奴”, “冥”, “秘”, “盲”, “纶”, “陋”, “邻”, “劣”, “怜”, “丽 2”, “酪”, “琅”, “菜”, “恳”, “遨”, “履”, “苛” and so on.

**Table 7** Same Monosyllabic Non-words in CCD5/6/7 Labeled as Words in CED3

<table>
<thead>
<tr>
<th>CED3</th>
<th>CCD5/6/7</th>
</tr>
</thead>
<tbody>
<tr>
<td>官 2</td>
<td>guān 名 organ: 感~ sense organ</td>
</tr>
<tr>
<td>稷 2</td>
<td>jī &lt;书&gt; delay; procrastinate</td>
</tr>
<tr>
<td>虏</td>
<td>lǔ 名 1 Captive; prisoner of war 2 &lt;古&gt; slave ②</td>
</tr>
<tr>
<td></td>
<td>&lt;古&gt; enemy: 强~ strong enemy 动② take prisoner</td>
</tr>
<tr>
<td>觑</td>
<td>mèi 名 1 hide; conceal: <del>良心 go against one's conscience 2 offend; risk: 冒</del> take the liberty ②</td>
</tr>
</tbody>
</table>

(2) Some headwords which belong to morphemes and words are improperly labeled as words. As shown in Table 8, CED3 marked “官 1” as a noun and adjective. There is no doubt that “官 1” in Modern Chinese can really be noun, however it cannot be used alone as an adjective. We use it as the word-building morpheme of “官印”, “官办 “, “官费”, “官道”, “官话” and other words. In the following table, “贱” has the similarly improper handling with “官 1” because they all do not distinguish morphemes with independent words. Besides, the five senses of “脑” are all marked as nouns in
CED3. While CCD5/6/7 takes steps to distinguish them between words and non-words. For example, the second, fourth, fifth senses of “脑” cannot be worked as an independent word but only for word formation. Compared with CED3, it is inaccurate to mark the three senses as nouns. Statistics show that there are 9621 monosyllabic lexemes in CED3 and only 35 of them are treated as mixture of words and non-words (Group 2). They are “岸、秘、螭、噁、柯、类、爨、面、氏、黼、超、反、分、溜、缓、略”、无、有、苦、然、昙、尔、三、勗、甸、第、粲、瞿、沱、耶、邪、綦、师”.

### Table 8  Headwords Belonging to Morphemes and Words in CCD5/6/7 are Labeled as Words in CED3

<table>
<thead>
<tr>
<th></th>
<th>CED 3</th>
<th>CCD 5/6/7</th>
</tr>
</thead>
<tbody>
<tr>
<td>官1</td>
<td>guān ① ② government official; officer;</td>
<td>guān</td>
</tr>
<tr>
<td></td>
<td>office-bearer: 高→high official ③ (Guān) a</td>
<td>高政府机关或军队中经过任命的，一定等级以上的公职人员：～员</td>
</tr>
<tr>
<td></td>
<td>surname ①② governmental; official:～官</td>
<td>指属于政府的或公家的：～印</td>
</tr>
<tr>
<td></td>
<td>run by the government ② &lt;旧&gt; public:～道</td>
<td>公用的：～道</td>
</tr>
<tr>
<td></td>
<td>public road</td>
<td></td>
</tr>
<tr>
<td>贱</td>
<td>jià ①② low-priced; inexpensive; cheap (opp.贵):～卖 sell cheap ③ lowly; humble (opp.贵):～</td>
<td>jià ①② (价钱)低（跟“贵”相对，下⑧同）：～卖</td>
</tr>
<tr>
<td></td>
<td>low-down; base; despicable:她很～！She's so cheap!③＜谦＞my:</td>
<td>地位低下：～贫</td>
</tr>
<tr>
<td></td>
<td>您贵姓？～姓王。Your surname, please?</td>
<td>卑鄙；下贱：～骨头。③谦辞，用于称跟自己有关的事物：（您）贵姓？～姓王。</td>
</tr>
<tr>
<td></td>
<td>My surname is Wang.</td>
<td></td>
</tr>
<tr>
<td>脑</td>
<td>nǎo ①② brain;用～过度 overtax once brain ② head ↓ brains; mind;</td>
<td>nǎo ①</td>
</tr>
<tr>
<td></td>
<td></td>
<td>动物中枢神经的主要部分，位于头部。人脑管全身。</td>
</tr>
</tbody>
</table>
(3) Word classes of some lexemes in CED3 are completely omitted. For example, “趿” and “钙” belong to this situation:

趿 tā 素
钙 gài 【化】 calcium (Ca): 补~ take extra calcium; take calcium supplements.

“趿” and “钙” in CCD5/6/7 are all labeled as words, for “钙” is a technical term in chemistry and it can be treated as a noun. Relatively speaking, CED3 just shows the usage field of “钙” and believes that “钙” can only be used as a word formation to constitute “补钙” and other words.

Based on the big data analysis, the difference between the four dictionaries in the treatment of monosyllabic headwords may be due to the fact that the compilers of CED3 mislabeled the monosyllabic headwords as words, which not only leads to the improper wordhood distinction in the dictionaries, but may also cause confusion to the dictionary users. This may be related to the codification theory on which the dictionary compilers are based. For example, in the preamble of CED3, it mentions that some characters are not used alone nowadays, but in ancient times they were words, for instance, “蝮” in “蝮蛇” was a monosyllabic word in ancient times. Lexicographers of CED3 still regarded characters like this as independent words, such as “蝮”, just added the style marker <ancient>. CED3 does not distinguish the use model of monosyllabic lexemes in Ancient Chinese and Modern Chinese, ignoring that language is dynamic in diachrony. Some of the uses of words are inevitably to be abandoned, while the new
uses of words continue to emerge. In addition, the Chinese characters and the words are not the same, we should not confuse them with each other. Studies show that only a very small number of characters in Modern Chinese can be independent words, the vast majority used only for word formation, such as morphemes “邻” in “邻居”，“怜” in “可怜” and so on. The handling of monosyllabic lexemes in CED3 to some extent confuses the essential difference between characters and words.

4.2 The Analysis of Separable Words

The separable word has always been regarded as a knotty issue in the study of Modern Chinese, whose definition is critical to the treatment of the headwords in the dictionaries. Separable words mean disyllabic compounds in Modern Chinese with a conventional sense, which can be separated or united, considered either as word or phrase. When not separated, it is a word; When separated, it is often used as phrase (Wang 2011). As the separable frequencies of them are different, it is difficult to define them either as words or phrases in a continuum. According to the Guide of CED3 and CCD5/6/7, some of the multi-syllabic lexemes can be inserted with other components, so its pronunciation notation is separated by double slash “//”. If the separable word has multi-class membership, it is dealt with in another way, that is to say, double slash “//” is added in the verb and adjective senses respectively. This kind of treatment does not explicitly show that the separable word has the scalar attributes of both words and phrases. Compared to the separable words collected in the four dictionaries (CED3 and CCD5/6/7) as shown in Tables 9 and 10 below, the findings are as follows: (1) For separable words of lower separable frequencies, such as “出席”, “节能” and so on, few of them can be used as separable words. They can only be treated as words in CCD5/6/7, while they are treated as separable words in CED3, which have both the use of words and phrases. (2) For separable words of higher separable frequencies, such as “吃惊”, “干活儿”. CED3 and CCD5/6/7 both handle them as the same scale with the use of the word and phrase. The lexemes are not distinguished between word and phrase, which brings great trouble to the teaching and learning of Chinese as a foreign language. In view of this, the separable frequencies of words should be taken into consideration in the compilation of dictionaries: for separable words of lower degree of separable frequencies, they are labeled as words which is favorable for the language output of the bilingual dictionary users; for the separable words of the high degree of separable frequencies, it’s better to label them as word and phrase considering its two levels of scalar attributes.

Table 9 Samples with Low Separable Frequencies in CED3 and CCD5/6/7

<table>
<thead>
<tr>
<th>Dictionary Lexeme</th>
<th>CED3</th>
<th>CCD5/6/7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

372
出席 chūxí be present (at a meeting, social gathering, etc.); attend (opp. 缺席): ~会议 attend a meeting / ~宴会 be present at a banquet ◇ ~人数 number of people present; attendance

节能 jié néng save energy; energy conservation ◇ ~灯 energy-saving light (or bulb) / ~法 energy-saving act / ~灶 energy-saving stove

Table 10 Samples with High Separable Frequencies in CED3 and CCD5/6/7

<table>
<thead>
<tr>
<th>Dictionary Lexeme</th>
<th>CED3</th>
<th>CCD5/6/7</th>
</tr>
</thead>
<tbody>
<tr>
<td>吃惊 chī//jīng</td>
<td>startled; shocked; amazed; taken aback: 吃了一惊 flabbergasted; surprised / 他那坚强的毅力使人～。 His will power is amazing.</td>
<td>吃惊 chī//jīng 受惊; 令人～</td>
</tr>
<tr>
<td>干活儿 gàn//huó-r</td>
<td>work; work on a job: <del>去吧。Let’s get to work. /他们都在</del>呢。They are all at work. or They are all on the job. / 今天你干什么活儿啊? What your job for today? / 他是干木匠活儿的。He works as a carpenter.</td>
<td>干活儿 gàn//huó-r 从事体力劳动，泛指做事、工作，帮妈妈多干些活儿</td>
</tr>
</tbody>
</table>

5. Conclusion

To summarize, it is found that there are both merits and demerits in the identification of wordhood in CED3. This dictionary realized for the first time in its history a comprehensive word class labeling on the basis of the distinction between words and non-words, especially the meticulous manipulation of monosyllabic headwords. However, based on the empirical study of all the lexemes in CED3, it is found that there is divergence in the treatment of monosyllabic headwords and separable words in this dictionary and in the mainstream Modern Chinese studies, which factually reflects the true picture of the word class studies in analytic languages.

Acknowledgment

This study is funded by the Chongqing Social Science Planning Project “A Corpus-based Study of the Compilation of Chinese-English Advanced Learner’s

References

CSF-based Processing of Near-synonyms in the EFL Dictionaries for Chinese Learners

Huixiu Liu
School of Foreign Studies, Central South University
liuhuixiu526@sohu.com

Abstract

As the cognitive target, the near-synonym series can be regarded as a kind of cognitive semantic frame (CSF), which is composed of category, category members, attributes and values, conceptual invariants and constraints. In such a cognitive semantic frame, the near-synonyms represent the categorization and sub-categorization of human experience. The activation of one category member always interacts with the activation of relevant knowledge and background. The CSF-based notion will help to solve many controversial and bewildering problems in the compilation of entries of near-synonyms in the English learner's dictionaries, including the establishment of near-synonym entries, the selection of the dominant, the salience of semantic attributes and values, the sequence of near-synonyms within the entry, the granularity of discrimination, the typicality of illustrations and the multi-dimensions of discrimination of near-synonyms.

Keywords: Near-synonyms; Frame; cognition; EFL learner's dictionaries

1. Paradox of the Discrimination of the Near-synonyms

On account of the anisomorphisms of English and Chinese near-synonyms, the teaching and learning of English near-synonyms is a hard job for most Chinese teachers and learners. Similarly, there have been quite a few scholars who complain the embarrassment of the discrimination of near-synonyms in the compilation of synonym dictionaries. There are several puzzling and bewildering problems and contradictions to be solved, including the indeterminacy in listing the number of near-synonyms, simplicity and vagueness of purely semantic discrimination and arbitrariness and subjectivity of the compiler and so on. In this research, the key point under discussion is how to discriminate the near-synonyms and how to represent the near-synonym information in the EFL dictionaries for Chinese learners.
2. Cognitive Semantic Frame

Cognitive semantic frames (CSF) are a kind of semantic theory founded on the notion of human cognition, semantics and frames. They approach the description of lexical meaning in quite a different way. All the entities in the world are the human cognitive targets. Frames serve as a type of cognitive structuring device that provide the background knowledge and motivation for the existence of words in a language as well as how they are used in discourse. The frames in questions can be simple—small static scenes or states of affairs, simple patterns of contrasts, relations between entities and the roles they serve, or possibly quite complex event types — which can be called scenarios — that provide the background for words that profile one or more of their phrases or participants. For example, the word *bartender* evokes a scene of service in a setting where alcoholic beverages are consumed, and profiles the person whose role is to prepare for and serve these beverages. In a sentence like *the bartender asked for my ID*, it is the individual who occupies that role that we understand as making the request, and the request for identification is to be understood against the set of assumptions and practices of that frames.

The abundance of near-synonyms in all the languages has resulted from the development of language vocabulary and the evolution of human cognition. The near-synonyms which have nearly the essential meaning and different connotative meaning can be regarded as the categories within the same semantic frame. The categories have different numbers of members, which have similar attributes and values as well as distinctive attributes and values. Owing to limitations of human perception and cognition, there are different numbers of near-synonyms in different cognitive field. For example, there are larger numbers of near-synonyms for the verb “die” in almost all the languages than the verb for “be born”. It is estimated that there are more than 100 near-synonymous saying for the concept of death both in English and Chinese. One of the various reasons can be attributed to the human different generalization and categorization of the time, place, targets, causes, state and style of death. However, there is a much too simplistic description of the time, place, state, style and targets of birth, thus reducing greatly the number of near-synonyms for the word “be born”. Similarly, as described in chapter 4, the movements and activities of the plants, animals and human beings in the world are manifold and diversified in terms of time, place, style, state, causes, results, scopes and purposes. With the limitations of human perception and cognition, there cannot be equivalent categorizations of all the movements and activities in terms of all the perspectives. Some of the movements and activities can be observed or perceived easily and then described in language, therefore generating a high degree of lexicalization. Comparatively, on one hand, the human movements and activities can be perceived or recognized than those of the plants and animals except that specialists have professional insights of movements and activities of plants and animals. On the other hand, the human movements and activities on the land can also be easily observed and
described in language than those of fish in the water or those of birds in the sky. Therefore, despite the fact that the movements and activities in the sky (flying) or in the water (swimming) may be equally diversified, in reality there are far more lexicalized near-synonyms for describing the human activities on the land than those for activities in the sky or in the water. That is why there are a lot of near-synonyms for walk and fewer near-synonyms for fly and swim in most languages. The manners of walking can be directly and ubiquitously perceived and then lexicalized. In contrast, the manners of flying high in the air and swimming in the water can hardly be observed or perceived directly, thus resulting in the low degree of lexicalization of such fields. Similar cognitive semantic frames can be applied to explain rich near-synonyms for look, speak, and hold, angry, laugh and poor near-synonyms for smell, taste, read, write and so on.

3. The CSF-based Processing of Near-synonym Entries in the EFL

The greatest advantage of the cognitive semantic frame is that it provides a special and significant approach to observe and explain the lexical meaning. And at the same time, it helps to explore and discriminate similarities and difference between near-synonyms. In such a frame, the lexical items represent the categorization and sub-categorization of human experience. The activation of one category member always interacts with the activation of relevant knowledge and background. Based on the theories of frame semantics and cognitive linguistics, we tend to place the discrimination of near-synonyms in the dynamic cognitive semantic frame (CSF). To discriminate near-synonyms in the entry is just to acquire knowledge of a cognitive semantic frame. In a great extent, the CSF-based notion will lay solid and scientific theoretical foundations during the process of compiling entries of near-synonyms in the English dictionaries for Chinese EFL learners.

3.1 Establishment of Near-synonym Entries

With the deep exploration into lexical sense, linguists and other scholars have made all kinds of distinctions of lexical meaning. For example, the British linguist Lyons (1995) distinguishes lexical meaning into sense, reference and denotation whereas Leech (1981) classified lexical meaning into conceptual, connotative, social, emotive, reflected and collocative meanings. So the problem of what is the similarity of near-synonyms comes into being. In the perspective of cognitive semantic frame, the near-synonyms at least belong to the same categorization members at the same cognitive level. The near-synonyms within the entry should share a high degree of similar attributes and values. On account of the difference of human cognitive categorization and lexicalization, the numbers of near-synonyms may vary from two or three to a dozen or even dozens. When establishing the near-synonyms entries, considerations should be taken into not only the lexical similarity of near-synonyms but also the cognitive needs of the semantic frame for the readers of English learning.
The latter is influenced and determined by such factors as the scale of the dictionary, the proficiency of the intended readers.

### 3.2 Selection of the Dominant

The dominant is the introductory and initial lexeme which is usually placed at the very beginning within the near-synonyms entries. According to Apresjan (2000), each synonym series begins with a dominant, that is, a lexeme which has the most general meaning in the given series, has the broadest application and co-occurrence, and is most neutral from the point of view of style, pragmatics, communicative values, grammar, and prosody, etc. In many cases the meaning of the dominant is fully present in the meaning of all other members of the series. However, the dominants may differ from the other synonyms by positive semantic components which the other synonyms lack. This applies in even larger measure to synonym series which have a more complex structure of meaning. In the perspective of cognitive semantic frame, the dominant should be the prototype of the human categorization, which has the largest numbers of attributes in common with other members of near-synonyms of the category and all the smallest number of attributes which also occur with members of neighboring categories. The dominants are almost always the foregrounded lexemes, which are most extensively used in a language and the most deeply rooted in the whole verbal culture reflected in that language. They retain semantic traces of the varied situations in which they alone may be used and in which they suggest specific senses and connotations. For example, within the cognitive semantic frame of the near-synonyms series amaze, astonish, shock, surprise, the dominant should be surprise, which is more typical than the others. Within the cognitive semantic frame of the near-synonyms series dash, hurry, hasten, run, rush, the dominant should be run, which appear as a general and universal lexeme, able to take the place of any other word in the series. Close analysis, however, reveals that it has a number of semantic properties which distinguish it from the other members of the series.

### 3.3 Sequence of Near-synonyms within the Entry

It is not reasonable and scientific to place the near-synonyms with the entry alphabetically or randomly in the practices of traditional near-synonyms dictionaries and English learner’s dictionaries. On account of the prototype theory and family resemblance principles, the dominants as the cognitive prototype should be placed in the very front. Then come the secondary prototypical members and last come the periphery and non-prototypical members. Such a sequence helps to emphasize the close semantic links between near-synonyms to achieve a clearer description of the semantic structure of the near-synonyms series. Thus the near-synonyms series draw, drag, haul, pull, tug, tow should be ordered as pull, draw, drag, tug, haul, tow in order to emphasize their close semantic links.
3.4 Salience of Semantic Attributes and Values

The process of attention is a well-known basic phenomenon in cognitive psychology. Attention appears to be closest to what Chafe (1994) calls the focus of consciousness. Attention comes in degrees and is usually modeled in terms of degree of activation of conceptual structures in a neural network model of the mind. The attention focuses on the human cognitive ability involved, but there are also natural properties of phenomena in the perceived world that lend themselves to being attended to by human beings, and these properties are said to enhance those phenomena’s salience to human being’s attention. Attention is a complex psychological ability whose different aspects can be most easily illustrated by visual ability: one can select one object or another to focus one’s attention on; focus of attention is surrounded by a scope of attention; one can take a more coarse-grained or fine-grained view of a scene; and one can fix one’s gaze on a scene or move one’s eye over it. As we have seen, cognitive semantic frames of near-synonyms can be conceived as a way of describing the cognitive context which provides the background for and is associated with cognitive categories. What the compilers want the reader to recognize most is the salient semantic attributes and values in the frame of near-synonyms.

3.5 Granularity of Discrimination

To introduce the notion of granularity to our discussion, we have to return to the problem of defining near-synonyms. Semanticists such as Ullmann (1962), and Cruse (1986), and Lyons (1995) have attempted to define near-synonymy by focusing on “propositional” meaning. Cruse, for example, contrasts cognitive synonyms and plesionyms; the former are words that, when inter-substituted in a sentence, preserve its truth conditions but may change the expressive meaning, style or register of the sentence or may involve different idiosyncratic collocations (e.g. violin and fiddle), whereas intersubstituting the latter changes the truth conditions but still yields semantically similar sentences (e.g. misty and foggy). Although these definitions are important for truth-conditional semantics, they are not very helpful for us. Moreover, a rigorous definition of cognitive synonyms is difficult to come up with, because it relies on the notion of granularity.

Lexicographers, on the other hand, have always treated synonymy as near-synonymy. They define synonymy in terms of likeness of meaning, disagreeing only in how broad the definition ought to be. For instance, Roget followed the vague principle of “the grouping of words according to ideas” (Chapman, 1992). And in the hierarchical structure of Roget’s Thesaurus, word senses are ultimately grouped according to proximity of meaning: “the sequence of terms within a paragraph, far from being random, is determined by close, semantic relationships (Chapman, 1992: 26).” The lexicographers of Webster’s New Dictionary of Synonyms define a synonym as “one of two or more words…which have the same or very nearly the same essential meaning or more words…. Synonyms can be defined in the same terms up to a certain point (Egan, 1942).” Clearly the near-synonyms must have the same essential meaning
but may differ in peripheral or subordinated ideas. Cruse (1986) actually refines this idea and suggests that synonyms (of all types) are words that are identical in “central semantic traits” and differ, if at all, only in “peripheral traits”. But how can we specify formally just how such similarity of central traits and dissimilarity of peripheral traits is allowed? To answer this question, we introduce the idea of granularity of representation of word meaning. By granularity we mean the level of detail used to describe or represent the meaning of a word. A fine-grained representation can encode subtle distinctions, whereas a coarse-grained representation is crude and glosses over variation. Granularity is distinct from specificity, which is a property of concepts rather than representations of concepts. For example, a rather general (unspecific) concept, say HUMAN, could have in a particular system, a very fine-grained representation, involving, say, a detailed description of the appearance of a human, references to related concepts such as EAT and PRCREATE, and information to distinguish the concept from other similar concepts such as ANIMAL. Conversely, a very specific concept could have a very coarse-grained representation, using only very general concepts.

Near-synonyms can occur at any level of specificity, but crucially it is the fine granularity of the representations of their meanings that enables one to distinguish one near-synonym from another. Thus, any definition of near-synonymy that does not take granularity into account is insufficient. For example, consider Cruse’s cognitive synonymy discussed above. On the one hand, at an absurdly coarse grain of representation, any two words are cognitive synonyms (because every word denotes a “thing”). But on the other hand, no two words could ever be known to be cognitive synonyms, because, even at a fine grain, apparent cognitive synonyms might be further distinguishable by a still more fine-grained representation. Thus, granularity is essential to the concept of cognitive synonymy, as which pairs of words are cognitive synonyms depends on the granularity with which we represent their propositional meanings. The same is true of Cruse’s plesionyms. So in the end, it should not be necessary to make a formal distinction between cognitive synonyms and plesionym. Both kinds of near-synonyms should be representable in the same formalism. By taking granularity into account, we can create a much more useful definition of near-synonymy, because we can now characterize the difference between essential and peripheral aspects of meaning. If we can set an appropriate level of granularity, the essential meaning of a word is the portion of its meaning that is representable only above that level of granularity, and peripheral meanings are those portions representable only below that level.

But what is the appropriate level of granularity, the dividing line between coarse-grained and fine-grained discrimination during the compilation of EFLD. We could only simply use our intuition—or rather, the intuitions of lexicographers, which are filtered by some amount of objectivity and experience. Alternatively, from a concern for the discrimination of lexical knowledge in a multilingual application, we can view words as (language-specific) specializations of language-independent concepts. Given a hierarchical organization of coarse-grained language independent concepts, a set of near-synonyms is simply a set of words that all link to the same
language-independent concept (DiMarco, Hirst, and Stede, 1993). So in this view, near-synonyms share the same propositional meaning just up to the point in granularity defined by language dependence. Thus we have an operational definition of near-synonymy: If the same concept has several reasonable lexicalizations in different languages, then it is a good candidate for being considered a language-independent concept, its various lexicalizations forming sets of near-synonyms in each language.

3.6 Typicality of Examples

Examples in reference works are used to illustrate a particular form or meaning in a wider context. They can be either based on objective evidence such as from a citation file or corpus or be invented by the compiler. Examples play a dual role: first they form a substantial part of the material upon which our description of a near-synonyms series is constructed; second, they demonstrate how the features of these near-synonyms are manifested in various styles and genres of literature and speech (Hartmann, 1998).

It should be noted that the examples include numerous instances or individual usage of near-synonym by writers and that this usage is characterized by departures from established co-occurrence norms, by unexpected metaphorical reinterpretation of a word, and so on. Although such materials are of great value as evidence of the great wealth of possibilities of use in the living language and as an indicator of the direction of potential change, they should not be considered when describing the semantic, syntactic or co-occurrence norms of near-synonym uses as set forth in the analytical part of the dictionary. The final purpose of examples in the near-synonym entries is for the provision for the patterns of near-synonyms discrimination and production. It can be said not all entries of near-synonyms discrimination need the examples. In some extent, whether examples are needed or not depend on the pattern and the granularity of discriminations. For instance, if the method of semantic discrimination is employed, examples are necessary to achieve a fine-grained discrimination. Therefore the examples in the entries of near-synonym discriminations should not only be authoritative and typical in language use, but also be capable of revealing signs of nuances between near-synonyms.

4. Conclusion

As the cognitive target, the near-synonym series can be regarded as a kind of cognitive semantic frame (CSF), which is composed of category, category members, attributes and values, conceptual structures and constraints. In such a frame, the near-synonyms represent the categorization and sub-categorization of human experience. The activation of one category member always interacts with the activation of relevant knowledge and background. The CSF-based notion will solve many controversial problems in the compilation of entries of near-synonyms in the English learner’s dictionaries, including the establishment of near-synonym entries, the selection of the dominant, the salience of semantic attributes and values, the sequence of
near-synonyms within the entry, the granularity of discrimination, the typicality of illustrations and the multi-dimensions of discrimination of near-synonyms.

References


Anxing Li
Shanghai University of International Business and Economics
anxingli@outlook.com

Abstract

The publication of the first volume of The Chinese-English Dictionary (Unabridged) was praised as “a dictionary that might generate earth-shattering changes”. This paper discusses two aspects of its theoretical contributions which made the dictionary innovative. Firstly, its design aiming at making the dictionary not only prescriptive but also descriptive makes the dictionary fuller in the inclusion of word entries. Since most Chinese words are loaded with not only cultural value but also linguistic value, it is important that the compilers creatively explore words and characters analytically in their particular cultural context and watch and detect neologisms. In addition, some translations of English words and phrases are listed as headwords in this dictionary. Secondly, by designing the dictionary to be a mine of information for both consultation and reading on the part of different users, the dictionary compilers try to make the dictionary fulfill its remedial function of satisfying their users. To achieve the aim, the dictionary compilers try to provide translational equivalents which focus on message not only gained from the discourse world founded on real external circumstances but also gained from the text world which requires memory and imagination of users for its understanding. In addition, the dictionary compilers take into considerations of how to describe meanings in context by translation and how meanings can be implicated rather than explicitly stated by translation. Also, to make this dictionary a mine of information, this dictionary is also designed to include Chinese dialectal expressions, slangy expressions with their English counterparts to reflect their respective history and culture. The implications of the dictionary for the making of the future Chinese-English dictionaries have also been discussed.

Key words: descriptive, prescriptive, remedial, satisfy users, meanings in context, a mine of information, consultation, reading
1. Introduction

The publication of *The Chinese-English Dictionary (Unabridged)* Vol. 1 (The Dictionary, Lu Gusun 2015) has attracted wide attention of the media from China and the rest of the world. It was praised as “a dictionary that might generate earth-shattering changes” by the Wenhui Scholars column of Wenhui Daily. More noteworthy achievements are its contributions to the theory and practice of the making of bilingual dictionaries in China. This paper attempts to describe its theoretical contributions in two aspects along with language exploration on the part of dictionary compilers and discuss its implications for the future Chinese-English dictionaries.

2. Language Exploration and the Descriptive Function of *The Chinese-English Dictionary*

As is known, the compilation of a bilingual dictionary, first and foremost, starts with the decision of choosing a list of words that can be used as headwords for the dictionary. Then the dictionary compilers describe them by translation. In this process, the dictionary compilers compare the meanings of the words in order to find out translational equivalents between lexical units in the two languages. With the availability of diverse and up-to-date corpus evidence, a descriptive approach to dictionary-making based on observed facts becomes possible. This advantage is obviously reflected in the decision of a Chinese-English dictionary compiler’s choice of headwords. However, finding out a real lexical unit in the target language for a source language item is a very difficult process because many Chinese lexical units are loaded not only with linguistic value but also with cultural value and need to be defined in their cultural context. As far as compilation of a Chinese-English dictionary is concerned, traditional methodology, as was also mentioned by my mentor Professor Lu Gusun (2000), is based on an acknowledged authoritative source-language mono-lingual dictionary if one exists for choosing headwords and hopefully a filing cabinet of citation files for reference. Although a source-language mono-lingual dictionary is a sine qua non for doing the work, it is sometimes not reliable. One of the problems facing The Dictionary compilers is that there is always a mismatch between the way the lexis behaved in the discourse world inhabited by an author and a reader and the way it was described in the existing monolingual SL dictionaries. For example, in the venerable The Chinese Dictionary (paperback edition) (Luo2000), we find the following definitions for the headword “永”:

永 yǒng

① 指水流长。《诗·周南·汉广》: “江之永矣。” 亦泛指长。《书·高宗肜日》: “降年
In this dictionary, the headword “永” was presented as sequences of numbered senses with each having one illustrated example taken from the original books. However, a careful scrutiny of the senses reveals that the third sense “犹度过、消磨” (to pass the time; to while away time) was incorrect — the sense recorded in the dictionary doesn’t mean the same with the sense found in actual use. According to poems concerning life of the nobility in XiaoYa (Minor Odes of the Kingdom) in ShiJing(Book of Odes), “絷之维之，以永今朝” presents a situation in which the host and hostess had the guests’ white pony tethered and tied up outside so that its master might prolong his visit during the day. Similarly, in the line “何不曰鼓瑟？且以喜乐，且以永日” (why not play daily your lute as part of your life to prolong the joy?) and the line “客来斯舟，可以御风，可以永夕” (visitors who step aboard this ship could find shelter against wind and stay longer for the night hours), “永日” and “永夕” don’t mean “to pass the time or to while away time”. As a matter of fact, “永” in this context means “to prolong a period of time”. “永日” means “to spend a prolonged period of daytime” and “永夕” means “to spend a prolonged period of night hours”. Since these Chinese lexical units are loaded with cultural value, we need to find out the real meaning of a lexical unit by exploring its actual use in its cultural context. It is common sense that a dictionary should contain the correct interpretation of signifiers. However, room for creative work remains ample (Lu Gusun 2000).

As mentioned above, to write a Chinese-English dictionary, the dictionary compilers have to rely on authoritative source-language mono-lingual dictionaries for choosing headwords as a rule. However, language change is constant. When we look at the change of the Chinese language in recent years, we will find that the number of new words is much larger than any authoritative and comprehensive Chinese dictionary would include. In the words of Professor Lu Gusun (2000), “Language is inexhaustible and a-changing; dictionaries are never complete or up-to-date in the true sense of the word”. Therefore, the dictionary compilers’ perspective on language and...
their lexicographical practice coupled with language exploration are also indispensable for the compilation of a Chinese-English dictionary. By “dictionary compilers’ perspective on language”, it is meant that “creative work on the part of a bilingual lexicographer is to be carried out in connection with watching and detecting neologisms in the source language.” (Lu Gusun 2000). With a new perspective on language and with the availability of the corpus evidence, the creative dictionary compilers find it no longer as difficult and time-consuming as in the past to include those new words that are not found in traditional Chinese dictionaries of any size. These include Chinese characters like “蛮” (literally “brutal”) and “泳” (to swim), for example, which are very active and productive in present-day Chinese usage as are found in such compounds as “蛮辩” (brutal debate), “蛮肉” (brutal-looking muscle), “泳龄” (length of time as a swimmer), “泳友” (swimming buddy), “泳迷” (swimming fan). In the words of Professor Lu Gusun (2000), “Albeit to a lesser degree, being firsthand in culling examples involves a discriminative eye and a creative effort on the part of a bilingual lexicographer”.

In sum, language exploration and the dictionary compilers’ perspective on language help to make a Chinese-English dictionary fuller and thus fulfill its descriptive function in meeting the users’ needs. In addition, for a Chinese-English dictionary compiler, a much wider range of reference resources is also an absolute necessity. For example, in writing entries for The Dictionary, the dictionary compilers carefully selected numerous undocumented words from People.cn online (人民网) and included them in the inventory of entries. In addition, many Chinese expressions were taken from a corpus of Chinese novels and translated versions of English novels to realize the fullest descriptive function of The Dictionary.

3. Language Exploration and the Remedial Function of Satisfying Dictionary Users

The goal of bilingual lexicography is to offer interlingual lexical equivalents that facilitate translation on the part of users. As Ladislav Zgusta put it (1984), “the dictionary should offer not explanatory paraphrase or definitions, but real lexical units of the target language which, when inserted into the context, produce a smooth translation.” However, in compiling a Chinese-English dictionary, lexicographers have to take into considerations of one major problem that confronts them: language can be used differently for different purposes. When they translate, they choose. To be specific, when translators translate they choose expressions that best serve their immediate purposes because they know that how one says something can be as important, sometimes more important, than what they say. For example, describing meaning of “与鲨鱼共泳” by the translation of “to swim with sharks” is of limited value to
experienced translators who need more useful translation in the context of “to swim in a shark-infested sea” to arouse in the minds of readers a similar sense of danger as they might think of the phrase “与狼共舞” (to dance with wolves). This requires that when providing translational equivalents, bilingual lexicographers give thought to such issues as how to describe meanings in context by translation and how meanings can be implicated rather than explicitly stated by translation.

The compilers of The Dictionary are radically innovative in this respect because they designed the dictionary with users’ different needs in mind. To achieve this, they attempt to give full play of the communicative function of language based on the fact that word meaning of a source language item cannot always been rendered precisely by a single translational equivalent or multiple translational equivalents without giving considerations to their purposes. For example, the headword “害人之心不可有，防人之心不可无” has been rendered into the translation of “one should not have the intention to harm others, nor should one forget to guard against possible harm from others; every Caesar has his Brutus”. The compilers provide the first translational equivalent on the assumption that ‘communicating meanings in particular contexts’ is seen by most people as the primary function of the language (Thompson 2004) and the compilers provide the second equivalent “every Caesar has his Brutus “on the assumption that “experienced translators will find pragmatic techniques for dealing with problems of literary translation” (Landers 2001).

Most people nowadays are aware of the importance of a corpus and working with a corpus is rapidly becoming indispensable in all branches of lexicography. As stated by Professor Lu Gusun (2000), “Considering state-of-the-art achievements in both lexicography and computer technology, it is truism to say that real revolution has taken place in the development of a citation file, which I call the infrastructure of a dictionary”. However, the difficulty of writing entries for The Dictionary lies in the fact that in most cases dictionary compilers have to approach the data without preconceptions about meanings or cultural contexts of words in the source language. In this case, a dictionary compiler has to be “an untiring word-muncher … a hair-splitting meaning explicator, a fault-finding proof-reader” (Lu 2007). This requires that The Dictionary compilers be both creative and hardworking. Though the work on the part of The Dictionary compilers is challenging, – we strongly believe –dictionary users will find many illuminating and representative examples with unique and fresh translations that illuminate the spirit of The Dictionary. The example of translations which require a better understanding of the words of the source language include “诗言志，歌咏言” (poetry expresses aspirations in words, and singing makes poetry remembered in perpetuity). In this line, “志” means “aspirations” rather than “sentiments”.

388
In other cases, the lexicographical treatment of lexical items taken from translated texts also proves to be a thorny problem. On the one hand, the very nature of the bilingual dictionary dictates that translational equivalents provided by dictionary compilers be “real lexical units of the target language” that the dictionary users will find in their readings and insert into their translation. On the other hand, dictionary compilers have to rack their brains to come up with a best or most creative translational equivalent that is capable of effecting the optimal transfer of the meaning of the words of the source language. In both cases, to make translational equivalents more useful to dictionary users, a dictionary compiler has to prove to be “a bibliophile, not only keen on reading with gusto but also capable of writing with brio” (Lu2007). As listed below, users will find in The Dictionary many examples of translations where the compilers have added additional information for the ultimate benefit of the users.

安 ān

……2 to set at ease: ～下心来 set one’s mind at ease ……11<lit> where, how [as a rhetorical question particle] 其故何在 wherein lies the cause | *燕雀知鸿鹄之志 how can a sparrow know the soaring ambitions of a swan goose? or how can the mediocre know the minds of great men? 12……

涌流 yōngliúto

flow rapidly: <fig>让创新源泉不断～ stay invigorated with the continuous outflow of creativity

宾至如归 bīnzhírúguī

<idiom> (guests arrive as if returning home) a home away from home [descriptive of the warm reception, convivial atmosphere of a hotel, guesthouse, etc]

Knowledgeable persons tend to look at a dictionary as “a mine of information, an encyclopedia in disguise” (qtd. in LuGusun2007). To fulfil the remedial function of satisfying users, The Dictionary compilers include in the dictionary a great number of the lexical items translated from English in which cultural knowledge is embedded. “失礼的言行” (a faux pas) and “勇气是压力之下的优雅” (courage is grace under pressure) are just two examples. In addition, The Dictionary is designed to include Chinese dialectal expressions, slangy expressions with their English counterparts that reflect the cultural values of a society at a particular time. For instance,
阿家翁 ā gūwěng

abbr 阿家阿翁<dial>1 (woman’s) parent-in-law: 不痴不聋，作不～pretending not to see or hear family brawls is part of being a parent-in-law 2 head of a household

屌丝 diǎosī

<slang>1 diaosi, loser, nonperson, underdog [usu. self-depreciatory] ……2……

In the entries like these two examples, dictionaries users will find English kinship terms and many other terms in the form of translational equivalents. Traditionally, dictionaries are designed to be consulted to solve semantic problems. In other words, dictionary users consult a dictionary to look for the answers. However, for both Chinese and English language learners, The Dictionary can also be read to acquire Chinese culture as Professor Lu Gusun mentioned in his preface to this dictionary (2015). For example, in The Dictionary readers will find the following entry:

宥坐之器 yòuzuòzhīqìalso

宥坐器 /yòuzuòqì/ <lit> tilted container for holding water placed on the right hand side of an emperor’s throne [served as a reminder that the last drop makes the cup run over]

The usefulness of this entry lies in the fact that by reading this entry advanced learners of English will find that in ancient China the Chinese people also had the notion of ‘nothing in excess’ which is equivalent to that inscribed on the temple of Apollo at Delphi, 6th century BCE. Yet, few readers perhaps know this fact. Presenting in an optimal way the information in a Chinese-English dictionary required by users is a complex task. The value of The Dictionary lies in the creativity of its compilers’ efforts to describe meanings of headwords and illustrated examples by translation to fulfill the remedial function of satisfying users. For example, equivalents offered in the following entries will meet different demands of different users. These equivalents are offered for both consultation and reading on the part of users.

功败垂成 gōngbàichuíchéng

<idiom>to fail when victory seems within reach; to fail in the eleventh hour; to fail on the verge of success; there is yet a possible slip twixt the cup and the lip

聰明一世，糊涂一时 cōngmíngyīshì，hútúyīshì
also 聪明一世，懵懂一时 / mèngdòng - / <familiar> to be clever all one’s life but stupid this once; smart as a rule, but this time a fool; a wise man is not free from momentary stupidities; the wisest are not always wise; no man is wise at all times; every man has a fool in his sleeve; even Homer sometimes nods

4. The Significance of The Dictionary and Its Implications

From what has been said so far, it is clear, as was already stated in previous sections that the very nature of Chinese words or characters demands that a Chinese-English dictionary compiler be a language explorer since many Chinese words or characters have not only cultural value but also linguistic value. The very nature of a Chinese-English dictionary demands that compilers provide real lexical units of the target language which, when inserted into the context, produce a smooth translation. The very nature of corpus decides that lexicographers be creative writers rather than merely compilers since the corpus, as a tool, is only instrumental (Lu Gusun 2007; 2000). The publication of The Dictionary has generated a change in the methodology of including headwords and offering translational equivalents on the part of dictionary compilers. Its design with its methodology aims at making the dictionary fuller and more useful to users, which fulfils not only its prescriptive function, but also the descriptive function and the remedial function of satisfying users. In preparing the word list for the inventory of headwords, lexicographers should watch and detect neologisms in the source language. In offering translational equivalents, the dictionary compilers should have a purpose in mind and give thought to meanings in context and describe them by translation to meet the needs of learners of English.

The making of The Dictionary has significant implications for future lexicographers. Compilation of a successful Chinese-English dictionary still needs a lexicographer with a flair for language. Language exploration on the part of lexicographers is indispensable. In addition, to make a dictionary more useful to readers, compilers should also establish the dictionary on firm theoretical and methodological grounds. Since language can be used differently for different purposes, the dictionary compilers should provide translational equivalents which put focus on message not only gained from the discourse world (Simpson 2004; Werth 1999) founded on real external circumstances but also gained from the text world (Simpson 2004; Werth 1999) which requires memory and imagination for its understanding, rather than direct perception.
References

Landers, Clifford E. 2008. Literary Translation: A Practical Guide. 172; 7; 11; ix Shanghai: Shanghai Foreign Language Education Press.
A Discussion on New-Word Selection for Chinese-English Dictionaries

Yongwei Gao
Fudan University
ywgao@fudan.edu.cn

Abstract

As Chinese-English dictionaries (e.g. *New Century Chinese-English Dictionary* and *The Chinese-English Dictionary*) grow in size, they tend to include as many new words as possible so that they can make wide coverage as their biggest selling point. However, the inclusion of new-word entries in such dictionaries turns out to be rather problematic, which can be exemplified by problems such as sin of omission, the imbalanced treatment of similar words, and the inclusion of topical words or proprietary names. In the final analysis, these problems can be more or less attributed to loose or even haphazard criteria adopted by compilers of Chinese-English dictionaries who usually base their choices of new words on random readings rather than vast corpus data. This paper intends to address the problems that exist in five major Chinese-English dictionaries now available. Meanwhile, the author will put forward several criteria for excluding words that do not deserve entry status, and these criteria will encompass principles of excluding political words that are topical in nature, compound words of doubtful headword status, purely online words or expressions, less frequently used abbreviations, etc.

**Keywords:** Chinese-English dictionaries, new words, exclusion criteria

As Chinese-English (hereinafter abbreviated into C-E) lexicography has undergone more than two hundred years of development, the layout of dictionary entries has become more or less entrenched. As a result, innovative lexicographic practices are something of a rarity. The latest innovative feature one can think of is related to the addition of part-of-speech labels for headwords, which, however, is still not widely adopted in monolingual Chinese and C-E dictionaries. The current scene of C-E lexicography is dominated by four major Chinese-English dictionaries, namely *New Century Chinese-English Dictionary* (NCCED, 新世纪汉英大词典), *New Age Chinese-English Dictionary* (NACED, 新时代汉英大词典), *The Chinese-English Dictionary* (TCED, 汉英大词典), and *A Chinese-English Dictionary* (ACED, 汉英词典). These dictionaries are all large-sized dictionaries with coverage of over 100,000
entries, and one of them even boasts 240,000 entries. Without exception all of them have undergone revisions, and their latest editions were published after 2010. As a rule, as they are revised, they grow in size. Surprisingly enough, these bilingual dictionaries usually record Chinese neologisms at a faster rate than their monolingual counterparts. Whenever a new edition of a C-E dictionary is published, it will inevitably advertise in its blurb those new words or meanings it has included. TCED, for instance, added 15,000 neologisms into the third edition, among which words such as 物流 (logistics), 房奴 (mortgage slave), 躲猫猫 (hide-and-seek), 形象大使 (promotion ambassador), 转点 (turning point), 愿景 (perspective view), 婚奴 (wedding slave), and 宅男 (Otaku) were cited as examples in the dictionary’s preface. When The Chinese-English Dictionary Unabridged (CEDU, 中华汉英大词典), a new kid on the lexicographical block, released its first part in 2015, many new words it recorded made headlines, such as 爆表 (to exceed the upper limit), 代驾 (rent-a-driver), 点赞 (to push the “like” button), 屌丝 (loser; underdog), 给力 (boosting; stimulating), and 空窗期 (vacuum).

However, the selection and inclusion of new words in the above-mentioned C-E dictionaries is rather problematic, and the main problems therein include sin of omission, the inclusion of words that should not be listed as headwords, the imbalanced treatment of new words of the same category, the inclusion of topical, short-lived lexical items, or even proprietary names, the use of new words as examples, etc. This paper will first address these problems, and then discuss strategies for excluding words that should not be covered, such as political words that are topical in nature, compound words of doubtful headword status, purely online words or expressions, and infrequently used abbreviations.

1. Existing problems in the coverage of new words

Problems abound when it comes to the inclusion of new-word entries in the current batch of C-E dictionaries, the most conspicuous of which is sin of omission. Failure to present a panoramic picture of the latest lexical changes is an omnipresent problem in dictionary-making, and this kind of problem is especially conspicuous in C-E lexicography. After the first edition of NACED came out in 2000, Chen Zhongcheng (2001: 46-47), a scholar noted for his incisive criticisms of bilingual dictionaries, wrote a review of the dictionary and listed dozens of new words that were not recorded, such as 风险投资 (venture investment), 熟食店 (delicatessen), 动作片 (actioner), 家庭暴力 (domestic violence), 脱口秀 (talk show), and 工作狂 (workaholic). Jin Qibin
(2009: 75) mentioned scores of Chinese neologisms that failed to grace the pages when he reviewed the first edition of NCCED, and such new words include 街舞 (street dance), 网络公司 (dot-com), 恶意收购 (hostile takeover), 过渡房 (starter home), 电子宠物 (digital pet), and 情感符 (emoticon). The second edition of NCCED was published late last year, but it is paled by comparison with the number of new words recorded by CEDU which was published one year earlier. Those new-word entries that are absent from NCCED include 腹黑 (scheming; calculating), 低头族 (phubbers), 呆萌 (silly and cute; adorkable), 东东 (thing, stuff), 建群 (to set up a group), etc. CEDU, though much lauded for its extensive coverage of Chinese neologisms, is not “sin-free” as its first volume failed to record a dozen new words that are now enjoying great currency, such as 畅聊 (chat freely), 打脸 (be embarrassed or ridiculed), 采购经理指数 (purchasing managers index, PMI), 电臀舞 (twerk), 公益旅行 (voluntourism), and 剧透 (spoiler alert). Failure to include the latest lexical items can be attributed to two reasons: first, the compilers of C-E dictionaries are not fully acquainted with Chinese neologisms which are now cropping up at an astonishing rate, and the lack of large-size Chinese corpora and the unavailability of inclusive Chinese new-word dictionaries are also to blame; second, as C-E dictionary-makers are usually unfamiliar with the new-word scene in the English language, they often fail to record the newly-minted translations of English neologisms, many of which soon become established lexical items in the Chinese language. A case in point is 家庭暴力, a word failed to be included in NACED. The Supplement to The English-Chinese Dictionary, published one year earlier than NACED, recorded the use of domestic violence and provided it with 家庭暴力 as its Chinese equivalent. If the compilers of NACED had the opportunity to browse the said supplement, this Chinese new word would have been included. This kind of “reverse translation” has been advocated by some practical lexicographers. Zhao Gang, a member of the editorial team of NCCED, discussed how to use “reverse translation” to better the translations of headwords and illustrative examples in C-E lexicography (2004: 63-68).

Another problem that stands out is related to the inclusion of words whose use has not been established in the Chinese language. These words can be roughly classified into two categories: words translated from English neologisms and newly appeared words whose use may not last. Because of the importance of the English language as a lingua franca, many new ideas, concepts, technical innovations, etc. first originate in the English-speaking countries, and then they are propagated throughout the world.
The Chinese language has been on the receiving end of linguistic borrowings for decades, and as a result, some bilingual dictionary-makers are under the illusion that whatever words (neologisms in particular) that are translated from English are very likely to catch on. That is why words such as 非异性恋者 (LGBT), 究责风暴 (blamestorming), 空气末日 (airpocalypse; airmageddon), and awkwardly sounding 伙伴加对手 (frenemy; froe; frienemy) have made their way into CEDU. Other C-E dictionaries are not immune from this problem. NCCED, for example, included dozens of new words whose status as separate entries was rather doubtful, such as 网络分析 (network analyzer), 网络出版 (publishing on internet), 网络大学 (e-college; e-university), and 电子探伤器 (electronic crack detector). TCED, famous for its wide coverage of technical terms, included many new words that contain at least four characters, many of which cannot be regarded as headwords, e.g. 网上交易平台 (online trading platform), 网上求职 (hunting for a job on Internet), 网络化国家 (network nation), 网络私教 (private network services), 电子环境污染 (electronic environmental pollution), and 电子数据处理 (electronic data processing). As a matter of fact, it would make much sense if the above-mentioned headwords are treated as illustrative examples. As for the second category of words, their number is astonishingly large as they are prevalent in dictionaries of Chinese neologisms. The 2015 edition of The Commercial Press’s neologism dictionary included 加班楼 which literally means “an office building in which many employees have to work overtime”. As words like this are coined mainly for jocular or ironical purposes, they always enjoy fifteen minutes of fame and fail to achieve wider currency. Similarly, new words such as 帮带费 (money given to one’s parents for taking care of his or her kids), 拆发户 (a household made rich because of compensations for resettlement), 创时代 (an era for innovations), 代包手 (people who make money through distributing red packets for others on WeChat), and 雾霾费 (fees paid by polluting manufacturers) should not be considered candidates for inclusion.

The imbalanced treatment of new words of the same category also plagues C-E lexicography. The problem of the inclusion of one new word and the exclusion of a
related one can be said to be a perennial one, as is attested by its frequent mention in dictionary reviews. Zeng Dongjing (2004: 93), though rather positive about NCCED’s coverage of new words, cited several pairs of words that were not all included in the dictionary, namely 反腐 and 防腐 (absent), 高端 and 低端 (absent), 大哥大 and 大姐大 (absent), 索贿 and 收贿 (absent). Jin (2009: 76) listed over a dozen words that were not included while their related terms were recorded in NCCED, such as 任务栏 (task bar), 借记卡 (debit card), 小盘股 (small cap), 蝴蝶效应 (butterfly effect), 卫星电话 (satellite phone), and 机会成本 (opportunity cost). Omissions of this kind can be found in all the other C-E dictionaries: ACED recorded the verb and adjective uses of 在线 (be online; online), but failed to include its antonym 离线 (be offline; offline); TCED included several compound words containing the suffix-like族 (such as “卧槽族 jog-hugging clan” and “赖校族 campus dwellers”), but failed to record the use of another popularly used word that was formed in the same way, namely 蚁族 (lit. ant tribe); NACED included several words containing 领 (e.g. “白领 golden-collar”, “蓝领 blue-collar”, “粉领 pink-collar”, and “白领 white-collar”), but failed to include a neologism formed in the same way—绿领 (green-collar); CEDU included the translations of several English neologisms formed with the suffix -ware (e.g. “恶意软件 malware”), but another frequently used neologism 间谍软件 (spyware) was conspicuously absent.

Sometimes C-E dictionaries may have included words of topical nature or even proprietary names. Unlike new-word dictionaries, general C-E dictionaries should adopt strict criteria as regards the inclusion of short-lived lexical items and proper names. Due to the lack of uniform standards or rigorous screening process, topical words or proprietary names are sometimes “smuggled” into C-E dictionaries. TCED’s inclusion of the translations of English new words such as metrosexual and ubersexual is indicative of this problem. The editors listed 都市美型男 and 都市粗犷男 as headwords, but these two words were transient uses that have already been replaced by 都市型男 and 阳刚男子. When it comes to the coverage of proper names, problems also abound. NACED, for instance, included over dozens of proprietary names such as 阿迪达斯 (Adidas), 耐克 (Nike), 微软公司 (Microsoft Corporation), 麦当劳
(McDonald’s; a hamburger), 肯德基 (Kentucky Fried Chicken, or KFC), and 沃尔沃集团 (Volvo Group). However, the absence of the IT giant Google (谷歌) and the retailer giant Walmart (沃尔玛) has revealed the haphazard nature of the dictionary’s inclusion criteria. The presence of lesser-known proper names such as 华侨日报 (Wah Kiu Yat Po, a Hong Kong newspaper), 华润集团 (China Resources Group, a mainland-funded corporation based in Hong Kong), 中国国际旅行社 (China International Travel Services), and 中国人民大学 (the Renmin University of China) throws more doubt on its criteria.

C-E dictionaries treat some new words as illustrative examples now and then, the effect of which is doubtful. Liu et al (2016) discussed the pragmatic way NCCED adopted in the treatment of neologisms and singled out 碳 (carbon) and its compounds for in-depth analysis. Although the authors (including editor-in-chief of NCCED) stressed the use of collocational ability as the basis to distinguish illustrative examples and headwords, how the compounds of 碳 were treated in the dictionary showed otherwise. NCCED put 碳足迹 (carbon footprint), 碳补偿 (carbon offset; carbon offsetting), and 碳减排 (carbon emission reduction) in the section for illustrative examples. As the headword is 碳 (carbon), these three examples by no means appropriate as the initial character does not refer to carbon but carbon dioxide. Therefore, separate headword status should be granted to these new words.

2. Strategies in including new words

In the recent decades, the Chinese vocabulary has experienced exponential growth as lexical creations are cropping up at an unprecedented rate. The publication of over a dozen dictionaries of Chinese neologisms is reflective of the rapid growth of the Chinese vocabulary. The Commercial Press even puts out a dictionary of neologisms every year, chronicling the latest additions to the language. The 2015 edition of the dictionary, for example, included 471 neologisms that were culled from a corpus of 1.2 billion words. Yang Xuming et al (2009: 97-98) identified six types of Chinese neologisms, namely new words, new meanings of existing words, dialectal or regional words that are enjoying wider currency, loanwords of all kinds, English initialisms and
acronyms or lexical hybrids that contain both letters and Chinese characters, and finally numbers used as words. As this paper excludes the discussion of new meanings for existing words, we will mainly deal with the remaining five types.

When it comes to inclusion of new entries in C-E dictionaries, several dictionary editors have elaborated on their own principles. Pan Shaozhong, editor of NACED, advocated the policy of “seeking newness” that pays equal attention to the inclusion of new words, new meanings, and new encyclopedic entries (2003: 86-93). The late Prof. Lu Gusun, editor-in-chief of CEDU, put forth the concept of “descriptivism with a grain of salt” in the preface to his dictionary which refers to the practice of excluding words that are too slangy, obscene, or even extremely unpleasant (2015: 1). Instead of proposing a set of criteria that guides the selection of new words, the author intends to put forward a series of exclusion principles that are aimed at weeding out new words unsuitable for inclusion in dictionaries such as CEDU.

The first exclusion principle involves political words that are topical in nature, topical words, and buzzwords. Ideally, a dictionary should be apolitical, showing no political or ideological preference whatsoever; but in reality, a dictionary is more or less influenced by the political era in which it is compiled. Words or expressions that are in essence political slogans or encapsulates government policies should be avoided by C-E dictionary-makers as many of such words or expressions are topical ones which entail revision after a short while. TCED’s inclusion of 八荣八耻 (eight honors and eight disgraces) is a case a point. While all the other C-E dictionaries shunned the political term, TCED devoted eighteen lines to its explanation. The term, first brought forth by China’s former president Hu Jintao in 2006, was bandied about for a while, but now it is seldom heard of. As lexicographical misfits, terms like this would have made a perfect entry in an encyclopedia. Therefore, new political terms of the same nature, such as 一带一路 (The Belt and Road), 工匠精神 (artisanal spirit), 文化自信 (cultural confidence), and 两学一做 (Two Studies, One Action) should not be excluded as well. The same can be said of topical terms and buzzwords such as 阅兵蓝 (lit. parade blue), 小目标 (a small target), 狗带 (go die), 校园霸凌 (school bullying), 小短假 (mini-vacation), and 二孩效应 (two-child effect).

The second principle is to exclude compound words whose status as headwords is rather doubtful. First, dictionary-makers should enforce stricter criteria when including compound words or derivatives formed with affix-like characters. Zhang Xiaoping (2008: 119-148) discussed both words formed with affix-like characters (such as 性 and 化) and words of the same ending or beginning characters (such as 吧, 超, and 防). As there is no consensus among Chinese linguists regarding the distinction
between Chinese affixes and ordinary characters, the author here will simply label such characters “affix-like”. Nowadays, one of the most productive affix-like characters is 族 which literally means “clan or race”. When used in combination with other words, it refers to a group of people who share the same interest or like to do the same thing. New Chinese words with 族 as their ending may have amounted to over a hundred, such as 爱邦族, 爱券族, 闪玩族, 恐会族, 有房族, and 钟摆族, but C-E dictionaries should not spare space for words like these. In the same vein, new words ending with 热 (craze), 控 (complex), 男 (male), and 迷 (fan) or beginning with 裸 (naked), 名 (famous), 陪 (accompany), and 云 (cloud) should be eschewed as well. Second, caution should be exercised when dealing with words containing four or more characters as most of such words would better serve as illustrative examples than headwords. There is a profusion of such examples, e.g. 促销大战 (sales campaign), 审美疲劳 (aesthetic fatigue), 人口爆炸 (population explosion), 销售策略 (marketing strategy), 绿色消费 (green consumption) and 生态危机 (ecological crisis). Multi-character words that were word-for-word translations of English neologisms should be shunned as well for the same reason, such as 阿尔法男 (alpha male), 病态建筑物综合征 (sick building syndrome), 次贷危机 (sub-prime mortgage crisis), and 低端市场 (low-end market).

The third principle involves the exclusion of some online words or expressions that are most used in online contexts rather than in daily life. The popularization of the Internet has spawned hundreds or even thousands of online expressions, some of which may have gone mainstream or been recorded by major C-E dictionaries, such as 菜鸟 (a novice), 腹黑 (scheming; calculating), 灌水 (to publish numerous posts on the Internet), 潜水 (to lurk), 给力 (boosting; stimulating; cool), and 点赞 (to click the “like” button; to praise). However, many popular online expressions are rather vulgar, and their prevalence may ultimately sully the purity of the language. In consequence, online terms such as 逼格 (pretentiousness), 逗比 (funny), 叫兽 (a homophone of 教授 [professor]), 你妹 (lit. your sister, same as frigging), 蛋疼 (lit. it pains the
balls), and 碧池 (the transliteration of bitch) should be kept at bay. Another type of online words or expressions that should be avoided is the homophones of existing words that are created mainly for jocular purposes. Words such as 板斧 (版副, meaning “assistant moderator”), 斑竹 (班主, meaning “moderator”), 大虾 (大侠, meaning “knowbie”), 小公举 (小公主, meaning “little princess”), 炫腹 (炫富, meaning “to show off one’s wealth”), and 言值 (颜值, meaning “glamor or attractiveness”), though quite frequently used online, should not deserve a place in serious C-E dictionaries.

The fourth principle concerns the exclusion of abbreviations that are used not as often as what they stand for. The tendency to shorten words or expressions, whether in real life or in the virtual world, is growing, which results in the generation of thousands of abbreviations, initialisms, and acronyms. In the Chinese language, with the absence of initialisms and acronyms because of its nature of being a non-alphabetic language, abbreviations proliferate, and they usually come in A+C or A+D combinations in four-character words, such as 顶豪 (顶级豪宅), 白奴 (白领奴隶), 财险 (财产保险), 车补 (乘车补贴), 妇运 (妇女运动), and 电阅 (电子阅览). However, many of these abbreviations are not frequently used, and therefore should not be included in C-E dictionaries.

The fifth principle is to exclude transliterations that were created for jocular purposes or that seems to be superfluous. 销品茂 is a case in point. As the transliteration of shopping mall, it is not widely used, which might be attributed to the fact that there has already existed a Chinese equivalent for the English term, namely 购物中心. Therefore, transliteration like 销品茂 and the more absurdly used 摩尔 (said to be the transliteration of mall) should not be included in C-E dictionaries. Similarly, transliterations such as 拜客 (bike), 奥菲斯 (office), 崩克 (punk), 稻糠 母 (dot-com), 甫士 (pose), and 卡曲 (car coat) should not make the cut.

The sixth principle is to exclude new words or expressions that contain or are related to proper nouns. The past decade has witnessed the creation of many terms related to proper nouns partly because of people’s preoccupation with celebrities and the show business. 葛优躺, for example, became a household word last year because of its association of the reclining posture of the protagonist played by Ge You, a famous movie star, more than twenty years ago. Although the GIF of the posture has
gone viral, the word itself should not be granted a place in dictionaries. The same can be said of words like 纲丝 (fans of Guo Degang, a comedian), 阿迷 (fans of the Argentine soccer team), 凡客体 (the style of the advertisement copy of VANCL), 淘宝体 (the style in which a seller describes his wares on the online shopping site Taobao), 偷菜族 (people who are addicted to stealing vegetables on virtual farms), etc.

The seventh principle involves the exclusion of words coined through analogy for humorous purposes or analogical formations that do not enjoy much currency. Tiger mother (虎妈), a word popularized by the 2011 memoir Battle Hymn of the Tiger Mother by Amy Chua, has so far spawned several copycat expressions, such as 狼爸 (wolf father), 猫爸 (cat father), 虎爸 (tiger father), and 獅妈 (lion mother). As these new words are usually characterized by ephemerality, they do not deserve a place in C-E dictionaries. Similarly, other analogical words, like 泥饭碗 (mud rice bowl, patterned on “iron rice bowl”), 纸饭碗 (paper rice bowl, patterned on “iron rice bowl”), 虎市 (tiger market, patterned on “bull/bear market”), 鹿市 (deer market, patterned on “bull/bear market”), and 饥怒 (hunger rage, patterned on “road rage”) should be given the same silent treatment.

The eighth principle is to exclude overly slangy words or expressions that are used to denote sex or may cause offense. Prof. Lu Gusun once wrote that “I am an advocate of banishing dirty words from dictionaries. As a dictionary-maker, one should not cross the filthy line, and adopt an extremely descriptive approach as Urban Dictionary does, as a way of showing the moral code of a society and reflecting a lexicographer’s respect of his readers” (Lu 2014). Therefore, it will be better if words such as “娘炮” (a sissy), “操” (to fuck), “打炮” (to have sex), “大姨妈” (period), and 鸭子 (a male prostitute) are barred from inclusion in C-E dictionaries.

The ninth principle is related to technical terms whose number in any language may exceed that of ordinary words and expressions. An all-inclusive policy regarding the inclusion of technical terms always backfires, as is evidenced by Chinese users’ preference of NCCED and NACED over TCED which may have distanced its users with its wall-to-wall coverage of technical terms of all kinds. Therefore, an exclusive policy is in order. Then what kind of technical terms should be kept out? As is the
practice of most C-E dictionaries and their E-C counterparts, common or frequently used technical terms always make the cut as they are often used in the mass media. Thus overly technical terms and scientific jargon should be excluded, such as 熊果酸 (ursolic acid), 磁穿孔 (magnetoporation), 绒毛膜羊膜炎 (chorioamnionitis), 天使综合征 (Angelman syndrome), 微生物组 (microbiome), and 系统发育基因组学 (phylogenomics).

3. Conclusion

The inclusion of new words is one of the most important tasks for a dictionary compiler or reviser. Only when a set of strict criteria is established and then fulfilled can the dictionary better reflect the latest lexical changes in a language. The compilers of the second part of CEDU which is expected to be out next year, will not only adopt the above-mentioned exclusion principles, but also make full use of their expertise in English neologisms whose translations will be made as possible candidates for the dictionary. Such efforts will no doubt enhance the dictionary’s “lookupability”, an attribute which Prof. Lu Gusun always championed. In the meantime, the exclusion of temporary or short-lived words or expressions can ensure the durability of the dictionary.

References

30-33.


Strategies Regarding the Translation of Neologisms in *The Chinese-English Dictionary*

Fushou Xiao  
Shanghai University  
fushaw@163.com

Abstract

Along with the advance of modern Chinese society, more and more neologisms have been brought into being and they are destined to find their way into a Chinese-English dictionary. However, translating Chinese neologisms into English is no easy job; it still remains a hard nut to crack for many a lexicographer and merits close attention in academe.

To this end, the present paper endeavors to delve into the neologisms included in Prof. Lu Gusun’s *The Chinese-English Dictionary (Unabridged)* (2015) to explore potential proven strategies which seem to be best suited for the translation of Chinese neologisms into English in a Chinese-English dictionary. As revealed in this study, there are a host of effective strategies that lexicographers might as well consider pursuing in their translation of Chinese neologisms into English, which include purely literal translation, integration of literal translation with annotation, integration of literal translation with free translation, purely free translation, integration of free translation with annotation, integration of free translation with transliteration, and formative imitation. What’s more, each and every strategy is well supported by abundant illustrative examples.

It follows that the findings of the present paper can well serve the purposes of guidance and reference for enhancing the quality of translating Chinese neologisms into English in a Chinese-English dictionary, promoting the cultural exchanges between China and the West, and surmounting such obstacles as the “five drawbacks of mediocre translation” and the “four areas of concern” in MTI education.

**Key words:** Chinese-English dictionary; neologism; translation strategy

1. Introduction

In the world today, science and technology are developing by leaps and bounds, international exchanges are becoming more and more frequent, and our social life is undergoing unprecedented changes. These changes will inevitably find their way into Chinese vocabulary. As a result, more and more neologisms have been brought into existence in all walks of life. Take the following monologue by a young lady for
example.

年龄，属于奔奔族，谈不上什么新新人类了；职业，教师应该属于白领阶层吧，但还真想成为朝九晚五的粉领；兴趣爱好呢，是忠实的果粉，痴迷的“剁手党”一员，典型的草根网民，而且还是月光公主。看我这把年龄，肯定是已婚，不过现在暂时还是丁克家庭；因工作关系，我和丈夫成了半糖夫妻。要问我的经济状况，月薪八千，所以买房子只能按揭，庆幸自己没有成为房奴。

From the above paragraph, we can see that neologisms like “奔奔族”, “新新人类”, “白领”, “朝九晚五”, “粉领”, “果粉”, “剁手党” have indeed posed new problems for linguists and lexicographers alike. These problems include: (1) How to define a neologism? (2) How to classify neologisms? (3) How do neologisms come into being? (4) How to translate neologisms? (5) What kind of neologisms can be included in a dictionary?

With the deepening of China’s reform and opening up, the translation of Chinese neologisms into English has become a most urgent task of the day. For one thing, neologisms, which reflect the new changes taking place in China, also constitute a mirror of the new culture and spiritual outlook of the Chinese nation; for another, the translation of Chinese neologisms enables foreign nationals to keep abreast of the constantly evolving changes in China, promoting China’s new culture to go global.

Up to now, only a few studies have ever been conducted on the translation of Chinese neologisms at home and abroad. Some deal with the classification and composition of neologisms (Fan Fuju, 2014; Liu Yali, 2004), some investigate the origin of neologisms (Zhao Weili, 2014), some explore the strategies or principles for the translation of neologisms (Zhou Hong, 2011; Xu Changhe, 2009), and some look into the translation of neologisms that appear in the newspapers (Zhang Jian, 2001). However, such studies are by and large almost fragmentary, lacking in systematicity and in-depth analysis. It is also sad to note that none has ever conducted any research on the translation of Chinese neologisms into English from a lexicographical perspective.

In view of the above-mentioned, the present paper endeavors to delve into the neologisms included in Prof Lu Gusun’s The Chinese-English Dictionary (Unabridged) (2015) to explore potential proven strategies which seem to be best suited for the translation of Chinese neologisms into English in a Chinese-English dictionary. Hopefully, it will provide new insight into the study of neologisms in the field of lexicography.
2. Strategies of Literal Translation and Its Applications

The mention of literal translation often reminds us of free translation. These are two hotly debated topics in translation studies. Both touch upon the very nature of translation, and it is hard to decide whether one is superior to the other. In the 1940s or 50s, Chinese scholars like Zhu Guangqian and Lin Handa concluded on the basis of their investigations that there exist no differences between literal translation and free translation; in the 1970s or 80s, translators like Xu Yuanchong and Wang Zuoliang argued that literal translation is different from free translation and that the use of literal translation or free translation depends upon the style of the source language.

In the present study, literal translation and free translation are considered as separate concepts. Literal translation, as its name suggests, refers to direct translation or word-for-word translation. It enables a translator to achieve “formal equivalence” or “functional equivalence” in translation. Through literal translation, we can better retain the imagery and national traits of the original text.

2.1 Strategy of Purely Literal Translation

Purely literal translation aims at direct translation based on the literal meaning of the text. In this way, we can find it easier to achieve the formal and ideational equivalence between the source language and the target language. Let’s take the following for example.

Ex. 1: 单亲家庭 single- or lone-parent family: 这是“～”一词使用前很久的事了;那时候还称“破碎的家” this was long before the term “single-parent family” came into use; back then it was “a broken home”

In this case, “单” can be directly translated as “single” or “lone”, “亲” as “parent”, and “家庭” as “family”. Through the above illustrative example, we may get an inkling of how the term “single- or lone-parent family” evolved. In fact, single-parent families have been in existence since ancient times, but the term “single- or lone-parent family” came about in Europe and America during the 1960s and 70s when divorce became much more common and the number of one-parent families increased substantially. It was not until the 1980s that the term gained currency in China. It counts as a relatively new term accompanied by other terms like “单亲爸爸” （single father）, “单亲贵族”（single parent） or “单亲妈妈” （single mother）.

Ex. 2: 获得感 sense of gain; feeling that one has truly benefited: 让人们有～ give
people a sense of gain

The term “获得感” is quite a new concept introduced by President Xi Jinping in 2015 when he said that “making people have a sense of gain is a decisive rule to measure the gold content of any reform”. In recent years, increasing people’s sense of gain has been a key phrase for governments at all levels across China. When it comes to “获得感”, how to accurately translate “获得”? In fact, “获得” can be translated in many ways, such as “achieve”, “acquire”, “earn”, “gain”, “get”, “obtain”, “secure” and “win”. We believe that “gain” is the best choice because it implies benefits and happiness.

Ex. 3: 网怒 Internet or Web rage: ～是当今的路怒 Internet rage is the road rage of our time

“网怒” as a new concept is an offshoot of the Internet age. When surfing the Internet, a user may be frustrated by such things as slow dial-up connections, busy servers, missing links, excessive results. This is what we called “Web rage”. Why do we translate “网怒” into “Internet or Web rage”, since “怒” can also refer to “anger” or “wrath” or “ire”? The reason is simple. “网路” is similar in nature to “路怒”(road rage), and “road rage” is a term that has already been included in the Oxford English Dictionary. It follows that “rage” can be best used to describe “怒”, which is also true of “桌怒”.


2.2 Strategy of Integrating Literal Translation with Annotation

Some neologisms are hardly intelligible because they are culturally imbued, reflecting
the distinctive features of a certain age. Such terms can be best translated by integrating literal translation with annotation. Let’s take the following for example.

Ex. 4: 淡紫语 lavender language [language peculiar to the gay community]

“淡紫” means “lavender” in English, so “淡紫语” can be translated as “lavender language”. Since the straight community may find it hard to make head or tail of this term, we may as well add an explanatory note to the translation of “淡紫语”. In fact, lavender language is only intended to be understood by the gay community, functioning as a kind of homosexual code and characterized by acronyms, plays on words and double meanings. It follows that annotation also plays a necessary role in translation.

Ex. 5: 灰天鹅 grey swan [figuratively, less unpredicted event which has uncertain outcomes]: 2011 年密西西比河创纪录的泛滥是一例 the record flooding of the Mississippi River in 2011 is a grey swan | 世上还有许多“～”呢，说这些事情不会发生吧，又相当可能，还是提前预计的好 there are lots of “grey swans” out there, unlikely occurrences that are just likely enough that they should be anticipated

Literally, “灰天鹅” is a swan which looks grey. However, a grey swan actually refers to an event which is considered to be unlikely to happen though potentially extremely significant. This term is in fact derived from the term “black swan” given by Nassim Nicholas Taleb in 2001 to extremely unlikely, unforeseen risk events with a major impact. Therefore, it is necessary to annotate the translation of “灰天鹅”.

2.3 Strategy of Integrating Literal Translation with Free Translation

Some culturally imbued neologisms are difficult to translate due to their cultural implications. It is obvious that literal translation alone will not solve the problem and should be complemented by free translation. In this way, our translation can be faithful to the original and innovative in form. Let’s take the following for example.

Ex. 6: 傍大款 1.to find a sugar daddy, to be a mistress to a rich man, to go with a rich man for the sake of money (of a girl): 她刚又傍上一大款 she has just looked another rich man | 她心甘情愿～ she is a willing kept woman 2.to fall back on a fat cat; to
throw oneself into the lap of a tycoon: 男性当中“～”的人 male “gold diggers” | 谁不想要～? Who doesn’t want a moneybags?

It can be seen that “傍大款” can be translated in five ways, and each version is characterized by the integration of literal translation with free translation. Literally, the best version for “大款” is “moneybags”, but we can use other versions when translating “大款”, such as “sugar daddy”, “rich man”, “fat cat”, “tycoon”, etc. Actually, the term “傍大款” has a short history, which emergence is due to a host of social and economic factors. It somehow fits in with the popular saying “做得好不如嫁得好” (better to marry well than do well).

Ex. 7: 脑残 <slang> (brain-impaired) 1.brainless; brain-dead; stupid: ～ 粉 a fanboy/fangirl 2.sth or sb unimaginable queer: 你干吗木呆呆地站在这里? ～吗? Why standing here like a dummy? Are you a nerdy freak?

Literally, “脑残” means the brain “脑子” is disabled “残疾”. Apparently, literal translation will not do because “脑残” has more profound meanings. This term first came about in a Japanese cartoon and had a pejorative sense. However, it is now often used in the Internet and subculture to describe someone or something stupid or unacceptable. Therefore, when translating this term, we should take into account both its denotative and connotative meanings by integrating literal translation with free translation.

Ex. 8: 脑洞 <slang> (brain holes) avenue to the brain; brain crease: ～大开 (one’s brain is open) experience an aha/eureka moment or have a sudden flash of insight | 她平时的～很大 normally, she’s quite imaginative

Literally, “脑洞” means the brain “脑子” has a hole “洞”. But what is a brain hole? Maybe nobody can understand it. Actually, “脑洞” is derived from “脑补”, signifying “the brain is cracked and should be filled with a strong imagination”.

410
Ex. 9: 脑子进水 <familiar> to have air bubbles in the brain; to have a brain fart: 我数学考砸了，脑子大进水啊！i just had a major brain fart on that math test

First used in North China, the term “脑子进水” can also be heard in daily life. It is used to describe someone who is not quite in his mind or to raise your objection to a novel phenomenon. When translating this term, we must ensure that the translation contains an idiomatic expression involving the use of the word “brain” (脑). On second thought, the best translation is “brain fart”. Interestingly, the Chinese phrase “进水” corresponds to the English word “fart”（放屁）.

3. Strategies of Free Translation and Its Applications

Some neologisms have profound cultural implications and cannot be translated literally. In this case, we can resort to free translation, i.e. using idiomatic English expressions in our translation instead of translating a text word for word. It is a strategy that can best bridge the gap between Chinese and English.

More often than not, free translation can help solve many of our problems in translation. Certain circumstances also dictate that we may as well integrate free translation with annotation and transliteration.

3.1 Strategy of Purely Free Translation

The strategy of purely free translation aims at translating Chinese neologisms into concise and lively English expressions without altering the intent of the original text. Let’s take the following for example.

Ex. 10: 爆表 <colloq> (to burst the indicator) 1. to break a race record 2. to go beyond index; to exceed the upper limit: 颜值～ be stunningly beautiful/handsome | 空气污染从“～”降到“危险”程度 the air pollution dipped to “hazardous” from “beyond index” | 他能力～了, 简直碉堡了！His abilities are beyond anyone’s wildest dreams. How amazing!

Literally, “爆表” means “to blow up a meter”, which makes little sense. Actually, the term “爆表” is a new concept used to describe someone or something phenomenal.
It is derived from a Japanese cartoon.

Ex. 11: 爆料 to expose, to publicize (sth surprising, personal, or sensational): 如此～的报道 such an eyebrow-raising revealing account/story | 自我～的书/回忆录/自传 a tell-all book/memoir-autobiography that creates/makes an exposé of someone | 有些记者事先接到～称，下周将再次爆发冲突 some journalists were tipped off in advance that a new conflict would break out next week

Literally, “爆” means “to appear unexpectedly”, and “料” means “material or raw material”. Obviously, literal translation will not work in this case. Over the years, the term “爆料” has been used to expose some news which is stunning or unexpected. Therefore, we can translate this term in many ways, such as “to tip off” or “to bring sth to light”.

Ex. 12: 高大上 <colloq> superb; top-notch; peerless; A-1: ～的设计 a high-end and classy design | 这些人从头到脚穿戴的都是名牌，可惜行为不那么～  they accoutered themselves with designer brand names from head to foot/toe; their behavior, however, is not always A-OK

Literally, “高” means “tall”, “大” means “big”, and “上” means “upper”. Obviously, the term “高大上” cannot be translated literally. This term is in fact an abbreviation of “高端” (high-end), “大气” (magnanimous) and “上档次” (classy). As a neologism, it was first used in 2013 and can be translated in many ways (see the above examples).

Ex. 13: 高级黑 (last word in calling sb black) 1.last word in being negative; clever put-down in a refined way; android or pokerfaced sarcasm: “哪儿有人写遗嘱，我想总～的有我一份好处” is a case of hilarious parody of an old saying | 那些表扬的话只不过是～ those positive remarks are merely sarcasm in disguise 2.to malign by innuendo (sarcasm, overstatement, etc)
Literally, “高级黑” can be translated as “high-grade black”. But this translation doesn’t mean anything. In fact, this is a new concept used to signify a kind of tactful sarcasm. It is through purely free translation that we can come up with a couple of its English versions.

Ex. 14. 给力 <colloq> 1. boosting; stimulating: 前冠军咋这么不～ the ex-champion is but a pathetic let-down | 这番话真～啊！ these remarks are a real pick-me-up！ 2. cool; awesome; nifty; bravo: 这首歌太～了！ this song rocks!

Strictly speaking, “给力” is not a new term; it was taken to mean a kind of corvée during the Northern and Southern Dynasties. Over the years, it has acquired many new meanings. A look at its intertextual context enables us to translate this term freely in many ways, such as “offer financial support”, “contribute to”, “happily”, “oh, my God”, “Can you believe it?”, “breaking news”, “awesome”, “bravo”, “cool”, “nifty” and “rock”. It must be noted that it is ill-advised to translate “给力” as “gelivable” because the translation sounds too abstruse.

3.2 Strategy of Integrating Free Translation with Annotation

Sometimes, free translation may lead to incomplete information. In this case, some explanatory notes can be used in our translation. Let’s take the following for example.

Ex. 15: 啃老族 dependent or boomerang kids, NEETs [young people who still depend on their parents economically]: 《广州日报》报道，调查显示，我国 30%的青年人是～，他们不上学，不上班，也不接受培训，主要靠父母供养 A survey shows that about 30% of Chinese youth are NEETs—people currently not in education, employment or training and are supported by their parents

Literally, “啃” means “to nibble”, and “老” means “senior citizens”. Obviously, literal translation does not work in this case. We must search for its most equivalent version in English. A careful study has shown that “boomerang kids” or “NEETs” (Not currently engaged in Employment, Education or Training) can best correspond to the term “啃老族”.

When translating other similar neologisms like “追星族”, “工薪族”, “打工族”,
“月光族”，“负翁族”，“毕婚族”，“乐活族”，“慢活族”，“SOHO 族”，“御宅族”，“彩虹族”，“晒黑族”，“飞特族”，“蚁族”，“考碗族”，“嫁碗族”，“新丁克族”，“恐归族”，“赖校族”，“捏捏族”或“奔奔族”，我们可能需要结合自由翻译与注释。

**Ex. 16:** 空窗期 1.[Med] window period [time between first infection and when the test can reliably detect that infection] 2.<fig> vacuum, interregnum [emotional and otherwise]: 2017 年直选前的～ the time pending the direct election in 2017 | 跟前一个妻子离婚，下一个还没找到前的～ during the emotionally uncommitted span between wives 3.<humour> period of time when one lives a sexless life [空窗 as intentionally erroneous form for 空床]

 literally, “空窗” means “an empty window”, which is hardly intelligible. How can a window be empty? It can be seen that literal translation will not work here. Actually, this is a new term and can be interpreted in different ways. It has at least three meanings, as indicated above, and each meaning needs to be annotated for better understanding.

### 3.3 Strategy of Integrating Free Translation with Transliteration

Some Chinese neologisms are culturally imbued and it is very difficult to find their English counterpart. In this case, we can resort to transliteration in our translation. Let’s take the following for example.

**Ex. 17:** 屌丝  <slang> 1.diaosi, loser, nonperson, underdog [usu self-depreciatory]: ～能给别人的只有一肚子气 what a lowly loser gives the world is a bellyful of bile 2.diaosi, disadvantaged male [in looks, income, career promise, etc]: ～文化 the culture of angst or the diaosi culture [ie the mentality and lifestyle of diaosi]

As a new concept, “屌丝” has profound cultural implications and it is hard to find its English counterpart. In this case, we can integrate free translation with transliteration in our translation. Now that there is no equivalent English version for this term, we may first translate this term as “diaosi” in pinyin, followed by its English
4. Strategy of Formative Imitation

Some Chinese neologisms can be translated by imitating some existing English expressions. In this case, we just change one or two words in an existing English expression when translating a Chinese neologism. Let’s take the following for example.

Ex. 18: 代驾 rent-a-driver; dial-a-driver: 酒后～ rent a driver to forestall DUI (driving under the influence) | 全家旅游～服务 a rent-a-car-and-driver service for whole-family tours

Literally, “代” means “to replace”, and “驾” means “driving”, but we cannot translate “代驾” as “to replace sb as a driver” which sounds awkward. Some people even translate this term as “designated driving”, which sounds puzzling. The best way is to search for some existing English expressions for reference. Then we hit upon the term “rent-a-car”. Isn’t it a good idea to imitate this term and translate “代驾” as “rent-a-driver” or “dial-a-driver”? It simply sounds too good to be true.

By the way, have you heard about “rent-a-crowd”? Can we translate “代孕妈妈” as “rent-a-mother” instead of “surrogate mother”?

Ex. 19: 究责风暴 blamestorming: 集思广益会演化成了～ brainstorming sessions became blamestorming sessions

Speak of “风暴”, and you will naturally think of “storm”. What first comes into our mind may be “头脑风暴” (brainstorm). Since “brain” and “storm” can be assembled to form “brainstorm”, why not translate “究责风暴” as “blamestorming”? It is quite a good idea to translate “究责风暴” by imitating “brainstorming”.

Ex. 20: 空气末日 aircapocalypse; airmageddon [Armageddon]: 呼声四起，说是要立法防止雾霾，延缓～的到来 people are clamouring for lawmaking to fight smogs,
postponing the advent of an airmageddon

The sight of “空气末日” reminds of “世界末日” (the end of the world). But can we translate “空气末日” as “the end of air”? The answer is no because it sounds too puzzling. In fact, “空气末日” as a new concept began to gain currency in 2013 when air pollution was so serious that people felt as if the end of the world were coming. In this case, we may refer to two words from The Bible: “apocalypse” and “armageddon”, which are suggestive of the end of the world. By imitating these two words, we may as well translate “空气末日” as “airpocalypse” or “airmageddon”. What a great strategy!

By the way, how can we translate “僵尸末日”? The best answer is “zombiepocalypse” or “zombiemageddon”.

5. Conclusion

This paper has investigated seven proven strategies for translating Chinese neologisms into English in a Chinese-English dictionary. These strategies include purely literal translation, integration of literal translation with annotation, integration of literal translation with free translation, purely free translation, integration of free translation with annotation, integration of free translation with transliteration, and formative imitation. Such explorations will definitely help promote the quality of translating Chinese neologisms by overcoming the seven deficiencies in MTI education mentioned by Prof He Gangqiang (2016), contributing to the cultural exchange between China and Western countries.

References


Abstract

The Chinese-English Dictionary (Unabridged) Volume I (Lu 2015) is a seminal dictionary compiled by a distinguished team led by the late Professor Lu Gusun. Since its publication, the dictionary has received many positive comments from reviewers for its intended promotion of Chinese culture, especially for the abundance in the Chinese culture-bound words. The translation of culture-bound words in the dictionary is well worth our attention, but not much research has touched upon the topic. We conducted text analysis of some items in the dictionary and the findings indicate that it not only achieves semantic, pragmatic and stylistic equivalences (Huang & Cheng 2001), but also goes “beyond equivalence” (Lu 2015). In other words, it takes its users from various cultural backgrounds into consideration in its explanation of culture-bound words and thus helps expand the use in different contexts. In addition, the dictionary reaches the goal of ‘acculturation’ proposed by Lu (2015), by striking a balance between foreignization and domestication (Venuti 1995). In this way, it functions as a bridge between foreign users and Chinese culture. Based on the above analysis, this paper suggests two ways to define culture-bound words, which are ‘literal translation (+[explanation]) +free translation ([substitution])’ and ‘transliteration + [explanation] (+substitution)’. The former is more suitable for idioms, and the latter can be applied to other types of culture-bound items. Besides, pictures are good aids. With the help of the previously mentioned translation methods, we are positive that Chinese culture can go global more effectively.

Keywords: culture-bound words, The Chinese-English Dictionary (Unabridged) Volume I, translation

1. Introduction

The Chinese-English Dictionary (Unabridged) Volume I (Lu 2015) (hereafter The Chinese-English Dictionary) is a seminal dictionary compiled by a distinguished team led by the late Professor Lu Gusun. It was published in August 2015 after fifteen years of hard work. Since its publication, the dictionary has received many positive
comments, especially for its promotion of Chinese culture which Professor Lu was dedicated to. Huang Yuanshen commented that the dictionary’s selection of culture-bound words bear ‘the Chinese mark’, Zhang Boran mentioned its ‘Greater China’ characteristic, and Huang Fuhai also considered it to be of ‘Chinese characteristic’ (Xu 2015). From their opinion, we can see the importance the dictionary attaches to culture-bound words. In its selection of entry words, *The Chinese-English Dictionary* displays the rich culture of China; in its definition of headwords, it keeps as much Chinese cultural imagery as possible.

Culture-bound words are resulted from cultural differences because sometimes there is one thing in one culture while there is no such correspondence in the other, or the other way around (Li and Zhou 2011). In reference to previous classifications such as Newmark’s (1988), this paper divides culture-bound words into four categories, namely daily, spiritual, natural and historical culture-bound words. Daily culture-bound words cover all aspects of life, including clothing, food, accommodation, transportation, so on and so forth. Examples are 旗袍, 粽子, 胡同 and 轿子. Spiritual culture-bound words are myths, legends, festivals, customs, philosophy, religions, aesthetics and values, such as 端午节, 临时抱佛脚, 助纣为虐, 黄梅戏 and 三从四德. Natural culture-bound words cover both animals and plants, like 梅兰竹菊 and 熊猫. Historical culture-bound words are mostly proper nouns with specific history, with 孔子 and 嵩山 being instances.

In order to show *The Chinese-English Dictionary*’s focus on culture-bound words, the paper compares it with three other Chinese-English dictionaries, which are *New Age Chinese-English Dictionary* (Wu and Cheng 2006) (hereafter New Age), *A Chinese-English Dictionary* (Yao 2009), and *A New Century Chinese-English Dictionary* (Hui 2002) (hereafter New Century) in terms of the number and distribution of their culture-bound words in their A section. The results are provided in Table 1 and Table 2.

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Number of culture-bound words</th>
<th>Number of entry words</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>The Chinese-English Dictionary</em></td>
<td>74</td>
<td>1539</td>
<td>4.8%</td>
</tr>
<tr>
<td>New Age</td>
<td>26</td>
<td>847</td>
<td>3.1%</td>
</tr>
<tr>
<td><em>A Chinese-English Dictionary</em></td>
<td>22</td>
<td>575</td>
<td>3.8%</td>
</tr>
</tbody>
</table>

1 The comments are originally in Chinese and are translated into English by the first author.
As is shown in Table 1, in the investigated section, The Chinese-English Dictionary has a larger number of as well as a greater proportion of culture-bound words than the other three dictionaries. This gives us an insight into its cultural focus.

Table 2 Distribution of culture-bound words in A section

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Daily</th>
<th>Spiritual</th>
<th>Natural</th>
<th>Historical</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>The Chinese-English Dictionary</em></td>
<td>8</td>
<td>19 (25.7%)</td>
<td>32 (43.2%)</td>
<td>15 (20.3%)</td>
<td>74</td>
</tr>
<tr>
<td><em>New Age</em></td>
<td>1 (3.8%)</td>
<td>2 (7.7%)</td>
<td>14 (53.8%)</td>
<td>9 (34.7%)</td>
<td>26</td>
</tr>
<tr>
<td><em>A Chinese-English Dictionary</em></td>
<td>2 (9.1%)</td>
<td>1 (4.5%)</td>
<td>15 (68.2%)</td>
<td>4 (18.2%)</td>
<td>22</td>
</tr>
<tr>
<td><em>New Century</em></td>
<td>1 (4%)</td>
<td>5 (20%)</td>
<td>13 (52%)</td>
<td>6 (24%)</td>
<td>25</td>
</tr>
</tbody>
</table>

Also, Table 2 is a demonstration of the distribution of culture-bound words in the four categories mentioned previously. Comparatively speaking, The Chinese-English Dictionary has a more evenly distributed selection of culture-bound words than the rest.

*The Chinese-English Dictionary* features in its abundance in the culture-bound words representing traditional Chinese culture. The cultural flavor of the dictionary makes the translation in the definition part a complicated process. However, there is not much research looking into the topic. Therefore, this paper attempts to fill the current research gap by investigating the translation strategies employed in dealing with culture-bound words in this dictionary.

2. Translation of culture-bound words in *The Chinese-English Dictionary*

Generally speaking, there are a few common translation methods used to define culture-bound words in bilingual dictionaries. The following table gives a rough review of each method.

Table 3 Review of translation methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Strength</th>
<th>Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transliteration</td>
<td>It is convenient and keeps the original pronunciation.</td>
<td>To those foreign users who do not know the Chinese language or its pronunciation, it does not help their understanding and might make them confused.</td>
</tr>
</tbody>
</table>
Literal translation | It keeps the cultural image and helps promote cultural communication and history maintenance. | In translation, sometimes the sum of parts does not equal the whole, so it is possible that dictionary users will not get the cultural connotation of the culture-bound words immediately.

Free translation | It makes the entry word more comprehensible. | The loss of culture is unavoidable.

Substitution | It reduces the level of comprehension difficulty, especially for foreign users. | Accuracy cannot be guaranteed.

Explanation | It enhances the understanding of dictionary users. | It makes the entry long and unfit for actual use. Thus, it is not in accordance with the ‘user-friendly’ principle advocated in dictionary compilation.

Apart from the five major translation methods listed above, there are other ways: abridged translation, acronymization and common translation. Sometimes, tautology is common in Chinese, especially in Chinese four-character idioms, so abridged translation can be employed. Acronymization is rarely used in Chinese-English translation of culture-bound words. Common translation is established translation agreed upon by the majority, such as the translation of some festivals, scenic spots and so on.

It is worth mentioning that the eight methods are not mutually exclusive. Instead, they overlap to some extent and together make up the definition part of an entry. Next, we will focus on ‘beyond equivalence’ and ‘acculturation’ proposed by Lu (2015) from the perspectives of equivalence theory and cultural translation theory respectively.

2.1 ‘Beyond equivalence’ from the perspective of equivalence theory

In terms of the translation of culture-bound words, the principle of equivalence has been valued in the academia. Heng and Cheng (1995) sum up Nida’s and Hartmann’s views on equivalence in translation: the translation of idioms is regarded as a continuum and is divided into word-for-word translation, literal translation, free translation and functional equivalence. Nida, as the representative of equivalence theory, holds that dynamic equivalence is the appropriate equivalence, because it is not only the equivalence of forms but also the equivalence of semantics and style. Huang and Chen (2001) take the principle of equivalence as the basic principle of translation in bilingual dictionaries, which can be reflected in three aspects -- semantic equivalence, stylistic equivalence and pragmatic equivalence. A balance should be stricken between foreignization and domestication to realize the equivalence, pragmatic equivalence in particular, of culture-bound words while achieving the
diversity principle of definition. Yong and Peng (2007) call for attention to the equivalence principle in the translation of culture-bound words in bilingual dictionaries, so they suggest using the following strategies to translate idioms depending on the degree of equivalence: literal translation, free translation, literal translation plus free translation, free translation plus explanation or transferred translation. On the basis of previous studies, this paper makes use of the three equivalences by Huang and Chen (2001) to analyze the translation of culture-bound words in *The Chinese-English Dictionary* and then attempts to point out how it goes beyond equivalence.

The most basic requirement for translation of culture-bound words is to achieve semantic equivalence, i.e. to maintain the core semantic feature of the source-language word. For instance, 胡同 in *The Chinese-English Dictionary* is defined as ‘hutung or hutong; lane; alley; small back street; bystreet’, 胡同 in *Xiandai Hanyu Cidian* (hereafter *Xiandai*) is defined as ‘巷; 小街道’, and 巷 in *Xiandai* is ‘较窄的街道’.

We can see that [-大] and [街道] are the two main semantic features of the word. In *Merriam Webster’s Collegiate Dictionary* (hereafter *Webster*), lane is defined as ‘a narrow passageway between fences or hedges’, alley as ‘a narrow street’, and bystreet as ‘a street off a main thoroughfare’. Together with ‘small back street’, these definitions show that [-big] and [street] are the overlapping semantic features. Thus, *The Chinese-English Dictionary* manages to achieve semantic equivalence in its translation of 胡同. Zugsta (1971) suggests that the equivalent word should be able to be directly used in the target language context. For some Chinese culture-bound words, there is no equivalent imagery in the target language culture. The uniqueness of those culture-bound words makes some dictionary apply only the translation method of explanation. For example, *New Age* defines 粽子 as ‘pyramid-shaped dumpling made of glutinous rice wrapped in bamboo or reed leaves (eaten during the Dragon Boat Festival on the fifth day of the fifth lunar month)’. This helps understanding but does help using. Aware of this point, *The Chinese-English Dictionary* provides pragmatic correspondent in its definition to facilitate the use of culture-bound words.

锅贴, as an example, is translated in *The Chinese-English Dictionary* as follows: [Chin Culin] guotie, pansticker [pan-fried Chinese dumpling with a minced meat or vegetable filling]. This definition offers both transliteration and literal translation with explanation, which enhances the target users’ understanding and can also be used directly.

Stylistic equivalence is shown in the match of rhetorical features between the source language and the target language. The four-character idiom 锦上添花 is a case in point, which is often mistranslated into ‘to gild the lily’. In fact, they are not
equivalent in style. The former is positive, which means making good better, while the latter is negative, which is in Webster ‘to add unnecessary ornamentation to something beautiful in its own right’ and is actually closer to 画蛇添足. The Chinese-English Dictionary sticks to the principle of stylistic equivalence well by translating 锦上添花 into ‘(to add flowers to brocade) to improve on perfection’ with literal translation and free translation and without wrong substitution. 画蛇添足 in this dictionary is ‘(to paint a snake and add legs to it) to ruin the effect by adding something superfluous; to over-egg the pudding; to gild the lily’. From this example, we can see that appropriate substitution, together with literal translation and free translation, helps achieve stylistic equivalence.

That The Chinese-English Dictionary goes ‘beyond equivalence’ means that it on the basis of semantic, pragmatic and stylistic equivalence ‘extends the cultural scope’ (Lu 2015). This is especially true for idioms: The Chinese-English Dictionary tends to use literal translation, then free translation and sometimes substitution. Literal translation in a definition of a culture-bound word conveys the explicit meaning, free translation shows implicit meaning, and substitution takes its use into consideration.

Take 班门弄斧 as an example. It is defined in the dictionary as follow: (to show off one’s axe before the door of Lu Ban— the master carpenter) to show off one’s meager skills before an expert; to display one’s lesser skills before a connoisseur; to teach one’s grandmother to suck or roast eggs. This definition first in brackets applies literal translation and then explains Luban, and then uses free translation followed by substitution. The explanation in the brackets well expresses to the dictionary users the Chinese cultural image in the four-character idiom, the free translation afterwards is aimed for the users’ better understanding, and the last substitution expands its cultural scope. Similarly, the definition of 害群之马 -- ‘(an evil horse in the herd) person harmful to a group or society; black sheep’ achieves the same effect. ‘The meaning and use of an entry word can be better clarified with the help of examples’ (Li and Zhou 2011), so the complex cultural nature of culture-bound words also requires examples as a way of illustrative explanation. For instance, the example sentence in the entry of 春节 is ‘~是家人团聚的日子 family members reunite at the Spring Festival’. The given example further illustrates the importance of this festival in the traditional Chinese family culture, on the basis of its definition ‘Spring Festival; Lunar New Year; Chinese New Year’.

In a word, The Chinese-English Dictionary handles with culture-bound words in an appropriate way: it takes the cultural backgrounds of different readers into consideration and explains the culture in a detailed way. In this way, it achieves the
goal of ‘beyond equivalence’.

3. ‘Acculturation ’ from the perspective of cultural translation

Venuti (1995) sums the two translation strategies on the basis of previous translation studies, which are foreignization and domestication. The former brings the readers home and the latter takes them abroad. That is to say, domestication helps cross-cultural understanding and the latter facilitates cross-cultural communication. Since the cultural factor is inevitable in translation, the two strategies also work for culture-bound words. Hu (2006) calls for focus on foreignization when dealing with the translation of culture-bound words so that Chinese culture can be better spread. Huang and Chen (2001) hold that there should be a balance in foreignization and domestication and there should be diversity in the definition of culture-bound words. In the same vein, Wan (2006, 2016) also suggests combining domestication and foreignization with regard to the translation of culture-bound words. In this respect, The Chinese-English Dictionary aims to present Chinese culture to maximum by using the two translation strategies according to different culture-bound words.

As is mentioned previously, domestication is intended to circulate the culture of the source language, so transliteration and literal translation come in handy. For instance, The Chinese-English Dictionary literally translates 黄鹤楼 into ‘Yellow Crane Tower’ and adds in square brackets the history behind it -- ‘originally built in 223 on Mt Sheshan (蛇山), Wuhan, Hubei Province’ -- including the time and place of its construction. This way of translation helps promote the global circulation of Chinese culture.

Although scholars like Venuti (1995) and Hu (2006) advocate foreignization in the translation of culture-bound words, The Chinese-English Dictionary takes a similar stance with researchers such as Huang and Chen (2001) and Wan (2006) -- a combination of foreignization and domestication. The domestication strategy includes such translation methods as explanation, substitution and free translation. Take 藏龙卧虎 as an example. Its definition ‘(hidden dragons and crouching tigers) talented people who remain obscure to the undiscerning eye; undiscovered talent’ is a combination of literal translation and free translation. Another instance would be 饺子. In The Chinese-English Dictionary, 饺子 is defined as ‘jiaozi, half-moon or Chinese dumpling, tortellini [usu with minced meat and vegetable filling]’. The first part is transliteration, the second part is explanation and the last part is substitution. This is what is mentioned in the preface of The Chinese-English Dictionary -- ‘acculturation’. The advantage for its combination of domestication and foreignization is that it
presents to its users similar images of two different cultures and thus enables them to ‘ferry between two cultures’ (Lu 2015).

In a word, The Chinese-English Dictionary pays particular attention to the cross-cultural communication. It firstly displays Chinese culture to its users through foreignization; then, to facilitate understanding, it sometimes explains the cultural image or substitutes it with an equivalent with the help of domestication, so as to ‘extend the cultural scope’ (Lu 2015). In addition, compared to other dictionaries, The Chinese-English Dictionary attaches more importance to explanatory and complementary notes.

4. Implications

With the development of corpus linguistics as well as other tools, the objectivity in the compilation of a dictionary has increased, but the subjective factors are still unavoidable. Therefore, dictionary compilers need take a ‘positive critical attitude’ (Wang and Lu 2006), so should dictionary critics. As an unabridged Chinese-English dictionary, The Chinese-English Dictionary is not without flaws. The 胡同 discussed previously lacks explanation; there seems to be an absence of literal translation in the definitions of 不僧不俗 and 带臊貂褂 -- ‘not one thing nor another; neither fish nor fowl’ and ‘highest ceremonial or formal dress [during the Qing Dynasty]’ respectively. These definitions do not facilitate understanding and thus cannot help spread Chinese culture. Another instance is that in the entry of 黄鹤楼, the example sentence is ‘手持绿玉杖，朝别~’, which is less well known than ‘昔人已乘黄鹤去，此地空余黄鹤楼’.

We think the latter poem would be a better option for the purpose of cross-cultural communication.

Therefore, based on the observation and analysis of the translation strategies of The Chinese-English Dictionary, we propose the following two strategies to translate culture-bound words in bilingual dictionaries.

(1) literal translation (+[explanation]) +free translation (+substitution)
(2) transliteration +explanation (+substitution)

The first applies to culture-bound words such as four-character idioms. It suggests using literal translation and then free translation. If the entry word is rich in cultural information, explanation can be used as complementary information. Although according to Newmark (2001), it is unrealistic to find complete equivalents in the source language and the target language, we still think that a bilingual dictionary should try its best to provide its users with an equivalent for culture-bound words. The equivalent might not have the same cultural image, but it need be as close to the entry
word in its meaning and style as possible and should not interfere with the users’ understanding and application of the entry word. Therefore, if there exists such an equivalent word, it is suggested that substitution be used to complete the definition of an entry. *The Chinese-English Dictionary* handles well in this respect, like the aforementioned idiom 班门弄斧. The substitution may be criticized for its mismatch in style, but the hard work of dictionary compilers in the translation of culture-bound words cannot be denied. In fact, if there is no such equivalent, then substitution is not necessary. For instance, 暗度陈仓 is defined in *The Chinese-English Dictionary* as ‘(to dispatch armies to attack Chencang in secret, a tactic by Liu Bang [刘邦] in defeating Xiang Yu [项羽] ) to feign attack; to make a feint; to do sth secretly while ostensibly doing sth else [eg to commit adultery]’. The definition does not include an English equivalent, but is still acceptable.

The second applies to other types of culture-bound words. It suggests using transliteration and then explanation. The advice for the choice of substitution is the same in the previous paragraph. In *The Chinese-English Dictionary*, 宫保鸡丁 is ‘gongbao chicken [Sichuan-style stir-fried diced chicken with chili sauce and peanuts]’ and 孔子 is ‘Confucius, Kongzi [literally “Kong the master”. 551-479 BC, philosopher, politician, moral teacher, and founder of the Confucian School]’. This definition is very appropriate. But there are also some definitions that could have been better. Take the previously discussed 胡同 as an instance. Its definition in *The Chinese-English Dictionary* is not perfect and we suggest a refined definition as follows: ‘hutong [a type of narrow street or alley, commonly associated with northern Chinese cities, most prominently Beijing]; lane’. In this way, the entry includes the corresponding English cultural image of the entry word and also keeps the Chinese cultural meaning through explanation.

Apart from the two main strategies mentioned above, some visual aids (i.e. illustrations) can also be made use of to increase efficiency. Illustrations in a bilingual dictionary on the one hand clarifies the entries and thus reduces the burden of compilers to some extent and on the other arouses the users’ interest in language learning and makes them consult the dictionary more frequently. Using illustrations in the definitions of culture-bound words can help visualize the entry, because some abstract culture is embedded in concrete objects which are difficult to be fully explained with language. Although illustrations will inevitably take up the precious space of a dictionary, it is sill worthwhile to include them in entries like 饺子, 鼎 and 中国结. However, most Chinese-English dictionaries, *The Chinese-English Dictionary*
included, do not provide illustrations for those culture-bound words that need pictures as complementary information. Take 旗袍 as an example. We suggest giving such a definition as ‘qipao, cheongsam [a tight-fitting dress worn by Chinese women, usually of silk or cotton, with a stand-up collar and a slit in one side of the skirt]’ and then providing the following illustration taken from the Internet.

![Image of qipao](image.png)

5. Conclusion

In the context of globalization, Chinese-English dictionaries should be aimed to make Chinese culture go global, be their target users Chinese or foreigners. Zhou (2014) mentioned that we lack authoritative translation. Dictionaries are supposed to be the medium of such translation. The translation of culture-bound words in The Chinese-English Dictionary well represents Lu’s (2015) ideas of ‘beyond equivalence’ and ‘acculturation’. While making sure the equivalence of core semantic meaning and style and the convenience of users, The Chinese-English Dictionary manages to make Chinese cultural images understood, applied and spread. Also, on the basis of the analysis of the dictionary, we hold that ‘literal translation (+[explanation]) + free translation (+[substitution])’ and ‘transliteration + [explanation] (+[substitution])’ are two appropriate ways to deal with culture-bound words in Chinese-English dictionaries. Illustrations can be used as additional aids to add to the definition of culture-bound words in bilingual dictionaries. In this way, cultural diversity can be maintained, the communication between cultures can be promoted and consequently Chinese culture will be better understood across the globe.

Acknowledgments

This study is funded by the Student Academic Scientific Innovation Action Support Plan of Fudan University. We would like to thank Dr. Wan Jiangbo for her guidance in the process of the study and her comments on the article.
References

黄建华、陈楚祥. 2001. 《双语词典学导论》. 北京: 商务印书馆.
陆谷孙. 2015. 《中华汉英大词典（上）》. 上海: 复旦大学出版社.
万江波. 2006. 《双语词典的翻译研究》. 上海: 复旦大学出版社.
吴景荣、程镇球. 2006. 《新时代汉英大词典》. 北京: 商务印书馆.
徐明徽. 2015. 搞七念三、钓鱼执法、官本位，《中华汉英大词典》怎么译？《澎湃新闻》http://www.thepaper.cn/newsDetail_forward_1316903.
姚小平. 2009. 《汉英词典（第三版）》. 北京: 外语教学与研究出版社.

周明伟. 2014. 建设国际化翻译人才队伍 推动中国文化走出去. 《中国翻译》(5): 5-6.
On Bilingual Dictionary’s Strategy of Fixing Lexical Form in Dealing with Lexical Image Transformed between Languages

Yuanke Tao
Research Fellow, Guangdong Federation of Social Sciences
taoyuank@aliyun.com

Abstract

Lexical image refers to the meaning reason of lexical forms, which has been named differently as inner form, speech image, verbal image, and just image or icon by different scholars. In order to express its sense of the image of specific word, I prefer to use the term of lexical image to call the conception. The conception of lexical image is interdisciplinary and concerns many disciplines such as art, literature, aesthetics, and inter media etc. That is why it is comprehensively valuable for us to study it as a common problem in bilingual dictionary making.

Although Chinese character belongs to semantic script, people catch the meaning of some Chinese graphic symbol string through the meaning of words in it but not through characters in it. And the differences of lexical image between two languages appear mainly in compounds. Lexical images are formed under certain background of nature, language and culture; therefore, people with different languages may create common lexical images for similar feeling of some objects, which could be confirmed by some word-to-word translations. But on the other hand, many different lexical images of same objects may be created in different languages for their different cognitional backgrounds of nature, culture and linguistic arbitrariness, and cause various translations, which lexical forms need to be reevaluated and fixed in transforming between languages, so as to keep or moderate lexical images as nearer to the original ones as possible.

Two rules based on Chinese lexical image are put forward in the paper as bilingual dictionary’s strategy to deal with lexical image transformed from Chinese to English. One is ‘Chinese compounding lexical image → English native lexical image + linguistic difference’, the other is ‘Chinese two-alliterated (or two-rhymed) characters image → transliterated + meaning explanation’. And other three rules, based on native English lexical components, are put forward as well. One is to make use of English morphemic components instead of words to translate Chinese lexical image into English formative-lexical image, another is to transform Chinese lexical image into English semantic loan to avoid too many transliterations, and the other refers to the object of Chinese lexical image with temporary English transliteration and additional short explanation. In all patterns of translation above, it is most helpful for cognizing and remembering the origin lexical images, if which is transformed into morphemic lexical image in target language.

Keywords: lexical image, bilingual dictionary, linguistic difference, transliteration,
There are different vision angle for cognizing the definition and explanation of ‘lexical image’ in linguistic research circles and literary art research circles.

Linguists considered that lexical image refers to a rational relationship between a lexical form and its semantic referent, which M. von. Humboldt called as inner form and referred to the etymological structure of word or its componential form-meaning relational rationale. Traditional Chinese linguistics also had a discussion about the cause of being named. Since the Qing dynasty, from ‘Mr. Duan Yucai (1735-1815), to Mr. Wang Niansun (1744-1832), Wang Yinzhi (1766-1834), Liu Shipei (1884-1920), and Yang Shuda (1885-1956), they made great endeavor in the academic field. Although Mr. Liu Shipei used ‘meaning image’ in his paper ‘My Explanation about the Origin of Chinese Character Meaning from Character Pronunciation’, Mr. Yu Min addressed it at first as ‘speech image’ in English translation, which also refers to etymological rationale of word to be named. After that, some scholar made use of the terminological conception to study the etymological relations within word family, and someone further applied ‘lexical image’ in probing the families of two-alliterated words, analyzed them into basic lexical image (common reason for the same family words to be named), accessory lexical image (with accessory meaning or meaning extended from the basic lexical image), and concrete lexical meaning (with concrete object which is depended by basic lexical image and accessory image), and so as to interpret the semantic the semantic relevance of etymological two-alliterated words and their extending relationship between each other.

On the other hand, in the field of literary art studies, scholars prefer to cognize and interpret image in the perspective of lexical expressing of image and lexical comprehending image, and consider that ‘lexical image is just the figure referred to by words’. According to the explanation given on Baidu encyclopedia platform as well as Hudong encyclopedia platform, ‘verbal image is specific term of the new criticism school, which is a lingual image interpreted directly as an image made of lexical materials, which is a lexical image word-based and different from the image staying in consciousness and imagine dropped out of language, such as the impressions caused by mountain, water, person, and spring mountain, autumn water, beautiful woman, jade-like arms, and golden hair etc.’ Shortly speaking, it is interpreted from the perspective of how comprehending words, which takes lexical image as an image formed by words.

However, since the term ‘image’ can refer to the image to be described, metaphorical image, symbolized image, situational object, and character image, and the American theorists of new criticism tended to substitute image with icon, which was
translated into Chinese as 语象 (lexical image) or 具词的象 (image of specific word). And there are some Chinese scholars who used 语言级的象 (image in lingual leve) or 具词的形象 (image of specific word) to express the meaning of lexical image, which image characteristic is just what the lingual form refers to and is just what we like to concern when discussing lexical image in the scope of lexical definition. So, I’d like to use ‘lexical image’ to refer to the image of specific word in my paper. 

Moreover, in the perspective of literary narration to interpret lexical image, there is a concept of ekphrasis or ecphrasis. It claims that ‘the Narration of lexical image is a trans-border term which core content is the relationship between word and image, so it is also taken as inter-media studies, inter-arts studies, picture theory and so on, and its synonyms or relevant words may also be poem like picture, pictorialism, iconography (iconology), spatial form and so on. In English, its scope has covered interdisciplinary and inter-media content such as literature, art, aesthetics and so on, and has multiple meanings in use but with no meaning mistake yet.’ In fact, being an expression of formational image in speech, lexical image is capable to refer to the words of different disciplines, which is just the common significance for us to discuss the phenomenon of lexical image transformed between two languages and the existing problems.

1. Difference of Lexical Image Reflected in Transforming Lexical Image between Two Languages

Although Chinese character is called as an ideography or semantic script, and there are great differences between Chinese character form and English letter form, people would never pay attention to the composition of Chinese character 家 (home) composed of a deep house 宀 and 臬 (which means ‘male pig’) when they read the idiom 家陡四壁 (a home has nothing but four walls), would not think the Chinese character 明 (bright) is composed of 日 (sun) and 月 (moon) when they read the stereotyped expression 明日復明日 (tomorrow and again tomorrow), and would not comprehend the meaning of Chinese word 吞 (swallow) as its composition with the same initial of 天 and a radical of 口 (mouth) when they read another idiom 氣吞山河 (filled with the heroic spirit that conquers mountains and rivers), and so on. Therefore, I will not base on graphic compositional

---

6 赵毅衡《新批评》，北京：中国社会科学出版社，1986年，第 132-136 页。
7 王安、程锡麟《语象叙事》，《外国文学》，2016年第4期。
level to discuss the difference of lexical image between Chinese and English, but focus on lexical level to recognize the difference of lexical image between Chinese and English.

Essentially, lexical image is something produced by people through perceiving and mastering of things in various natural conditions and different cultural backgrounds, so, a same sensory organ may produce similar lexical image. For example,

(1) 着凉: catch a cold→重感冒: a heavy cold; 后门: back door; black horse→黑马; honeymoon→蜜月; hornet’s nest↔马蜂窝;

In the above examples, catching cold may leads to have a touch of flu, both Chinese speakers and English speakers may have such an experience, that is why English also say ‘catch a cold’ to mean 着凉, and obviously both Chinese and English hold the same lexical image, and even express 重感冒 with a same lexical image of ‘heavy cold’, a same ill feeling. Moreover, Chinese 后门 has the same lexical image as English ‘back door’ as well, and also extends a lexical image as an extended near sense referring to a way for doing embezzlement. And in the last two cases in the example (1), English and Chinese also share the same lexical images with similar compounding structures, which reflect that morphemic translation may create new compounds of the same lexical image with highly identity to that of the original language.

On the other hand, the lexical images of different languages are also something perceptually cognized by people in different natural environment and different cultural environment, therefore, different languages may produce different lexical images of the same objective referents. For examples,

(2) 深呼吸: a heavy breath; 大雨: a heavy rain; thin-skinned→脸皮薄; 看家狗: watchdog, housedog; 粤孱鸡婆: mother hen

The first case in example (2) reflects that the basic lexical image of English ‘heavy’ is ‘having great weight’, which is similar to Chinese lexical image of 重, however, ‘heavy’ can collocate with ‘breath’ or ‘rain’ in English to form compounding lexical images, which Chinese 重 could not collocate with corresponding words of ‘breath’ nor ‘rain’ to form correspondent lexical image. While the last two cases in example (2) tell us that same things may cause to produce those lexical images having something
different in English and Chinese. The English words ‘watchdog’ and ‘housedog’ caught lexical images from different vision angles, which come together to form a Chinese lexical image 看家狗 (guarding-house dog). And similarly getting lexical image from ‘hen’, Cantonese 打鸡婆 is quite different from English ‘mother hen’ as lexical image, for the former refers to someone unreasonable and rude, while the latter has analogy of someone cherishing others too carefully. So, in the light of those cases, which may lead to a basic law, that a common basic lexical image and an accessory lexical image can form a concrete lexical image, the translation of Chinese lexical image into English one may be induced as the following pattern.

**Chinese lexical image → English native lexical image + lingual difference**

Furthermore, in some aspects of foundational common lexical image, different cultures may produce some lexical images with absolutely different implied meanings of the lexical images, especially in idioms, which reflect cultural difference among the comparative languages. For examples,

(3) Behave like a hog 像猪一样举止粗鲁 → as independent as a hog on ice 像冰上的猪一样独立

In example (3), the lexical image before the arrow implies the same meaning of lexical image in both Chinese and English, while the English lexical image behind the arrow does not exist in Chinese, which obviously reflects a cultural difference of lexical image. In the last case, we hardly find a correspondent Chinese idiom to translate such an English lexical image of independent pig, but have to do word-to-word translation plus some additional interpretation.

Moreover, some structuring characteristics of a language may also lead to such untranslatable lexical image with native idiom lexical image in another language. For instance, there are large amount of alliterated or rhyming twin-syllable words in Chinese, most of which are adjectives, as a salient feature of Chinese vocabulary. If we want to translate those twin-syllable words into English, we have to use a specific method. Here I’d like to generalize the pattern of interpreting alliterated or rhyming twin-syllable Chinese words in Chinese-English dictionary into the follows: ‘**Chinese lexical image of alliterated or rhyming twin-syllable words → transliteration of Chinese Pinyin + liberal translation (interpretation)**’. For examples,

(4) [望洋] wangyang (in a muddle): 望洋向若而叹 → （The river master）sighted in a
muddle toward him.

[望阳] watch the sun 〈wangyang〉武帝望阳 → The Wu emperor watched the sun.

In example (4), the last case applies a word-to-word literal translation to deal with the original word with a note in angle brackets to imitate the pronunciation of the twin-syllable word, while the first case directly offers the spelling of original word with a note in round brackets to reveal its meaning by literal translation. And both the cases are equipped with examples and translations.

From the analysis above, we may know that, because of the influence of different cultures, different languages catch and express the same meaning images not only with common aspects, but also with more different side aspects. Different languages use different words to catch and reflect same object figure, which may form different lexical images by different lingual words and produce the lexical image with different components. The Method of morphemic literal translation and the method of translational interpreting native lexical image plus lingual difference can transform lexical image of different lingual words better to some extent, so they can offer some strategic support in bilingual treatment of lexical image transformation.

2. Strategy of fixing lexical image in introducing translation and description

Since there are cultural differences and even lacks in the cultures on which languages are founded, we often hardly find out native-lingual lexical image to translate directly the original lexical image into target language, and have to make use of other native lexical image of target language to catch and describe the lexical image of such foreign language. For example,

(4) 玫瑰： rose, rosier（法语）/ rose（英语）;

月季： rose de Chine ou rosier del Chine/ rose d’Inde ou rosier d’Inde（法语）/ Chinese rose（英语）;

山茶（花）： rose du Japon→caméllia (Camellia japonica)（法语）/ camellia（英语）

In the example (5), we may find that, both English and French have only the lexical image of Chinese 玫瑰, i.e. ‘rose’ in English and ‘rose’ or ‘rosier’ in French, but do not
have suitable correspondent lexical images to express Chinese lexical image of 月季 and 山茶花, so they need to borrow those native words of target languages to form compounds of ‘Chinese rose’ in English and ‘rose de Chine’ or ‘rose d’Inde’ and ‘rose du Japon’ and so on, which pattern of the translations may be expressed as ‘location + lexical image of the same kind’. Although the French translation of Chinese 山茶花 is finally fixed as ‘camellia’, Chinese 月季 is still translated as its early expedient name and lexical image in English and French up till now. Maybe it will last forever.

With regards to those lexical images shared commonly by both original language and target language, morphemic literal translation could be used to reach the purpose of maintaining original lexical image well, so people prefer to select it in practice. However, considering the problem more deeply, we may find that there are also problems in the perspective of morphemic literal translation to make the original lexical image meet with the lexical rules of target language; especially those new cultural and political terms require effort to promote the translation to agree with the lexical rules of target language. For example,

(6) 三个代表: three representatives →tri-representative theory; 五律: poem with eight five-character lines;

八字宪法: Eight-character Charter for Agriculture →Eight-point Charter for Agriculture;

In example (6), the current Chinese political term 三个代表 should be comprehended as a theory or concept instead of persons, so, there is a distinction between syntactic collocation of words and lexical compounding of morphemes, with the latter expressing theoretical integrality more suitably. And the number ‘five’(五)in 五律 refers to five Chinese characters in a poetry line, so five-character line is a basic feature of that kind of poem. If at lexical level, it could be also translated as ‘quingue-character lines’ which is of primitive simplicity. Moreover, the Chinese term with eight 八字宪法 actually includes eight aspects, i.e. soil, fertilizer, water, seed, density, protection, management, implement, so, ‘eight-point’ would be better than ‘eight-character’ to express the meaning of original term. From the analysis of example (6), we may see that comprehending and translating lexical image at morphemic level
would be able to maintain the original Chinese lexical images in English to some extent, and we should also improve the lexical images at the morphemic level if it is needed.

There are more and more new terms appearing now, so that people cannot find suitable lexical images in target language for them, therefore, tentative transliteration is often applied first, and some of them continue to be employed year by year, for instances, 工夫→gongfu, gene→基因, Macdon→麦当劳, and so on. But others of them are influenced by the target lingual formation tendency and later also appear some other literal translation with Chinese morphemes. For examples,

(7) Taxi→的士（粤语）→计程车（台湾)/出租车（大陆); Internet→英特尔→互联网。

In example (7), ‘taxi’ was transliterated into Chinese (especially Cantonese) as 的士, but substituted by another word 计程车 (car with distance record) in Taiwan and 出租车 (car for rent) in China’s mainland through liberal translation. And ‘internet’ was introduced into Chinese through half-transliteration 英特网 at first, however, in recent years it has been substituted gradually by 互联网 (net for mutual connection). The development and fixing of these cases confirm that Chinese as a target language holds an inner tendency of meaning lexicalization in dealing with borrowing words. So, in example (7), 计程车 and 出租车 present that Chinese in different places caught different functions of ‘taxi’ as their rational meaning in literal translation and formed different new lexical images. While 互联网 literal translated English term ‘internet’ at morphemic level, so it has a quite similar lexical image as the original one.

Both translational words and other lingual units should abide by economic rule, so, in the process of borrowing words, if there are certain abbreviations of terms in original language, the abbreviations may be adopted and applied directly in target language, and in the meantime, morphemic literal translation is used as an interpreted-translation in dictionary or other documents. For examples,

(8) EMS (Express mail service)→特快(邮政)传递; WTO (World Trade Organization)→WTO 组织→世界贸易组织→世贸组织; Wi-fi (Wireless fidelity)→无线(局域)
In example (8), the morphemic meaning compounding rationality of the original terminologies above presents some lexical image characteristics and all have abbreviations, which were used directly when the terms were introduced into China at first stage, and most of them have formed Chinese terms through morphemic literal translation. However, except the first item which was finalized as 特快传递 and shorter form 快递 soon, the other Chinese translated terms above are usually applied as interpreting words in bracketed discourses. Among them, the abbreviation WTO once was used as WTO 组织(WTO organization), so that Chinese readers who were not familiar with the original term may also know what the term mean roughly just according the attached Chinese lexical forms, though actually both WTO and the abbreviation 世贸组织 of its Chinese literal translated term 世界贸易组织 have just four syllables, and the Chinese form is more adaptive to Chinese discourses.

With a comparison of ‘Wi-fi’ with Chinese borrowing word 无线局域网, we find that though ‘Wi-fi’ has much less syllables than its Chinese translation form and is popular up till now, the abbreviation 无线网 of 无线局域网 is not a long string as well, and comparatively another Chinese brief translation form 无线保真 maintains its lexical image of original language but is not so popular as 无线网, therefore, the latter seems to be more survivable with the lexical image of some referent character. And except ‘G20’ which has been used for a short time yet and is quite brief for speaking, the other abbreviations are all accepted as letter words in ‘Contemporary Chinese Dictionary (6th edition)’ and have been used in relevant scopes, however, their Chinese word-to-word translation forms are being used without being mutual exclusively, and more easily accepted especially by those readers of the Chinese discourses spreading in outer fields.

From the analysis above, we may conclude that there is a formational force in target language which is playing role potentially always in borrowing lexical images,
though transliteration is applied tentatively at first stage of translation and introduction.

3. Some more points: how taking account of tentative explanation of lexical images and finalized lexical images in bilingual dictionary making.

Summarizing the above analysis, we may find that the lexical image composed according the lexical rationality of the original language is an important cultural characteristic; however, different languages would not be able to successfully transform all lexical images through translation, but needs to settle the problem caused by lingual difference and mutual restriction through image adjustment in bilingual dictionary making.\(^9\)

Firstly, to maintain original lexical images today when Chinese culture is spreading out to the world, Chinese compounds or phrasal words, which compose rational meaning with morphemic components and reflect Chinese cultural elements to some extent, should be transformed through morphemic literal translation, so as to be able to hold the original images and also agree with the formational rule of the target language. For example,

(10) 中国梦: China dream; 八项规定: eight-point rules; 知青: educated youth; 丢脸/丢面子: lose face; 中华（白）鲟: Chinese sturgeon/ paddlefish; 三个文明: three civilizations→ tri-civilization; 铁饭碗: iron bowl→ iron rice bowl; 丝绸之路: silk road→ silk route; 兵马俑: terracotta figures→ terracotta warriors and horses; 洋鬼子: foreign devil→ devil foreigner; 中秋: Moon cake festival→ Mid-Autumn Festival.

If investigating the composition of lexical images at morphemic level, we may find that last five cases have arrows directing to the meanings with lexical image compositions closer to the original ones, which reflect the applicable effect of lexical image transformation through morphemic translation.

Secondly, if target language lacks certain suitable lexical images, transliteration is adopted to maintain the original lingual lexical images attached with some explanations which are necessary to reveal the cultural contents of the original lexical images. For examples,

(11) 钓鱼岛: Diaoyu Island; 宣纸: Xuan paper; 毛南族: Maonan ethic; 茅台: 

---

9本节若干语料见章宜华《中国特色词在中外媒体中的使用特点及影响因素》，《学术研究》，2016年第7期。
Maotai → Maotai spirits;
太极：Taijiquan or t’ai chi ch’uan→ a kind of traditional Chinese shadowboxing→ Taiji boxing;
国子监：Guozijian → directorate of the Imperial Academy in dynamic China → Guozijian Imperial Academy.

Because there are differences between Chinese phonetic system and target language phonetic system, if large amount of Chinese transliterations entered the target language, it would certainly cause difficulty into native people’s reading, comprehending and remembering. However, the means of transliteration plus liberal translation would make the transliterated part become a complementary component in the lingual difference of target image formation. All the items in example (11) agree with this modal of translation ‘phonetic image of original word + core-word image of target language’, which are not difficult for English users who are familiar with English vocabulary to understand the lexical rationality of the translations. Among the translations, if 钓鱼岛 was translated as ‘fishing island’, it would lost the original implied meaning as a place name of China, so, the partially imitating of the original word pronunciation in target language may play an important role in maintaining the original lexical image as well.

Thirdly, it is reasonable to make a well use of lexical images of target language but not restricted to the compounding rationality of original language, in other words, it is suggested to base on the referent object to reveal the lingual difference of lexical formation in target language, so that the users of target language could approach to the original lexical images through their native lexical images, but avoiding the production of too many transliterations. For example,

(12)冬至: the Winter Solstice; 长城: the Great Wall; 大妈: Dama→ aunt/ madam (an affectionate or respectful address of elderly woman); （什锦）炒面: chow mein→(mixed) fried noodles; 坐月子: Zuoyuezi→ month confinement in childbirth; 高铁: Gaotie→ metro-liner/ bullet train; 八卦: Bagua→ Eight Trigrams; 二胡: erhu→ Chinese violin→ two-stringed (Chinese) fiddle/ erhu fiddle; 翰林: Hanlin→ literary-officials (around emperor).
In example (12), the first four terms concern common things. If the users of target language are drawn into consideration, though the lexical images of target language have different compounding rationality, they all should not be transliterated because there are correspondent native words in English, since they refer to correspondent things. For instance, ‘Eight Trigrams’ with initial capitals as a translation of Chinese term 八卦 was selected into the series of Oxford dictionary, the English word ‘trigram’ extends a sense, ‘a figure of three lines, i.e. three 爻(whole and broken lines which form the eight trigrams in ‘the Book of Changes’), in addition to the original sense ‘three letters represent one sound’. It proves that, those lexical images of target language, if similar to those of original ones, may be chosen to express the referent objects, so as to make the original lexical images merge into the target language adaptively.

Fourthly, the lexical principles and spelling rules of target language are carried on to maintain Chinese lexical images, so that the translation can agree with the derivative and compounding laws and maintain the original lexical images as more as possible in spreading Chinese cultural images. For example,

(13) 龙芯: dragon chip; 蛟龙: Jiaolong→flood dragon; 枭龙: Xiaolong→elite dragon;
单位: danwei→unit; 叩头/叩首/磕头: kowtow; 元气: Yuanqi→original Qi/ Chi (vital energy)/ vitality; 鱼米之乡: land of milk and honey→land of fish and rice; 春晚（春节联欢晚会）: Chunwan→the Spring Festival Eve Gala.

In example (13), all Chinese lexical images are almost maintained as similar word-to-word translations in target language, so that the Chinese lexical images could influence English lexical rationality. ‘Dragon (龙)’ is not a lucky animal in English, however, international society has accepted ‘dragon’ as a correspondent word of Chinese 龙, so, other Chinese compounds with 龙 (or~龙/龙~) as component should be also translated with the same lexical image of ‘dragon’. And the Chinese word 单位 has some specific feature of Chinese organizations, but the English word ‘unit’ is qualified to be its word-to-word translation because 单位 is a basic independent working unit with Chinese social characteristics, and in the interpreting phrase here, ‘unit’ is a core word. Moreover, both ‘kowtow’ and ‘Qi / Chi’ have been involved into the series of Oxford dictionaries, however, especially the last one is a Chinese cultural basic conception and has become an independent morpheme in English, but there is also
English inhere native word correspondently, we should not use its alphabet spelling form anymore in a common native discourse. So, especially, the last two cases clearly reflect that it is necessary to restore Chinese lexical images in some lexical forms of target language, which is called lexicalization.

Fifthly, transliteration should be taken as a tentative strategy to deal with difficult images, with a complement of literal translation as an interpretation to reconstruct lexical images for referent objects, so as to establish a semantic relationship between the original morphemic meanings and the lexical meaning of target language. For example,

(14) 户口： Hukou→ registered residence; 牌楼： Pailou→ decorated gateway; 高考（高等教育考试）： Gaokao→ university entrance examination; 针灸： Zhenjiu→ acupuncture and moxibustion.

In example (14), all the four Chinese words refer to things with Chinese cultural characteristics. To maintain their cultural characteristics, some media transliterated them into English. Actually, however, the cultural characteristics of things and that of lingual components belong to different cultural levels. Although these Chinese words refer to those referents which English have no correspondent words, there are word groups in English almost corresponding to those referents (e.g. the first two cases), and even morphemic literal translation can be applied to the last two cases. Actually, the China’s cultural characteristics of registered residence and university entrance examination require long discourses to describe their abundant contents. While ‘Zhenjiu’, acupuncture and moxibustion are important concepts in Chinese medicine which also need a long paragraph to explain their main connotations, but they don’t have specificity in the relation between lexical meanings and referents, so, transliteration is not necessary for transforming the original language images into target language images.

Actually, the rational structure of native lexical image is an easier symbol for native speakers to learn and remember the new conceptions than pure transliteration forms without rational lexical meanings. (a draft)

---

陶原珂《汉英词典中的文化义差及语差释义》，《深圳大学学报》2004年第2期。
The Translation of Keju Vocabulary in Chinese-English Dictionaries

Shan Ma  
Shaanxi Normal University  
mashan0509@snnu.edu.cn

Abstract

Chinese culture-specific words, always loaded with unique cultural information, hardly have any equivalents in English. Thus they pose a huge challenge to Chinese-English dictionary makers when they translate them in the dictionary. Through a case study of the translation of keju vocabulary in Chinese-English dictionaries, Charles Fillmore’s Frame Semantics is found to be helpful in tackling the problem. It is proposed in this paper that we make use of cognitive frames in the translation. Human beings, regardless of their races, categorize the world in more or less similar ways, which forms the basis for cross-cultural communication. As cognitive frames are not bound to any language, they may provide a valuable platform for cross-linguistic comparison. In order to evoke similar conceptual experiences of target users, firstly, this paper analyzes relevant data in Chinese corpora, extracts the propositional model of keju vocabulary and finds that the frame keju vocabulary fits in is “Examination”. Then this study refers to FrameNet and takes core frame elements of “Examination” as the minimum elements to be covered in dictionary translation. After that, culture-specific information of keju frame is analyzed and non-core frame elements that should be made salient are settled in this paper. Finally, this study states that dictionary makers should reconstruct the semantic frame of keju vocabulary in the dictionary and reveal the semantic frame of it as a united whole by showing its structure and inner connection to target users to enhance their understanding of the cultural information that keju vocabulary conveys in Chinese-English dictionaries.

Keywords: Frame Semantics, Chinese-English dictionary, culture-specific words, translation, keju

1. Introduction

Monolingual dictionaries provide the definition of words, and bilingual dictionaries usually translate the word from the source language to the target language. Traditional bilingual dictionaries have always made it their major task to find equivalents in the two languages. However, as a matter of fact, most full equivalents exist only in technical terms. Therefore, due to the isomorphism of the linguistic schemata and conceptual schemata between Chinese and English, Chinese culture-specific words, can hardly be translated directly into English, as there are often no ready equivalents in the
Concerning the translation of Chinese culture-specific words in Chinese-English dictionaries, in recent years, scholars in China have done plenty of research and exploration, which can be roughly divided into two groups. The first group of scholars focuses on individual words and phrase, aiming at solving the problem from bottom up. For example, Huang Jianhua, Chen Chuxiang (1997: 102) stated, through certain efforts, Chinese culture-specific words are translatable, “we should distinguish cases where transliteration, or free translation, or the combination of transliteration and free translation is applied. If necessary, notes and pictures can also be used…” Wu Jianping (1997: 150-157) developed the corresponding relation of cultural information between Chinese and English, and put forward that dictionary translation of above mentioned words should be concise and brief. Wan Jiangbo (2006: 137-147) classified Chinese culture-specific words and proposed different methods to be adopted in translating each category of words, such as free translation, transliteration, the combination of both, translation based on semantic analysis of morphemes, and translation using homophones and puns, etc. The other group of scholars tried to find out a theory that can be applied to the translation of Chinese culture-specific words in Chinese-English dictionaries generally. For example, Zhang Yihua (2009:272) states that “Equivalence in conceptual meaning is only part of the definition in a bilingual dictionary”, and “meaning representation is a multi-dimensional semantic structure.” He argues, “… in bilingual dictionaries, the headword is not defined directly in the target language; rather, it is represented by the cognitive-semantic structure hidden in the word.” Therefore, the definition in bilingual dictionaries involves the transformation of all semantic components of the word from the source language to the target language. Zhang (2006: 365) also states that “Interpretation in bilingual dictionaries …its essential feature is translation, not interpretation. By ‘translation’, I mean, it is no longer restricted to seeking for equivalents, but reconstruct the structure of the cognitive-semantic structure of the word in the source language.” Wei Xiangqing(2005 b: 38) points out that the basis of bilingual dictionary compilation is the reconstruction of the semantic system of the source language in the target language, whose fundamental purpose lies in the cognitive transformation between two sets of semantic systems in two different languages and cultures. To sum up, research on the translation of Chinese culture-specific words has developed from individual words to the whole, from micro level to macro level, and from specific practice to the working out of a general theory.

This thesis will continue the theoretic study in the translation of Chinese culture-specific words in bilingual dictionaries, from the perspective of Frame Semantics, through the case study of keju vocabulary, to further the reconstruction of semantic structures of Chinese culture-specific words in the target language.

\[1\] All the Chinese citations are translated by the author of this paper.
2. Problems in the Translation of Keju Vocabulary in Chinese-English Dictionaries

The translation of *keju* vocabulary, which causes strangeness or dislocation, in its linguistic form, is not able to effectively evoke target language users’ similar cognitive experience, making it hard to achieve the communicative function of the dictionary. In table 2, the translation of *xiucai* and *tongsheng* do not provide target language users with an effective frame for reference, and it will be difficult to activate relevant prior knowledge stored in the target language user's mind that can help them understand the two words. "Graduate" in the translation of *xiucai* may mislead target language users, making them think that *xiucai* is related to “diploma” in modern education system, as the translation cannot reflect *xiucai*’s relationship to *keju*. Likewise, “scholar” in the translation of *tongsheng* is too broad, thus fails to describe distinctive features of *tongsheng*, and consequently poses obstacles to target language users.

**Table 2 xiucai and tongsheng**

<table>
<thead>
<tr>
<th>秀才</th>
<th>Cultivated talents-a graduate of the former first degree or <em>hsiuts'ai</em></th>
<th>Mathews’</th>
</tr>
</thead>
<tbody>
<tr>
<td>童生</td>
<td>scholar without a <em>xiucaititle</em>[in the Ming and Qing dynasties]</td>
<td>New-Century</td>
</tr>
</tbody>
</table>

In general, to understand the above two words, Chinese native speakers have amassed considerable prior knowledge in the default position, so, mandarin monolingual dictionary definitions do not have to explain in great detail. However, for non-Chinese-native-speakers, to facilitate their understanding of Chinese culture-specific words, Chinese-English dictionaries need to first specify the propositional model, or the genus, or the prototype, so as to providetarget language users with a frame for reference, to reduce the efforts that non-Chinese-native-speakers use when they understand the words (Li Yunlong 2010: 68).

Due to the usually adopted alphabetic order of words in dictionaries, the construction of a complete system does not get enough attention, as a result, definitions of words are often isolated, lack of a clear, systematic and interconnected framework. Monolingual dictionary users are less influenced by the negative effects of this practice in dictionary making, because in mother tongue users’ cognitive experience, there is a basic framework of this kind of things, from which they can get reference to the relationships between certain words. However, when consulting the dictionary for Chinese culture-specific words, target language users, to a large extent, rely on the information carrier—the translation of the words, to help them build up the cognitive framework.

For instance, according to Encyclopedia of China (2009), *zhuangyuan, bangyan*,...
and *tanhua*, refer to the top 1, top 2, and top 3 at the palace examination respectively. But in Chinese-English dictionaries, the translations of the three words are not consistent in form, nor are they unified in wording, so they fail to reflect their interconnections (Table 3).

**Table 3 zhuyuan, bangyan, and tanhua**

<table>
<thead>
<tr>
<th>Headword</th>
<th>Translation</th>
<th>Dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>zhuangyuan</td>
<td>highest graduate of the Hanlin Academy</td>
<td>Mathews’</td>
</tr>
<tr>
<td>bangyan</td>
<td>the “eye of the list”—a graduate who came out second in the list at the former palace examination</td>
<td></td>
</tr>
<tr>
<td>tanhua</td>
<td>title of the third graduate on the list at the finals for the Hanlin Academy</td>
<td></td>
</tr>
<tr>
<td>zhuangyuan</td>
<td>the NO. 1 of the national examinations</td>
<td>Lin Yutang’s</td>
</tr>
<tr>
<td>bangyan</td>
<td>NO. 2 at palace examinations (“on bull’s-eye”)</td>
<td></td>
</tr>
<tr>
<td>tanhua</td>
<td>number three in former national examinations</td>
<td></td>
</tr>
<tr>
<td>zhuangyuan</td>
<td>Number One Scholar [title conferred on the one who came out first in the highest imperial examination]</td>
<td>CED-3 (Wu)</td>
</tr>
<tr>
<td>bangyan</td>
<td>the second place at palace examinations</td>
<td></td>
</tr>
<tr>
<td>tanhua</td>
<td>number three in national civil examinations (in feudal China)</td>
<td></td>
</tr>
</tbody>
</table>

Similarly, Chinese-English dictionaries have the same problems in the translations of *xiangshi, huishi, and dianshi* (table 4). First of all, in CED-3 (Wu), the translation of *huishi* is not accurate, and "the examination for the selection of juren (举人) should be changed into “the examination for the selection of gongshi (贡士) from juren(举人)". In addition, the hierarchy of the examination is not clear. If *xiangshi* is translated into “provincial examination”, *huishi* should be translated into “national examination”, and *diangshi* into “palace examination” plus the enclosed note “the highest imperial examination”. In this way, the hierarchy of *keju* can be clearly represented in the translation, and target language users can tell the relationship between the three words easily at a glance. Likewise, the translations of *xiangshi, huishi, and dianshi* are not consistent in the language form either, which requires more cognitive efforts form target language users to see their relationships, or even causes failure in identifying the hierarchical relationships in the three words.
<table>
<thead>
<tr>
<th>Headword</th>
<th>Translation</th>
<th>Dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>乡试</td>
<td>provincial examination (for the selection of juren(举人) from xiucai(秀才), under the Ming and Qing civil service examination system, held triennially in the provincial capital)</td>
<td>CED-3 (Yao)</td>
</tr>
<tr>
<td>会试</td>
<td>metropolitan examination; general examination (under the Ming–Qing civil service examination system, the examination for the selection of juren（举人）, held triennially in Beijing, the national capital)</td>
<td></td>
</tr>
<tr>
<td>殿试</td>
<td>palace examination (the highest imperial examination, presided over by the emperor)</td>
<td></td>
</tr>
<tr>
<td>乡试</td>
<td>(during the Ming and Qing dynasties) a civil service examination held at a provincial capital once every three years</td>
<td>New-Age</td>
</tr>
<tr>
<td>会试</td>
<td>a general examination (for the successful candidates of all the provincial civil examinations held once every three years in the capital during the Ming and Qing dynasties) →会元</td>
<td></td>
</tr>
<tr>
<td>殿试</td>
<td>a palace examination—the final imperial examination presided over by the emperor in the palace</td>
<td></td>
</tr>
</tbody>
</table>

Because of the alphabetic order that the dictionary adopts, the above words are placed on different pages. When users look up the dictionary for an individual word, the problem may not be so obvious. But suppose the user is reading an article about keju in ancient China, the co-occurrence rate of the above words is inevitably high. Or suppose a native speaker of Chinese is using Chinese-English dictionaries to translate an article about keju from Chinese to English. In such cases, the lack of integrity and hierarchy in dictionary translation will aggravate the problems in cross-cultural communication.

3. Dictionary Translation of Keju Vocabulary from the Perspective of Frame Semantics

Meaning comes from human experience of life. It is the result of human interaction with the outside world. Human beings’ cognition of the outside world has to undergo a conceptualization process, and semantic structure is a product of human mind, derives from and reflects the conceptual structure of human beings. The study of the conceptual framework and the answer how language use reflects the conceptual
framework, is the premise of dictionary definition (Tian Bing 2003: 345). Bilingual dictionaries, which involve two languages and cultures, should, in theory, reflect the cognitive structure of the outside world. To put aside the differences among language groups within the same language, the two conceptual structures represented by Chinese and English are not of parallel relationship, but in net-like distribution (figure 1). When separate words in two languages represent the same concept, the lines in the bilingual structure will meet, sharing a node, and the concept in the source language can be easily identified by target language users. When separate words in two languages represent the similar concepts, the bilingual conceptual structures will be parallel, coming infinitely close to each other, in which case, target language users can make use of their conceptual structure in their mother tongue to help them understand the concept in the source language. Then again, when the concept represented by the source language does not exist in the target language, namely, when the word is culture-specific, we may assume that there is a point in the conceptual network of the target language, which, in appearance, is discrete and without any connections, and its distance from the reference point is random. However, empirical study in “categorization” and “prototypes” show that human cognition shares something in common that sets limit to the range and scope in which languages differ (Wang Renqiang 2006: 89). Moreover, languages differ only in forms, not in the essence of the mind (Gao Yihong 2000). Therefore, people speaking different languages with different cultures all share the basis of understanding and communication. That is to say, by making efforts, human beings can understand an alien concept anyway. In other words, dictionary compilers need to take measures to make the two discrete points acquire connection with each other in the user’s bilingual conceptual network—to find a reference point in target user’s conceptual network and draw it to the concept in the source language as close as possible. In this way, target language users can understand the concept in the source language via the reference point. In fact, this is also the way how the communicative function of dictionaries is realized. To put it another way, the realization of the communicative function of bilingual dictionaries relies on the basic cognitive semantic framework human beings share.

![Figure 1 Bilingual Conceptual Network](image)

(Full line represents source language, and dotted line the target language.)
3.1 Frame Semantics and FrameNet

Frame semantics, put forward by Charles J. Fillmore in 1970s, is one branch of cognitive linguistics. “Frame” is “any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits” (Fillmore, 1982: 373). For instance, to understand the word “buy”, we need to understand the frame “Commercial Transaction” that “buy” fits in. At the same time when “buy” is understood, other concepts like “buyer”, “sell”, “seller”, “money”, “pay”, “cost” are also activated and the frame “Commercial Transaction” is evoked.

FrameNet is a lexical database housed at the International Computer Science Institute in Berkeley, California. Up till November 11th, 2016, FrameNet covers roughly 13,000 lexical units (word senses) for the core English lexicon, associating them with roughly 1,200 fully defined semantic frames; the frames are documented with 200,000 manually annotated examples.

In Chinese-English dictionaries, the translation of Chinese culture-specific words depends on the semantic frame that they fit in, which requires that compilers locate the semantic frame of the word before translating them. To do it, we need to analyze related empirical schemata of Chinese native speakers and to summarize, abstract and idealize the propositional model of the headword in the source language and describe it in the target language. Then, we need to evoke the empirical schemata and background frame with the name of the frame in the target language. Whether the above mentioned steps are successful or not depends on the extraction of the propositional model and the match of it with the semantic frame in the target language.

3.2 Propositional Models and Semantic Frames

Propositional Models are semantic features that are not directly described but can be concluded by analysis, reasoning, abstraction, or empirical study (Zhang Yihua 2002: 188). For example, cognitive models of “breakfast” vary in different cultures, but with its propositional model unchanging--the “the first meal of the day, eaten in the morning.” Again, “east wind” in Chinese culture and “west wind” in British culture share the same propositional model—“the spring breeze”, although their cognitive models are different due to different geographical locations of China and UK (Wu Ping 2006: 45). As for the categorization of concepts, the propositional model is an ideal cognitive model. Chinese and English are two completely different languages with distinctive cultures. As a result, Chinese culture-specific words hardly have any equivalents in English, but they are not untranslatable. In fact, as long as we study the category they fit in and find out their propositional models, we can see the inside convergence through the superficial differences. To extract the propositional models of Chinese Culture-specific words, Chinese-English dictionary compilers need to identify and define the semantic frames they fit in.

In the context of Frame Semantics, the essence of dictionary definition is to reveal and describe the semantic frame of the word (Tian Bing 2003: 342). “Since Frames are supposed to be conceptual structures not tied to any particular language, and given the
great amount of work that has already been devoted to identifying numerous frames and their interrelationships, they may seem to provide a valuable platform for cross-linguistic comparison” (Goddard 2011: 81). Of course, new frames can be invented where necessary, especially in highly culture-specific domains.

3.3 Core Frame Elements and Minimum Definition Elements

Propositional models are shared by many nations, so Chinese culture-specific words can find the same propositional models in English. It is their cognitive models that truly reflect the culture-specific meanings of these words. When translating these words, it is suggested that Chinese-English dictionary compilers start from covering their propositional models first, and then further analyze their cognitive models that need to be made salient, so as to describe the culture-specific meaning accurately.

In Chinese Dictionary, from the definition of keju, one can easily tell that its propositional model is “examination”, which corresponds to the semantic frame “Examination” in FrameNet. Therefore, when translating keju related words, dictionary makers may set it as a basic reference system, to evoke similar cognitive experience of target language users. In this way, the culture-specific information of keju is de-alienated as much as possible, and the communicative function is half-realized. In FrameNet, Core Frame Elements (Core FEs) are Examination, Examinee, Examiner, Knowledge, and Qualification. Non-Core Frame Elements (Non-Core FEs) are Manner, Means, Place, Purpose, and Time (Table 5).

<table>
<thead>
<tr>
<th>Core FEs</th>
<th>Non-Core FEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examination</td>
<td>Manner</td>
</tr>
<tr>
<td>Examinee</td>
<td>Means</td>
</tr>
<tr>
<td>Examiner</td>
<td>Place</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Purpose</td>
</tr>
<tr>
<td>Qualification</td>
<td>Time</td>
</tr>
</tbody>
</table>

Table 5 Frame Elements of Examination

As the propositional model of keju is the same as that of the semantic frame “Examination” in FrameNet, the core FEs should be the minimum definition elements in a Chinese-English dictionary, which solves one aspect of the problem. The other aspect is that minimum definition elements must represent the attributes of the concepts defined, because an individual’s calculation of the attributes is a kind of gestalt perception—to integrate the most important and most noticeable parts of the object into a complete visual-mental representation. Only in this way can the language user identify the category of it (Li Yunlong 2010: 69-70).
3.4 The Culture-specific Meaning of “Examination” in Keju Vocabulary

After analyzing the literature and copra data on keju, it is found that the culture-specific meanings of keju are represented in the following four aspects:

(1) Knowledge tested: the writing of eight-part essays and trial poems. The subjects of the eight-part essays are from Four books and Five classics.
(2) The qualifications conferred: xiucai, juren, jinshi, etc.
(3) Purpose: To select officials for the feudal rule.
(4) Time: From Sui and Tang to the end of Qing dynasty, abolished in 1905.

Therefore, when the semantic frame "Examination" is used in the translation of keju vocabulary, among its Core FEs (Table 5), “Knowledge” should highlight that Confucian Classics are the content of the test, and “Qualification” should highlight the hierarchy of xiucai, juren, and jinshi—titles conferred by the government. However, the cognitive model of keju differs from that of “Examination”. That is to say, the content, manner and purpose are different within the two different cultures. Just as Atkins and Rundell (2008: 427) said, “Culture-specific words are often loaded with cultural associations and ‘form part of the native speaker’s knowledge about words.’ A definition should contain enough encyclopedic information that the users can identify and distinguish one headword from other semantically related headwords.” Due to the limited space in the dictionary, lexicographers need to choose carefully the encyclopedic knowledge to be included in the entry, based on accuracy and economy. We can start by defining the core frame elements of the semantic frame, and then consider highlighting some non-core frame elements to embody its culture-specific meanings. To be more specific, in the semantic frame constructed for keju vocabulary, we may start by defining the core frame elements of “Examination”, and then make salient non-core frame elements: “Purpose”-- to select officials for the feudal rule, and “Time”-- from Sui and Tang to the end of Qing dynasty. In this way, the semantic frame of keju can show the culture-specific meaning of “Examination” frame in FrameNet, which not only includes the minimum definition elements that guarantee the evocation of similar cognitive experience in target language users, but also highlights some non-core frame elements that embody the culture-specific meaning, in order to convey the cultural information accurately and effectively to target language users.

<table>
<thead>
<tr>
<th>Core FEs</th>
<th>Non-core FEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examination</td>
<td>Manner</td>
</tr>
<tr>
<td>Examinee</td>
<td>Means</td>
</tr>
<tr>
<td>Examiner</td>
<td>Place</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Purpose</td>
</tr>
<tr>
<td>Qualification</td>
<td>Time</td>
</tr>
</tbody>
</table>
Through comparison and analysis, current Chinese-English dictionaries use “imperial” in the translation of kejù to define the time span in which kejù was practiced. Although not specific enough, it can help target language users to distinguish it with examinations in today’s world. Compared with early Chinese-English dictionaries, it is more accurate than the previous translations like “the old examination system” in Mathews’, or “the civil service examination system in old China” in Liang Shiqiu’s. According to Merriam-Webster’s Collegiate Dictionary (Eleventh Edition), the definition of “imperial” is “of, relating to, befitting, or suggestive of an empire or an emperor”. In other words, with the use of “imperial”, target language users can easily conclude that kejù was an examination system in the imperial era in China. However, “imperial examination” does not show the purpose of the examination, which bears precisely the culture-specific information that can by no means be neglected. So it needs improvement. Wikipedia translates kejù into “Imperial Examination” or “Civil Service Examination”, the former with its emphasis on the time, and the latter with its emphasis on the purpose. Kejù is a system in ancient China that was used to select officials for the feudal rule, and Chinese-English dictionaries should give it a clear description, to avoid misleading target users to confuse it with modern civil service examinations.

Although existing Chinese-English dictionaries have already provided the semantic frame “Examination” in the translation of kejù, which serves as a reference frame for target language users, the translation should, as much as possible, reflect the culture-specific meanings, and at the same time, match it with target language user’ cognitive experience, so as to optimize the understanding of target language users to the greatest extent.

To sum up, in Chinese-English dictionaries, the translation of kejù vocabulary should not only represent the semantic frame “Examination”, but also identify the Time and Purpose. The translation “Imperial Civil Service Examination System” satisfies all the above conditions, thus should be adopted. Some dictionaries (Liang Shiqiu’s, New-Century, New-Age-2) have already done so, but without an apparent system, nor has it been used consistently in all words in the kejù frame. Moreover, dictionary compilers cannot expect target language users to automatically relate a word in kejù vocabulary they see in the dictionary to the “Imperial Civil Service Examination System” without sufficient hints just like native speakers of Chinese do. For instance, in New-Age-2, the translation of kejù is “Imperial Civil Service Examination System”, while the translations of tongsheng, xiucai and juren only use “imperial examination”. Likewise, the translation of xiangshi uses only “civil service”, not “imperial civil service examination”. It is not hard to imagine that, without proper reference, target language users cannot associate them with “Imperial Civil Service Examination System” automatically when they see them on different pages of the dictionary. As a result, there might be confusion, and target language users might take them for different examination systems mistakenly.
4. Conclusion

Through the case study of keju vocabulary, this paper argues that, when translating Chinese a culture-specific word, Chinese-English dictionary compilers should first indentify its propositional model, locate the semantic frame it fits in, and make the semantic frame a starting point of translation, so as to evokesimilar cognitive experience of target language users, and help them save the twists and turns in understanding. In addition, Chinese-English dictionaries should also provide a systematic, consistent description of the semantic frame, and use certain language hints to show the structure and hierarchy of the frame, to make up for the isolation of words resulted from the alphabetic order. In this way, Chinese-English dictionaries will be more successful in presenting the inherent semantic network of culture-specific words, under the help of which, target language users can not only see the “trees” but also the “forest”.

References

Dictionaries


Other References


万江波. 2006. 双语词典的翻译研究. 上海：复旦大学出版社.

王仁强. 2006. 认知视角的汉英词典词类标注实证研究. 上海：上海译文出版社.


魏向清. 2005b. 双语词典译义研究. 上海：上海译文出版社.


吴平. 2006. 文化模式与对外汉语词语教学. 中央民族大学博士论文.


章宜华. 2009. 语义·认知·释义. 上海外语教育出版社.
### Abbreviations of Dictionaries Used in this Paper

<table>
<thead>
<tr>
<th>Dictionaries</th>
<th>Compilers</th>
<th>Abbreviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathews’ Chinese-English Dictionary</td>
<td>R.H. Mathews</td>
<td>Mathews’</td>
</tr>
<tr>
<td>Chinese-English Dictionary of Modern Usage</td>
<td>Lin Yutang</td>
<td>Lin Yutang’s</td>
</tr>
<tr>
<td>A New Practical Chinese-English Dictionary</td>
<td>Liang Shiqiu</td>
<td>Liang Shiqiu’s</td>
</tr>
<tr>
<td>The Contemporary Chinese Dictionary (Chinese-English edition)</td>
<td></td>
<td>CCD-CE</td>
</tr>
<tr>
<td>A New Century Chinese-English Dictionary</td>
<td>Hui Yu</td>
<td>New-Century</td>
</tr>
<tr>
<td>The Chinese-English Dictionary (3rd edition)</td>
<td>Wu Guanghua</td>
<td>CED-3 (Wu)</td>
</tr>
<tr>
<td>A Chinese-English Dictionary (3rd edition)</td>
<td>Yao Xiaoping</td>
<td>CED-3 (Yao)</td>
</tr>
<tr>
<td></td>
<td>Cheng Zhenqiu</td>
<td></td>
</tr>
</tbody>
</table>
J. Huang’s *Grand Dictionnaire chinois-français contemporain* (2014) and His Lexicography

Bing Tian
The Center for Lexicographical Studies, Guangdong University of Foreign Studies
201510083@oamail.gdufs.edu.cn

Abstract

*Le Grand Dictionnaire Chinois-Français Contemporain* was published in China in 2014. It is at present the largest dictionary of its kind and contains a number of innovative features. To know a dictionary and its design features, it helps to know its editor and his academic work. This is especially true of this Dictionary in that dictionary-making projects in China are in most cases chief-editor-oriented. Jianhua Huang, the chief editor of this Dictionary, accumulated rich experience in lexicographical practice and spent two decades exploring all aspects of theoretical lexicography before starting to compile the Dictionary. This paper will first address Huang’s lexicography as a preparation and underpinning for the Dictionary and will then examine the Dictionary itself. The examination is carried out at two levels. At the macrostructural level, the focus is mainly on its coverage and headword selection, in comparison to its sole source dictionary, *A Modern Chinese Dictionary* (6th edn, 2012). At the microstructural level, the focus is on sense translation, illustrative example allocation and translation, and grammatical and pragmatical labelling, in comparison with *A Chinese-English Dictionary* (3rd edn, 2010), a well-established Chinese-English counterpart. Our examination reveals that the Dictionary is lexicographically well-designed in all major respects and that it has successfully achieved the goal of combining a reference dictionary and a learner’s dictionary into one.


1. Introduction

*Le Grand Dictionnaire Chinois-Français Contemporain*, i.e. The Grand Contemporary Chinese-French Dictionary (*GCCFD* for short) (汉法大词典) (*hàn fǎ dà cídiǎn*) was published in 2014. Its editor-in-chief is Jianhua Huang, a professor at the Center for Lexicographical Studies, Guangdong University of Foreign Studies. He is also the editor of *A New Concise French-Chinese Dictionary* (1983), the author of two monographs, *Studies in Lexicography* (1987; 2001) and *An Introduction to Bilingual Lexicography* (co-authored with Chen: 1997; 2001), and the translator into Chinese of four French academic classics. The *GCCFD*, the latest and most voluminous work of its kind, could, in a sense, be viewed as the culmination of Huang’s four-decade pursuit of lexicography. It is also a reflection of his lifelong experience in learning, teaching, and researching the French language, and a distillation of his first-hand experiences in translation and
interpretation between French and Chinese, including a four-year period as a full-time senior translator at UNESCO, shortly after China opened its doors to the outside world in the late 1970s.

In the past four decades, China has witnessed a dramatic change in its economic development and has steadily broadened its cultural communication with the outside world, relying on the flourishing of foreign language learning and teaching. Bilingual dictionaries, as a by-product of that process, have also undergone a steady and unprecedented growth. The GCCFD is representative of this change, in that both the dictionary itself and the lexicography it embodies, as well as its editor, are all key parts of the state-of-the-art of bilingual dictionary-making and bilingual lexicographical studies in China.

The GCCFD, a large-scale bilingual dictionary with French as a minor foreign language in China, bears some noticeable features that distinguish it from its counterparts in the area of Chinese-English (C-E) dictionaries. For instance, its target audience in China is much smaller, which means the dictionary market will sustain only one or two of its kind, unlike the huge market for C-E dictionaries, which has already nourished five large-scale dictionaries. However, GCCFD also has some common ground with its C-E counterparts. Since they have the same source language, they can be compared against the same benchmark: how well they depict the vocabulary of the contemporary Chinese language. It is thus necessary for this study to refer both to the Chinese monolinguals and to the C-E bilinguals.

2. GCCFD’s Predecessor and its Source Dictionary


The CFD project was initiated at the 1975 Dictionary Meeting, an official meeting held at Guangzhou for the nationwide planning of dictionary compilation and publication, concerning chiefly philological dictionaries, both Chinese monolinguals and foreign-Chinese or Chinese-foreign bilinguals. The CFD’s sole source for the description of the Chinese language was the Modern Chinese Dictionary (MCD) – first its ‘trial edition’ (1965) and then its first edition (1978). The CFD is a medium-sized dictionary, with 5,300 single-character entries and 50,000 multi-character sub-entries, roughly the same as the MCD-1 (1st edition), which has 56,000 sub-entries.

The CFD is only the titular predecessor of the GCCFD, because of the large time span of 23 years between their publication. The GCCFD, in fact, has 10,786 single-character entries and over 100,000 multi-character sub-entries, occupying 2,206 pages (as compared with the 968 pages of the CFD).

The GCCFD project was initiated and sponsored by the Foreign Language Teaching and Research Press (FLTRP) and it is one of the Press’s ambitious projects that aim to produce large-scale bilingual dictionaries covering the world’s major languages. The project was formally started in 1998 and completed in 2014, thus spanning sixteen years, long enough for its sole source dictionary, the MCD to have
undergone three editions (2002; 2005; 2012), each with an intake of new entries (1,200, 7,200 and 3,000 respectively). The project schedule for the GCCFD had to be adjusted and readjusted accordingly. In the process, the compilers seemed to have experienced time and again the unique journey of dictionary-makers: that of ‘chasing the sun’. The dictionary size was also enlarged: from the planned 6 million Chinese characters to 7.2 million upon completion, so as to ‘capture the basic situation of and the latest changes in the contemporary Chinese language’ (Huang 2014: 6; Huang & Zhang 2000 [1998]).

3. Preparatory Work and Planning

The role of the editor-in-chief is of crucial importance for the making of a dictionary and this is especially true as regards current lexicographical practice in China. In most cases, Chinese dictionary projects are chief-editor-oriented. The preparatory work and planning by the editor-in-chief is decisive to the fate of a dictionary – its timely publication, whether it meets the expected standard, and its success in the market.

3.1 A New Concise French-Chinese Dictionary (1983): some insights gained

Huang once recalled his embarking upon the career of a practising lexicographer:

*I had not even heard of the word lexicography until 1979. In fact, when I was in charge of the compilation of A New Concise French-Chinese Dictionary, I was in reality as well as in name ‘learning the art of war from fighting actual battles in a war’, i.e. learning the work of dictionary-making in the process of making it.* (Huang 1992: 9)

After its completion, Huang reexamined his experience and editorial work in a paper published in 1979 in the first issue of Lexicographical Studies, the Journal of ChinaLex. The paper was titled ‘A tentative exploration into the headword list, sense definition, and illustrative example of a French-Chinese dictionary’ (Huang 1979: 143-154). It is an in-depth reflection upon and a distillation of his lexicographical practice and lexicographical project-managing experience, which in a sense foreshadows the GCCFD.

The article directly addresses the three core parts of a dictionary: its headword selection, its sense definition, and the sorting out and allocation of its illustrative citations. First, the headword selection. The coverage of a bilingual dictionary can directly borrow from that of a monolingual dictionary for native speaker. Thus, a lot of time and labour can be saved which would otherwise be spent in collecting citations to determine whether a given word should be included, whether the word is polysemous, which one of its senses is most frequent, which usages are most typical, and what factors should be considered decisive when it comes to the inclusion and ordering of particular words and their senses. In choosing the source dictionary, these factors are necessarily already taken into account. For the Concise French-Chinese Dictionary (the Concise for short), the predecessor of A New Concise French-Chinese Dictionary (the New Concise), its sole source dictionary was a small French-Russian dictionary published in the former Soviet Union. The Concise had adopted ‘quite a few obsolete or outdated words and phrases no longer in use in present day French’, ‘quite a few unnecessary French “loan words” from Russian’, and it had neglected ‘a large number of common French words’ (Huang 1979: 143-145).
The second issue is the sense definition, or to be more specific, sense translation. Huang is one of the earliest Chinese scholars to use 译义 (yiyi: sense translation) for bilingual lexicography, reserving 定义 (dingyi: sense definition) and 释义 (shiyi: sense explication) principally for monolingual or general lexicography. Such a treatment highlights the essential difference between making a bilingual and a monolingual dictionary. The fundamental task of bilingual dictionary-making is to translate, i.e. to find an expression in the target language equivalent to the headword in the source language. One common error frequently encountered occurs when the lexicographer coins or invents an equivalent while in fact there is one which he or she has merely failed to detect. Another problem is that the equivalent available is not exact but exhibits nuances in connotation or usage that need to be resolved or explained in sense translation. A third problem is the lexical gaps between the two languages, which call for new naming and novel sense expounding. Such problems are found not only in common words but in words of a technical nature from different fields. Huang also addresses five types of errors in sense translation. Some of these are common with translation in general, e.g. hasty translation upon reading the word literally, being satisfied with nearly exact translation, and not being meticulous enough with the wording in Chinese. Others are peculiar to bilingual dictionary-making (Huang 1979: 149-151), e.g. ‘adding word meanings that are nonexistent in the source language’ and ‘transporting “equivalents” from some French-Japanese dictionaries’. When there exists no strict equivalence between two languages, one preferred way of treatment is to add notes whose function is ‘to highlight some necessary restrictions on sense translation’. Another type of note is introduced to explain or advocate the editor’s opinions. The usefulness of this latter type is doubted by Huang and, consequently, he recommends applying them with caution.

The third issue concerns the illustrative examples. In this part, Huang points out three types of error in sorting out and arranging examples. The first type is that the headword in an example sentence bears a different meaning from that which it is intended to illustrate. The error mainly comes from the practice of ‘emphasising the multiple and different sources for illustrative example collection [to reduce the risk of plagiarism] but with less attention given to discrimination before adopting them for use’ (Huang 1979: 152). The second type is the mismatching of parts of speech between the headword’s labelling and its use in the example sentence. The third type is that the illustrative example violates the norms of usage in the target language, which might result from ‘the practice of inventing illustrative examples in making a bilingual dictionary’ (Huang 1979: 153). As to the function and assessment of illustrative examples, Huang has laid emphasis on their role in supplementing the deficiency in sense definition, illustrating collocations, or forming idiomatic multi-word expressions.

When supervising the compilation of the New Concise, Huang made extensive explorations into dictionary use, bilingual translation, and comparative analysis of the French and Chinese languages, but at the same time he conducted little research into lexicographical theory. It is from the actual experience of dictionary-making that he gained a better understanding of the process, a more comprehensive and systematic grasp of the lexis of the French language, and a more in-depth comparative perspective on the vocabularies of the French and Chinese from the point of view of translation. This provided him with a solid foundation for exploring theoretical lexicography in both general and specialised fields in the following two decades.
3.2 ‘On the Dictionary’ series (1983-85) and ‘A Plan for recompiling the New Concise’ (1987a)

In the following decade, Huang endeavoured to explore the general theory of lexicography. He published two dozen papers in the journal Lexicographical Studies. The most influential ones are the fourteen articles in the ‘On the Dictionary’ series (1983-85), later integrated into his monograph Studies in Lexicography (1987b). Theoretically better equipped, he turned to planning to recompile the New Concise. There are six points that he develops in this connection (Huang 1987a: 1-8):

1) Observing the principle of strict synchronicity
2) Adopting the method of distributional analysis
3) Perceiving the vocabulary as an organic whole
4) Revealing word function in particular contexts
5) Translating citations for functional equivalence
6) Highlighting the specific information of cultural words

We are going to elaborate on these six points in the following discussion.

3.2.1 The principle of strict synchronicity

In practice, as far as headword selection, sense definition, and illustrative examples are concerned, this principle requires that obsolete headwords, outdated word senses, and old-fashioned citations should all be discarded, and etymological information excluded, and whether some headwords are to be treated separately or together, should be assessed from a synchronic perspective (Huang 1987a: 3). This desideratum was proposed just to address the misconception then popular among dictionary-makers that ‘the more data put into a dictionary, the more beneficial it is to its users, resulting in an ardent pursuit of extensive collection of diverse lexical data, and a large proportion of entries being packaged with mixed information of various kinds – old and new, common and rare’ (Huang 1987a: 3). To review the principle three decades later, we can see more clearly that to emphasise strict synchronicity for dictionary-making was of both practical and theoretical significance at that time.

For a long time before opening its doors to the outside world in the late 1970s, China had been isolated and its foreign language teaching was principally restricted to the written language. The written materials were, in large part, outdated and the materials available to dictionary-makers were basically the same, lagging far behind in capturing the changes in the ‘foreign’ languages in use outside China. Owing to the relative normativeness of written grammar and the stability of lexical meanings and usages, the outdatedness of a dictionary’s content could easily be ignored when used only in learning or translating written materials. The outdatedness and bookishness of the foreign languages widely taught and learned during that period became more noticeable when the time came for international communication, especially for face-to-face interaction. Huang was fortunate to have worked as a senior official in translation in UNESCO for four years (1977-81) and his recognition of this defect in foreign language teaching and learning in China could not have been clearer. That the principle of strict synchronicity is put forward as the first principle for dictionary-making proves how much it meant to him.
3.2.2 Distributional analysis

If the desideratum of synchronicity is to address the practical problems in dictionary-making, then the purpose of adopting distributional analysis (DA) is to seek a better motivated methodology for lexicographical work. Huang, siding with structuralist linguists, emphasises that:

_For lexical semantic description [by DA], the meaning of a word should not be dealt with in isolation but in relation to other words, for a word acquires its meaning when it co-occurs with other words and forms particular paradigmatic and/or syntagmatic relations with them. The meaning and usage of a word can only be revealed through dissecting and analysing these relations._ (Huang 1987a: 3)

With the adoption of DA, a variety of problems and errors could be picked out from the _New Concise_ and subsequently solved and corrected by means of illustration. For instance, the French verb _estimer_ is transitive. Its first two senses are: (i) to think highly of; (ii) to assess, to estimate’. From the corresponding Chinese equivalents of these two senses, as depicted in the _New Concise_, one gets no idea about what types of word could serve as the objects for each sense. In other words, the distributional restrictions for the object position of the verb _estimer_ in these two senses are not specified. In fact, the object of the first sense could normally only be a noun designating a human being, and that of the second sense, a noun designating a non-human. Such restrictions should be identified through DA and be specified for each sense or usage in a bilingual dictionary.

DA can also be applied to identifying and arranging the senses of polysemous words (Huang 1985; 1987a). For sense identification, the analysis mainly focuses on the following discovery procedure: i) identifying the typical distributional slots or patterns of a headword in a particular part of speech; ii) establishing whether the use of the word in each pattern could be substituted with a different word without changing the meaning of the sentence; iii) examining the substituted words to see whether they are synonymous. If the answer is yes, the two patterns may share the same sense; if not, the two are probably distinct senses. As an example, consider the Chinese word _气候_ (_qihou_: ‘climate’, both literally and figuratively) in the following three sentences (Huang 1985: 41):

Xinjiang diqu qihou ganzao. (qihou → tianqi)
Xinjiang region climate arid. (climate → weather)

_In the Region of Xinjiang, the qihou (= climate) is arid. (qihou → climate)_

Dangshi de zhengzhi qihou ling ren zhixi. (qihou → qifen)
Then ‘s political atmosphere make men suffocate (atmosphere → conditions)

_The political qihou (= atmosphere) was suffocating then. (qihou → atmosphere)_

Wo kan zhe haizi cheng buliao shenme qihou (qihou → dashi)
I see this kid to be not as what [affect] climate (climate → big matter)
_I do not see this kid as qihou (= influential) [when he grows up]. (qihou →_
The same Chinese word qihou could be used as a noun in different distribution slots: i) as subject with ‘weather conditions’, i.e. ‘arid’ as subject complement; ii) as subject with ‘political’ as a pre-modifier and ‘suffocating’ as subject complement; iii) as object complement indicating the [potential] personality of the object (as a human being). The same word qihou can be substituted with ‘weather’, ‘atmosphere/ conditions’ and ‘influential’ respectively. The three substitute words can in no instance be synonymous and/or exchangeable. Thus, the conclusion can be drawn that qihou has three distinct senses.

As for sense arrangement, DA is especially useful for polysemous verbs. The sense arrangement thus follows a ‘structural sequencing’, i.e. the order of the combinatorial elements of the verb: the subject, the particle (preposition, adverb), and the object. This is well illustrated in the article (No. 14 of the ‘On the Dictionary’ series).

3.2.3 The vocabulary as an organic whole

This desideratum addresses the paradigmatic or substitutional relations a word may have, either formally or semantically. Paradigmatic relations may be presented in a dictionary in the following three ways: i) the derivatives and compounds stemming from the same headword may be listed as its sub-entries or run-on entries, thus forming an embedded bi-level macrostructure; ii) near synonyms and related words may be provided for each sense so as to add the extra function of a thesaurus; and iii) bound morphemes, such as prefixes and suffixes, may be listed as headwords, enjoying the same status as free morphemes, which makes it possible for a dictionary to embrace the morphology of the language. These three design features will benefit the dictionary users in learning the various linguistic relations between words either formally or functionally, overcoming the inherent defects of alphabetical listing of headwords in most of today’s dictionaries. This desideratum resonates with systemic functional grammar, which focuses on revealing the meaning potentials of words, the unity of meaning and grammar, and the systematicity of the vocabulary of a language (Halliday 1985).

3.2.4 Word function in particular contexts

Huang observed closely the rapid development of linguistic studies on context and pragmatics in the 1970s and early 1980s, attempting to make use of them in informing the practice of bilingual dictionary-making. He considered that ‘in traditional lexicography, a word is normally considered as an abstract entity and its description is carried out purely on logical grounds’, and as a result ‘the word meanings a learner has commanded [...] are mostly free from contexts’. For the improvement of a dictionary, ‘a word should be portrayed in typical contexts where its meaning, usage and functions could be much better captured as elements of full and authentic sentences’ (Huang 1987a: 6). His high appreciation of complete sentences as illustrative examples and his emphasising the role of context and the functions of a word in full sentences have a resounding echo in two prominent dictionaries published that year in Britain, namely Collins COBUILD (1st edition) and LDOCE-2. In the former, all the illustrative examples are full sentences, and in the latter ‘the achievements of the newly emerging
pragmatics have been fully embraced at all levels of linguistic information presentation’ (Tian 1999: 74).

The shift of focus from form and structure to function and context, and further to the speaker’s meaning and implication in particular contexts was chiefly motivated by the demands of face-to-face verbal communication in the same language and culture, as well as in cross-lingual and cross-cultural communication. It happened worldwide after WWII and nationwide after China’s opening of its doors to the outside world. To meet the novel need of cross-lingual face-to-face interaction, Chinese lexicographers had to urgently renew their conception of dictionary-making. ‘Revealing word functions in concrete contexts’ was proposed as a way of targeting the new need in language learning in a new mode of society and in a new world of verbal communication.

3.2.5 Functional equivalence

This desideratum pertains to the process of selecting and translating citations for a bilingual dictionary. Huang demands that in selecting and allocating citations to each sense, the following five functions be taken into account:

i) to indicate some nuance or sophisticated shade of the headword meaning, i.e. to instantiate what is lost in the ‘sense translation’;
ii) to show the common collocates of the headword, i.e. its grammatical restrictions;
iii) to demonstrate the different equivalent expressions of the headword in the target language, i.e. to provide translations for the headword occurring in particular collocations;
iv) to help foster the ability to produce analogous sentences;
v) to provide the formulaic expressions co-occurring with the headwords, such as fixed phrases, idioms, or proverbs.

(Huang 1987a: 7)

Of the five major functions of illustrative citations, i) is to compensate the part of meaning lost in translation; ii) and v) are to illustrate the usage of the headword, i.e. how free it is in combination with other words or the concrete restrictions on the headword when it collocates with other words; iii) is to explicate the correlation of headword combinations between the two languages; iv) is to develop the learners’ encoding ability other than the mere imitation of example sentences.

For illustrative example translation, Huang especially emphasises that in a bilingual dictionary, a word, when posited as a headword, calls for rigorous semantic equivalence in translation, whereas when it appears in an example its translation should aim at functional equivalence. The reason for this is that the headword loses its independence once it enters into the context of an example sentence (Huang 1987a: 7). The underpinnings for the emphasis on semantic or functional equivalence may be found in Nida’s translation theory, where formal versus dynamic or functional equivalence are similarly differentiated (Nida: 1949; 1964; 1984; 1986; Nida & Taber 1982).

There are two thorny problems frequently confronted in bilingual dictionaries: i) the sense translation of a headword is phrased with a weightier consideration for citation translation at the cost of a more rigorous semantic equivalence; ii) in sorting out candidate illustrative sentences, the preferred ones are usually those better matching the sense translation, and as a result, the function of the illustrative sentence is likely to be
reduced to merely reinforcing the sense translation rather than encouraging the reader to be more productive in headword use (Huang 1987a: 7). To overcome these two problems, lexicographers should treat headword as entry-head versus headword in illustrative sentences as different translation units at different levels: at the entry word level, the goal is rigorous sense equivalence and at the illustrative sentence level, the goal is mainly functional equivalence. The main task of citation selection is to pick out sentences with their headword translation demanding more than merely a substitution of the sense translation. It will naturally be a greater challenge for the lexicographer to translate those illustrative sentences. The more a lexicographer puts into the work of translation, the more a dictionary user can get out of it. To take a less trodden path so as to benefit the users is the ideal for lexicographers to pursue.

3.2.6 Information on cultural words

Huang pays special attention to words whose cultural significance is for the most part implicit. The ‘illocutionary forces’ of these words fall into the categories of local conditions and customs, social conventions, modes of life, etc. Native speakers have little difficulty in comprehending these kinds of information and they do not require an explication in monolingual dictionaries. But the non-native speakers do not have access to the relevant information, and the problem cannot be solved simply by means of sense translation or usage labelling. The solution proposed by Huang is to give well-thought-out notes, concisely expounding those aspects of the cultural information in question (Huang 1987a: 8).

The above six desiderata for recompiling the *New Concise* are not only of considerable significance for bilingual dictionary-makers but may also be of value for some monolingual dictionaries.


In the following decade, Huang continued his lexicographical exploration, aiming at a theory of bilingual dictionary-making. In 1997 he co-authored *An Introduction to Bilingual Lexicography* (Huang & Chen: 1997), a monograph which laid a sounder foundation for making the *GCCFD*. In the following year, at a lexicographical symposium, he delivered his *Plan* for the *GCCFD*, which was then titled *A Modern Chinese-French Dictionary* (Huang & Zhang 2000[1998]). The *Plan* is a direct statement of how the dictionary is to be designed and compiled. The general guiding principle is briefed as ‘user-oriented: the natives first, the foreigners second’. The target audience is both Chinese native speakers and foreign language learners. Dictionaries for the native speakers are internally orientated and those for foreign language learners, externally orientated. ‘The natives first’ means that the needs of the native users should be prioritised in the dictionary. The needs of foreign language learners should be put second, especially when in conflict with those of the native speakers.

3.3.1 The internal orientation

For ‘internal orientation’, the specific objectives are to serve the needs of translating
Chinese into French and to serve the needs of learning French. To better serve translation,

*the perspective of the Chinese language should be adopted in setting up the criteria for determining headword status. For instance, phrases of a free combinatorial nature in Chinese should not be profusely ranked as headwords for the sake of catering to the corresponding words in the target language. [...] In some cases, the lexical units, such as homonyms, should be separately treated as different headwords. Otherwise, the users might mistake them as the same or mistake one as a misspelling.* (Huang & Zhang 2000[1998]: 230)

For headword coverage, the umbrella should be extended to embrace more words:

*more sci-tech words, newly coined words and newly emergent senses, slang and colloquial words, abbreviated words and phrases, and phrases and expressions that enjoy more translation potentials in the target language, and their inclusion will give the users more translation alternatives to consider, to compare, and to choose from.* (Huang & Zhang 2000[1998]: 232)

To better serve ‘learning French’, there are two objectives to achieve. One is ‘to control the vocabulary as strictly as possible in selecting the French equivalents for the Chinese headwords’ (Huang & Zhang 2000[1998]: 232). For instance, it is stipulated that the vocabulary from which the French equivalents are chosen should normally fall into the headword list of *Dictionnaire du français contemporain*, and no French equivalents should go beyond that of *Le Petit Larousse* unless they are sci-tech words or some few very special types of words.

The other objective is ‘to specify the usage and word class of the equivalent French words [and] if possible give a concise indication of their pragmatic restrictions’ (Huang & Zhang 2000[1998]: 233). Such vocabulary control and a specification of grammatical and pragmatic information surely benefits native Chinese users more than it does foreign users.

### 3.3.2 The external orientation

For ‘external orientation’, the objectives relate to three levels: the single-character entry level, the multi-character entry level, and the intra-entry information level.

At the single-character entry level, the aims are: i) to learn from ‘the practice of teaching Chinese as a second or foreign language: the bound morphemes, such as bound stems, affixes, and character-morphemes that merely represent speech sounds, will be taken as candidate members for the “head-character” list’ (Huang & Zhang 2000[1998]: 234); ii) to enlist a few low-frequency characters and character variants so as to meet the receptive needs of foreigners in learning or using Chinese at more advanced levels.

At the multi-character entry level, there are three things to consider: i) ‘to cover additional dialectal words circulating mainly in Taiwan, Hong Kong, Macao, and Guangdong Province’ (Huang & Zhang 2000[1998]: 235); ii) to label those two-character words into which another character can be inserted without a change of meaning with a ‘//’ between the two component characters; iii) to treat another particular type of two-character word whose component character can be repeated in one way or another without changing the meaning, which will also be labelled accordingly.
At the intra-entry information level, there are six issues for special consideration (Huang & Zhang 2000[1998]: 237-240). The first issue is part-of-speech labelling, the most controversial part of a Chinese-foreign dictionary. It is a long-standing problem for Chinese philologists and linguists to classify the parts of speech of the Chinese lexis, in that Chinese is highly analytic and grammatical functions are chiefly realised by word order rather than inflection. Lexicographers, therefore, have less reliable means to systematically label the parts of speech of the vocabulary. The policy adopted here is: not to be committed to obstinately labelling each and every word of the lexicon [but ...] to consider each instance more carefully, in order to guarantee the correspondence of the parts of speech of the sense translation with that of the word in the original language, i.e. to make use of a French word with the same part of speech in rendering a Chinese word, so as to comprehensively capture the different grammatical behaviour of the words in different usages. [...] In short, a feasible measure for getting around the controversial problem might be to indicate the parts of speech of a Chinese word by means of choosing the French equivalents with the same parts of speech (Huang & Zhang 2000[1998]: 237).

The second issue concerns the selection of illustrative examples. More full-sentence examples are expected for learners to model their utterances on. The third concerns the description of phonological or spelling variants of the headwords, i.e. the variants which are provided to indicate their occurrence in the major varieties of the Chinese language. The fourth is the treatment of noun/verb classifiers. Special labels such as S and SV are used for the common classifiers of both types (further discussion in 4.6.2). The fifth is contextually substitutable synonyms. These synonyms will be dynamically rather than statically described, i.e. the candidates will chiefly be confined to those whose substitution for the headword in use will not dramatically change the meaning of the illustrative sentences. The sixth is element-order-reversible words, i.e. those whose component characters are exchangeable in position without affecting the status of the word and its essential meaning. They should be distinctly labelled.

Almost all the above objectives have been fulfilled in the actual process of compiling the GCCFD.

4. Design Features

The following discussion will focus mainly on the five design features summarised by Huang in the Preface of the GCCFD (2014: 5-8).

4.1 User-orientation: ‘natives first, foreigners second’

This feature, established at the very outset of planning, can be regarded as a fundamental policy. In the Plan, Huang points out that:

_In principle, each dictionary should be designed to meet one or more of the needs of a particular group of target users, and it is reasonable to design a series of dictionaries to meet the different types of needs for each different group of users._ (Huang & Zhang 2000[1998]: 228)

There are also other scholars who share the opinion that one of the major problems for
the makers of either a medium-sized or a large-scale bilingual dictionary is:

*a mis-targeting of its anticipated audiences [...] – almost every large- or medium-sized Chinese-English dictionary has claimed that it will meet both the needs of domestic learners/users of English and the needs of the foreigners in learning Chinese and/or learning about China. But such a misunderstanding will surely lead to uncertainty in dictionary classification and indecision in the whole process of dictionary-making. (Zhao & Jiang 2014: 11)*

The fundamental policy, i.e. ‘natives first, foreigners second’ has, however, been persistently adhered to by Huang from the planning stage to the dictionary’s completion. What is the motivation for this design feature?

This question can be further divided into: What are the major needs of native users? How can those needs be satisfied in a large-scale Chinese-foreign dictionary? What are the major needs of foreign users? How can their needs be satisfied? Are the two different types of need compatible (in other words, do native speakers and foreign language learners have some needs in common)? Can they be satisfied in a single-volume dictionary?

The GCCFD has its answers ready: the major needs of native speakers are to learn French, translate and interpret Chinese into French, and increase the learners’ fluency in the use of French through using the dictionary. These needs could be met by:

*working out a headword list in a more standard and scientific way, so as to reflect the lexicon of contemporary Chinese[; [...] the lexis and wording of sense equivalents or translations in French being concise and sufficient enough to be easily accessible, comprehensible, and usable[; [...] the illustrative example phrases and sentences being of great value for reference, consultation, and imitation. (Huang 2014: 5)*

The major needs of foreign learners are:

*to solve the potential difficulties they may encounter in the process of learning Chinese as a foreign or second language, especially the complex relations between morpheme characters, non-morpheme characters, and words and phrases, and the difficulties involved in learning the usages of Chinese quantifiers. (Huang 2014: 5)*

These objectives are to be realised in the detailed and comprehensive labelling of the grammatical and pragmatic information of the headwords in Chinese.

In short, the GCCFD should be both a high-quality bilingual dictionary for native speakers and a learners’ dictionary with the headwords explained in detail and labelled explicitly for foreign learners of Chinese. Is such a ‘two-in-one’ design appropriate to the GCCFD? To serve the needs of the native speakers, the main difficulties for the GCCFD to address are those encountered in translation either from Chinese into French or from French into Chinese, and in communication in either spoken or written form. In other words, when a Chinese speaker comes to a C-F dictionary, the most typical situation is that s/he has some idea to translate into French but does not know the French word or expression for that idea. What is available to them is only the corresponding Chinese word. How can they retrieve a French word via a Chinese word? The starting point is the Chinese headword list in a C-F dictionary, or in general the headword coverage of the dictionary.
4.2 Headword selection

The first activity to be addressed by a lexicographer in compiling a dictionary is to determine the scope of coverage and the criteria for selection. The standards for determining the headword coverage of the GCCFD are ‘to be normative, scientific, and contemporary’.

Being ‘normative’, for the Chinese language, is largely an issue of whether a headword and its pronunciation is orthographic, whether it belongs to the standard language (i.e. putonghua) or nonstandard language (i.e. the different regional dialects), and whether a descriptive or prescriptive policy should be adopted in labelling the different types of linguistic information about it. Owing to the fact that its sole source dictionary, i.e. the MCD, has a heavily prescriptive bias, the GCCFD inevitably inherits a prescriptive orientation. Some measures, however, have been taken to increase the descriptive element. For instance, it has been specifically proposed that ‘some considerations should be taken for covering some regional dialectal words, and words popular among the overseas Chinese and/or in Hong Kong, Macao, and Taiwan’ (Huang 2014: 8).

Being ‘scientific’ is an issue that mainly concerns the source dictionaries. The latest versions of a number of Chinese philological dictionaries are referred to, with the MCD taking the lead. In headword selection, more than a dozen Chinese text corpora have been consulted regularly. As for the ratio of one-character entries to bi-character entries, a good balance has to be struck. According to some studies (Aitchison 2003; Pawley 2001), multi-word expressions (chiefly collocations) are dominant in most languages, being about ten times more frequent than single words (Tian & Chen 2009: 221). The GCCFD has 10,786 one-character headwords and over 100,000 multi-character headwords. The ratio is thus about 1:10, which corresponds to Aitchison’s (2003: 92) estimate.

Being ‘contemporary’ is consistent with the previously emphasised ‘principle of strict synchronicity’. The main difference resides in the extension of ‘synchronicity’ here to embrace more ‘classical Chinese words’ with a relatively high frequency of use, ‘literary words’, and ‘outdated words’, matching the capacity of a large-scale bilingual dictionary such as the GCCFD.

When the standards of being ‘normative, scientific, and contemporary’ are met in headword selection, the dictionary coverage can guarantee that what the target users turn to the dictionary, most of the terms they seek assistance about will most probably fall into its scope. Both the native users and the foreign users, in fact, share the same basic needs for consulting a dictionary – in translation or interpretation to and from French, in spoken or written communication in French, or vice versa.

From the perspective of the dictionary user, the most relevant aspect apart from headword selection is headword arrangement, which is also a core activity in designing the macrostructure of a dictionary.

4.3 Headword arrangement

The GCCFD has inherited the bi-level structure design for entry arrangement from the MCD: at the first level are found the main entries headed by single-character words and at the second level the sub-entries headed by multi-character headwords sharing the
same initial character with the first-level headword. Such a bi-level design can be traced back to Giles’s *Chinese-English Dictionary* (1892, edited by Kelly and Walsh) (2002: 133; Yang 2010: 389). At that time, the first-level head-characters were arranged in phonetic order, i.e. according to the Wade-Giles System. Head-characters sharing the same pronunciation would be ordered by the number of strokes each has, from the simplest to the most complicated. In Chinese dictionaries nowadays, both monolingual and bilingual, such phonetic ordering is also adopted, but the difference is that the prevalent phonetic system now is the *Hanyu Pinyin* system.

The practice of phonetic sequencing for headword arrangement in Chinese dictionaries can be traced back to ‘speech priority’ in the early period of bilingual dictionary-making in China, that is, when western missionaries first came to China the sole means of communication with the native inhabitants were the Chinese vernaculars. The principal aim for making a Chinese-foreign dictionary at that time was to meet their needs for cross-lingual face-to-face communication. The practice of head-character arrangement in traditional Chinese dictionaries is obviously unfit for looking up words phonetically. The early bilingual dictionaries, mainly externally orientated ones, naturally adopted the ‘phonetic sequence’ at the first-level single-character word arrangement. Such a design would inevitably destroy the functions of the radicals and components of the Chinese characters – the meanings indicated by them would be completely diluted. This is especially true for multi-character sub-entries, in that the shared initial character mostly functions as a ‘pre-modifier’ in the multi-character words. The terminal characters are different from each other, but they are in most cases the heads of multi-character constructions, determining both the part of speech and the core meanings of these constructions. In other words, the multi-character headwords grouped under one single-character entry are in most cases unrelated semantically.

Let us take the single-character entry 田 (tian [i.e. the pinyin]: field) as an example. There are as many as thirteen bi-character words governed by this in the *MCD-6* (2012), including 田产 (~ *chan*), 田地 (~ *di*), 田赋 (~ *fu*), and 田野 (~ *ye*). The terminal characters are 产 (estate), 地 (land), 赋 (tax), 契 (contract). These terminal characters are in most cases the heads of the constructions and dominate the meaning of each construction, but their meanings are by no means related to each other. By contrast, those headwords that share the same construction head 田, altogether twenty three, are scattered at different places in a dictionary, such as 坝田 (ba ~: small productive flatland), 丹田 (dan ~: an acupuncture point), 井田 (jing ~: Well-Field (System)), 瓦田 (qi ~: embanked field), under the letters b, d, j, and q respectively. This defect in macrostructure design has unfortunately been inherited by the *GCCFD* from the *MCD*.

A dictionary that bases its macrostructure on phonetic sequence is doomed to having the lexical semantic relations disconnected. This will also confuse foreign learners using the dictionary in order to learn Chinese. As for the compilation process, the semantically related headwords would normally have to be assigned to different compilers and thus would be treated separately, which surely affects the efficiency of dictionary-making and the quality of the final product. Meanwhile, under one single-character headword entry with a polysemous headword, the multi-character headword sub-entries can in no way be grouped to correlate with each of its senses. As a result, the whole dictionary inevitably becomes less logical and less tidy.
4.4 Illustrative examples

In order to make up for the defects of phonetic sequencing and to take account of the character-based nature of the Chinese language, the GCCFD has purposefully added a radical index for single-character retrieval and increased the number of illustrative examples for single-character entries. The examples are listed by the senses of the polysemous head-character, which is intended to help the users learn the character, its meanings and usage. Let us examine the following, a part of the entry 毛 (mao: fur, hair, feather, wool), first, in the MCD and then in the GCCFD for a comparison. The relevant part of MCD-6 (2012) reads as follows:

毛 1 名 ① 动植物的皮上所生的丝状物；鸟类的羽毛；羊毛 | 鸡毛 | 枇杷树叶子上有许多细毛。

The following is a bilingual version of that part:

毛 1 noun ① 动植物的皮上所生的丝状物；鸟类的羽毛 the silk-like thing growing on the skin of animals or plants; the feather of birds; hair; wool; fur; down; fluff: 羊毛 wool of sheep | 鸡毛 feathers of chicken | 枇杷树叶子上有许多细毛。Loquat tree leaves are covered with down.

The GCCFD, by contrast, has 16 illustrative examples:

毛 1 ① poil m ; plume f ; duvet m ► 兔毛 poil de lapin || 山羊毛 poil de chèvre || 骆驼毛 poil de chameau || 软毛 poil mou || 刚毛 poil raid || 全身长毛 avoir du poil sur tout le corps || 鸡毛 plume de coq [de poule] || 羽毛 plume d’oiseau || 掉毛 perdre ses plumes || 拔毛 arracher des plumes || 小鸡的绒毛 duvet de poussins || 桃毛 duvet de pêche || 长毛犬 chien à long poil || 短毛犬 chien à poil ras || 一套全毛西服 un complet de laine || 像猴子那样浑身是毛 être poilu comme un singe

For the convenience of discussion, I have provided a counterpart English version as follows:

毛 1 n. ① the silk-like covering which grows on the skin of animals or plants; the feather of birds; hair; wool; fur; down; fluff ► 兔毛 rabbit hair/fur; lapin || 山羊毛 goat hair/wool || 骆驼毛 camel hair/wool || 软毛 soft hair/wool || 刚毛 bristle; seta || 全身长毛。The whole body grows hair. || 鸡毛 chicken feather || 鸟毛 bird feather || 掉毛 molt or moul [BrE.]; shed feather/hair || 拔毛 pluck feather/hair || 小鸡的绒毛 the downy feather of a chick || 桃毛 the down of a peach || 长毛犬 long haired dog || 短毛犬 short-haired dog || 一套全毛西服 a woollen suit || 像猴子那样浑身是毛 the whole body covered with hair, like a monkey

Mao (毛), as a single-character entry, is treated in the GCCFD in the same way as in the MCD, both in headword arrangement and in sense demarcation. The most noticeable difference resides in the analysis and allocation of illustrative examples. The GCCFD, in accordance with the conventions of Chinese lexicographical practice, has grouped the free combinatorial phrases under the corresponding single-character entry, which helps pack more useful information into a limited space.
After a careful examination of some randomly selected entries in the GCCFD, we come to realise that the illustrative examples are well analysed and allocated to implicitly compensate for some of the inadequacies of the MCD in sense discrimination and definition. Let us look at the illustrative examples of the entry-毛 in more detail. The first group of examples contains some typical ones about ‘the hair or wool of the mammals’, i.e. 兔毛 (rabbit hair/wool; lapin), 山羊毛 (goat hair/wool), and 骆驼毛 (camel hair/wool). Then, two terms of a more scientific classification of hair/wool, 软毛 (soft hair/wool) and 刚毛 (bristle; seta). And then a sentence example: 全身长毛。 (The whole body grows hair). The second group consists of ‘the feathers of birds’, typically 鸡毛 (chicken feather) and 鸟毛 (bird feather). Then, two typical examples of verb-noun phrase, 舍毛 (shed feather/hair; molt or moult [BrE]), 拔毛 (pluck feather/hair). Then, there are two extended usages, 小鸡的绒毛 (the downy feather of a chick) and 桃毛 (the down of a peach). The third group concerns ‘pets’, 长毛犬 (a long haired dog) and 短毛犬 (a short-haired dog). Lastly, the artefacts of everyday life, 一套全毛西服 (a woollen suit) and a typical usage, 像猴子那样浑身是毛 (the whole body covered with hair, like a monkey). This well-thought-out treatment of illustrative examples will be very beneficial and facilitative to both native speakers and second-language learners in understanding French or Chinese, as well as in translating and interpreting.

The above discussion, centring around ‘user-orientation’, is concerned mainly with two aspects of the macrostructure of the GCCFD, namely the coverage and selection, and the analysis and allocation of headwords from the perspective of information retrieval. The microstructure of the GCCFD, such as sense translation and illustrative example, will be discussed in detail later.

4.5 Capturing recent changes in Chinese

This feature involves several aspects in the design of both the macro- and microstructure of the GCCFD. The guiding principle here is to emphasise the philological nature of the dictionary and to pay adequate attention to the encyclopaedic knowledge which an understanding of each headword entail requires. The philological nature has twofold significance: one is to accurately depict the basic status of the vocabulary of contemporary Chinese and the other is to depict the latest changes in everyday vocabulary – the new words and the new usages of existing words. To adequately depict the encyclopaedic knowledge, means to take account of both the words of science and technology and the nomenclatures of the major academic disciplines.

4.5.1 Coverage

The GCCFD headwords fall into two categories, the philological and the encyclopaedic. There are 10,786 single-character entries and about 90,000 multi-character sub-entries of the philological type. The sci-tech words and academic terms together number over 10,000. Can this word list and these proportions really depict the basic status of contemporary Chinese?

How well the GCCFD can represent the basic status of contemporary Chinese to a large extent depends on how successful the MCD-6, its sole source dictionary, has been
in fulfilling the task. While fully capturing what is presented in the MCD-6, the GCCFD has made some substantial changes and some daring innovations. According to our statistics, the GCCFD has adopted policies different from the MCD-6 as regards both single-character entries and multi-character sub-entries. For the former, the main objective was to reduce their number from 13,000 to 10,786, mainly discarding the outdated ones. For the latter, the main task was to embrace more words and expressions – here the number has increased from 69,000 in the MCD-6 to over 100,000 in the GCCFD, i.e. by 45%.

Let us randomly take a single-character entry ‘决 (jue: decide; finally)’ as an example to observe how it retains much of the original MCD-6 treatment and yet includes more multi-character sub-entries in the GCCFD. In the MCD-6, there are altogether 22 entries at the second level, such as 决标 (~ biao: final bid), 决策 (~ ce: policy-making), 决战 (~ zhan: fight a decisive battle). The GCCFD has 29 entries and the ones added are: 决不 (~ bu: on no account), 决出 (~ chu: contest), 决堤 (~ di: burst a dyke), 决非 (~ biao: definitely not), 决明 (~ ming), 决赛圈 (~ sai-quan: final round), and 决狱 (~ yu). These are all non-technical words except for 决明 (~ ming: Cassia tora), a botanical word, and 决狱 (~ yu: try a law case), a word of the judiciary with a shade of classic literary Chinese.

The increase of multi-character sub-entries in the GCCFD is largely an expansion in scale but not necessarily an intake of newly coined words. How can the latest changes in vocabulary be authentically depicted in the GCCFD? The following section tries to answer this question.

4.5.2 Neologisms

The GCCFD project lasted for sixteen years, during which its source dictionary, the MCD, witnessed three major revisions (2002, 2005 and 2012). It was thus necessary for the GCCFD to renew its headword list three times. For the intake of new words and senses, the GCCFD has taken under its umbrella about 3,000 new items, mainly from the fields of politics, economics, culture, and science and technology. While compiling the dictionary, the lexicographers also kept a sharp eye open for recent newcomers, such as 和谐社会 (hexie-shehui: a harmonious society), 低碳经济 (ditan jingji: a low-carbon economy), 电子书 (dianzi shu: e-book), 房奴 (fangnu: mortgage slave), 二维码 (erwei ma: two-dimension code), 平板电脑 (pingban diannao: panel computer), 网购 (wanggou: online shopping), 微信 (weixin: WeChat (literally, micro-message/letter)). Quite a few of these emerged very recently and had not been registered in the latest edition of the MCD-6 (2012).

The intake of new words and senses is not only a core task for compiling a new dictionary, but will be a principal activity for its future revision. However, the need to identify new items and to formulate criteria for their inclusion remains a persistent problem for lexicographers.

First, no accurate assessment of the yearly influx of new words and new senses into Chinese is available. There has been no such research on this issue in Chinese studies to date. In English literature, there are three relevant studies on the vocabulary of English. Barnhart (1978) indicates that in the quarter of a century between 1947 and 1973, about 800 new words entered the ‘common or working vocabulary’ per year and roughly 500 of them ‘managed’ to appear in a variety of dictionaries (Landau: 2001: 472...
202). The second study is by the OED which ‘during the twentieth century [...] recorded around 90,000 new words, or roughly 900 per year’ (the Centre Section (page 2) of the COD (11th edn, revised in 2008)). The third and latest study was carried out as a result of work on the third edition of the OED. According to statistics published on the OED website, the dictionary has regularly added batches of new words, having done that altogether 67 times by September 2016, averaging 500 words quarterly, or around 2000 words annually at the turn of the 21st century9. The statistics for the neologisms in English might give some hints of the situation in Chinese.

Second, there is a lack of commonly accepted criteria for the identification and selection of new words and phrases for inclusion. A prerequisite is to judge whether an expression in Chinese has been lexicalised or not. For instance, by what criteria can we judge whether a bi-character expression is a word or a phrase, that is, a free combination of two characters? For the purposes of simplification, an expression can be regarded as a word if its meaning is more than the mere sum of the meanings of its two component characters. Then, what are the criteria by which to decide whether a new word is to be listed in the dictionary? The selection should take into consideration its use: its actual distribution in time, space, and register. Only when a new word can meet most or all of the above distributional requirements can it be valued as common or conventional enough to be included. But in Chinese studies, the statistics for verifying the commonness of newly coined words are not yet available. There is still more work to be done in preparation for the GCCFD revision programme in time to come.

The intake of new words also involves sci-tech words and the nomenclatures of a variety of academic disciplines or sub-disciplines, which is dubbed by Huang (2014: 6) “the most heavily contaminated part” of most philological reference works’.

4.5.3 Ordinary versus technical words

How can a dictionary guarantee that the criteria for its inclusions of neologisms are scientifically accurate and valid? Different dictionaries adopt different policies. For instance, some American college dictionaries include the important terms frequently encountered by graduate students. Higher standards of general education and the impact of the advance of science and technology on everyone’s lives have accelerated the process of turning specialist terms into common words. In practice, the lexicographer should have some knowledge of the difference between ordinary words and technical terms. As I have argued elsewhere (Tian 2010: 18-20), the dependence of technical terms upon specialised theories has provided a perspective which allows scholars to see the differences more clearly.

4.6 Definition and labelling

The task of describing, explicating and labelling the vocabulary of contemporary Chinese, is a sophisticated one. It involves working both objectively and scientifically to capture the various aspects of the meaning of each lexical item. The GCCFD has proposed a general postulate: that definitions should be as concise and applicable as possible and the labelling system as detailed and explicit as possible, so that the dictionary can better help the users to learn and grasp the essentials and patterns of the vocabulary of a language. The postulate can be further employed to guide the practice of sense translation and arrangement, the analysis and allocation of illustrative examples,
and the labelling of grammatical information, rhetorical colouring, or cultural information of the lexical items.

4.6.1 Lexical semantic description

There are four points to note when treating the lexical semantics of a word. The first involves the control of vocabulary used for sense translation into French. In most cases the French equivalents of ordinary Chinese headwords are everyday expressions, and the words employed should be controlled and not go beyond the coverage of *Le Petit Larousse or Dictionnaire du français contemporain*. The sci-tech words are different. The point here is the depth of the user’s knowledge, which determines how user-friendly the wording of a definition should be.

The second point is to provide more alternatives for the translation of certain headwords, i.e. idioms or idiomatic phrases. Their translation might be realised as: i) literal or word-to-word ‘sense or definition translation’ into French; ii) further expounding or explication in French; iii) corresponding idioms or idiomatic phrases in French; iv) making full use of examples to illustrate typical usage in typical contexts.

The third point is to use cross-reference devices to link the headwords to their synonyms, antonyms, or other semantically related headwords or senses, helping the users to establish mental lexical networks for the target language. This is applicable to both native and foreign users. The issue here in the effort required to detect and establish the semantic networks of the associative meaning of the lexicons, both intra- and inter-linguistically.

The fourth point is to label the grammatical and pragmatic meanings of headwords in a detailed and explicit fashion, helping to realise the learning function of the dictionary. The detailed labelling of grammatical information (such as parts of speech) and rhetorical (or register) information on the French equivalent is chiefly designed to help the native Chinese users in their production or translation into French.

4.6.2 Grammatical and pragmatic labelling

As for helping those learning Chinese, the ‘full and detailed’ labelling pertains chiefly to headword labelling, with special attention paid to the following: i) the labelling of Chinese characters that cannot be used individually, such as root- and affix-characters as bound morphemes, and those non-morpheme characters for representing speech sounds only; ii) the labelling of headwords (‘separable compound words’) which allow a particle character to be inserted into them, with a special symbol ‘//’ indicating this; iii) the labelling of those headwords, especially bi-character ones, whose character position is exchangeable without altering the meaning of the compounds; iv) the labelling of quantifiers and nominal classifiers, which constitute a major problem for foreign learners of Chinese.

The classifiers in the GCCFD are meticulously treated in two ways. One is to list the classifiers that frequently collocate with a certain headword; the other is to label those classifier characters and indicate the scope of their usage. The classifiers treated in the GCCFD can be further divided into two major types: the nominal and the verbal ones. The former will be briefly discussed here as an illustration. Classifiers are normally elusive, even for Chinese people from different dialectal regions. Let us examine a set of nominal classifiers in standard Chinese. When an individual item of
livestock is referred to, the classifier most frequently used is 头 (tou: head). It can be used for a cow, a donkey, a sheep, or a pig, but it is seldom used to refer to a horse. For a horse, the most frequently used classifier is 匹 (pi: [vacant]), which can also be used for a wolf, but not for a dog (that requires the use of 头 or 只 (zhì: [vacant])). In addition to 头, another classifier for a pig is 口 (kǒu: mouth), which can also be used for a human being, but not for other animals. For animals such as wolves, tigers or lions, 只 is the most frequently used classifier. In ancient China, living beings were roughly classified into four types: birds, beasts, insects, and fishes, and 只 could be used for almost any instance of one of the first three types. Moreover, in contemporary Chinese, the classifier 只 can also be used to refer to one of a pair of things, such as an eye, hand, foot, ear, shoe, stocking or glove, etc.11

The discussion here involves the different aspects of ‘definition’ and ‘labelling’. Each treatment is especially designed to serve the needs of a particular group of the target audience, either the Chinese or the foreign users, and quite often other groups of users may more or less benefit from it as well.

4.7 Illustrative examples revisited

The GCCFD emphasises the role of context in vocabulary learning, elaborating that:

For non-native learners, the equivalent words in the target language can only serve as the first step towards the lexical meanings, and the more in-depth and dynamic aspects of word meaning can only be captured in their actual use in each authentic and typical context. ... The illustrative examples should be extracted from an authentic large-scale corpus and will need to be modified to some extent so as to make it easier for the learners to better appreciate the usage differences of a word in different contexts. Such a treatment may also do some good to the translators, in that more flexible translation candidates and expressive alternatives are provided for them to learn and choose from. (Huang 2014: 7)

In 4.4, we used 毛 to illustrate the analysis and allocation of illustrative examples for single-character entries. Here we will take 工业 (gōngyè: industry) as an example of a multi-character sub-entry. Let us examine 工业 in the GCCFD:

【工业】
化学~ industries chimiques ||轻(重) ~ industrie légère (lourde) ||设备 équipement industriel || ~ 体系 système industriel || ~ 城市 villes industrielles || ~ 区 une zone industrielle || ~ 产品 articles [produits] industriels || ~ 化学 chimie industrielle || ~ 美术 art industriel || ~ 技术 technique industrielle || ~ 污染 pollution industrielle || ~ 布局 répartition géographique, répartition industrielle || ~ 界 monde industriel || ~ 学校 école des arts et métiers; école polytechnique

For the convenience of comparison, I have again translated the entry into English:

【工业】industry
化学~ chemical industry ||轻 (重) ~ light/heavy industry || ~ 设备 industrial equipment
The corresponding entry in *A Chinese-English Dictionary* (*3rd* edition, 2010) reads as follows:

【工业】
industry

~病 occupational diseases of industrial workers || ~国 industrialized (or industrial) country || ~化 industrialize || ~品 industrial products; manufactured goods || ~园 industrial park || ~权 industrial property right || ~尘 industrial dust || ~革命 Industrial Revolution || ~基 industrial base || ~酒 industrial alcohol || ~企 industrial enterprise || ~气压 technic atmosphere || ~染 industrial pollution || ~总产值 gross value of industrial output || ~产阶级 the industrial proletariat

With regard to the type of examples, the *CED-3* has just one type, i.e. "工业 (industry)" as a premodifier in the illustrative constructions, whereas the *GCCFD* has a new type where the headword functions as the head of the illustrative construction, e.g. ‘化学~chemical industry’ or ‘轻 (重)~light/heavy industry’.

With regard to the number of examples, the *CED-3* has 15 and the *GCCFD* has 14. But upon careful examination, we find that four illustrative examples: ‘工业国, 工业化, 工业产权, 工业革命’ in the *CED-3* are treated as multi-character sub-entries in the *GCCFD*, following the *MCD-6*. In other words, the *GCCFD* has, in fact, three more illustrative examples than the *CED-3*.

With regard to the content, the two dictionaries share only one item, i.e. "工业污染". This can be regarded as proof of the independent compilation of the large-scale bilingual dictionaries in China. It should be added that the *MCD-6*, the source dictionary of the *GCCFD*, provides no illustrative examples for the sub-entry "工业".

Another practice which distinguishes the *GCCFD* from other large-scale Chinese-foreign dictionaries, is that of giving ‘the synonyms and/or near synonyms of the headword in the illustrative examples rather than barely listing them directly following the sense translation’ (Huang 2014: 6). Let us see some examples:

【病】

① maladie f: mal m ► 生病 [得病*, 患病*] attraper une maladie; tomber malade; devenir malade

【收录】

employer v.t; engager v.t; recruter v.t ► 我厂刚收录[录用*]了 100 名工人。Notre usine vient de recruter cent ouvriers.

【收买】
A language is a vehicle for culture and to learn a foreign language is essentially to learn and understand another culture. [...] In addition to presenting semantic and pragmatic information, a dictionary should make an extra effort to expand its dimensions to cover the cultural information of a word. Therefore, we will see to it that an eye is to be always kept on cultural significance while making decisions on the wording of sense definitions, collocating phrasal combination, labelling pragmatic information, or selecting illustrative example sentences. [...] For those words and phrases with specific cultural import, in addition to their sense translation into French, the relevant cultural background knowledge needs to be provided so that the user could have a more comprehensive understanding of the content of the headword. Examples of such words are ‘阿Q’ and ‘一二·九’运动. (Huang 2014: 7)

In treating culturally loaded words in bilingual dictionary-making, Huang believes that their cultural content needs to be revealed. This is similar to Guzhuang’s views on academic translation. Guzhuang (1983) maintains that an academic translator is responsible not only for translating the senses of the words but also for explicating their content fully. In other words, s/he has to introduce a particular set of concepts from one culture into another. In rendering an academic work from a foreign language into their native one, the translator should shoulder roughly the same responsibility as the original author in putting forward a new concept or theory: to ‘fully explicate that concept or theory and have it established’ in his native language, i.e. to simulate the author’s originality in reformulating ‘the theory and its governing concepts’ in the target language. Guzhuang further points out that the translated senses given in a bilingual dictionary are usually what the general public conceives of the word. In most cases, they are formed by a process of conventionalisation among the community members during a certain time span. This is different from terms or technical words whose meanings are derived from the academic theories to which they are related. The treatment of culture-specific words and phrases in large-scale bilingual dictionaries is, to a large extent, analogous to what Guzhuang has proposed as the method for treating discipline-specific technical words in translating academic texts.
dictionary-making, is a perfect circle of practice to theory and then theory to practice. His experience in dictionary-making and exploration into general and bilingual lexicography are, in a sense, a direct reflection of the progress of Chinese lexicography in the past four decades.

The analysis of the GCCFD in this study has focused upon headword coverage and selection, sense translation, and citation selection and translation. The study concludes that the current situation and the latest changes in the lexis of the Chinese language have been successfully captured by adopting the macrostructure design of the MCD and by reducing the number of its single-character words while increasing that of the multi-character ones. Unfortunately, through the adoption the GCCFD has necessarily inherited some inherent defects of the MCD.

At the microstructural level, semantic equivalence has been rigidly pursued in headword sense translation and the vocabulary for translation reasonably controlled; functional equivalence has been pursued in illustrative citation translation and synonymous alternatives of a headword given in its citations whenever possible; and the labelling of grammatical and pragmatic information for headwords has been required to be explicit and in detail to better help its target users.

The GCCFD, through its innovative designs and rigorous execution has successfully realised its fundamental goal of combining a reference dictionary and a learner’s dictionary into one. These design features and remarkable innovations may also benefit other bilingual dictionaries of a similar size or nature.

Notes

1 In traditional Chinese linguistic studies, the term ‘multi-character word’ is normally used to refer to words that consist of two or more characters. The present paper follows this convention.

2 Unless stated otherwise, the quotations from Huang’s works, originally written in Chinese, such as this one, have been translated into English by the present author.

3 释义 (yiyi), as a bilingual lexicographical term in Chinese, could be ambiguous when translated into English. There are two other related terms in Chinese. One is 定义 (dingyi: sense definition) and the other is 释义 (shiyi: sense explication). In general, 释义 (literally ‘meaning translation’) could be translated as ‘sense translation’, which means finding the ‘equivalent word or expression’, or simply ‘equivalent’ in the target language. In some contexts, it could also mean the translation of 定义 proper, thus we have ‘sense definition translation’, or simply ‘definition translation’ or the translation of 释义 proper, then we have ‘sense explication translation’ or simply ‘explication translation’.

4 In the late 1970s and early 1980s, bilingual Japanese dictionaries were important references for bilingual Chinese dictionary-making. Compared with western languages, Japanese and Chinese share a similar tradition in dictionary-making and more common ground and conceptual ideas in life and in learning. Consequently, it was often thought that Japanese bilingual dictionaries are a convenient shortcut for making Chinese bilingual dictionaries.

5 For large-scale philological dictionaries, either monolingual or bilingual, it is hard to decide whether and to what extent to cover some dialectal words and phrases.
Different dictionaries adopt different polices. A similar policy is shared by a newly published large-scale Chinese-English dictionary: ‘[In The Chinese-English Dictionary (Unabridged)], we have paid special attention to recording words and phrases from regional dialects in China, such as Yue [including Cantonese and Taishanese], Min [including Fuzhouinese, Hainanese, and Taiwanese], and Wu [including Shanghaiese, Suzhouese, and Wenzhouese]. And the facts and phenomena of the Shanghaiese [and the people and their lives] have been more sufficiently represented, for the Dictionary is compiled [by the people] and published [for the people worldwide but not necessarily] here – a waterside pavilion gets the moonlight first, as a Chinese proverb says’. [from the Preface to The Chinese-English Dictionary (Unabridged) by Gusun Lu 2015]

Some scholars (Qigong 1985; Qiu 1988) question the qualification of multi-character words as words and identify at least some of them as fixed phrases or collocations. For convenience of discussion, we will follow the criteria for word identification adopted by the MCD-6 throughout this study.

For the majority of the bi-character words listed here, the first character 田 functions as a premodifier. The case of 田地 is slightly different. In classic Chinese, 田 and 地, as indicated in Shuowen Jiezi (The Origin of Chinese Characters) (121 AD), are two single-character words for cropland and non-cropland respectively. In contemporary Chinese, they have established a “co-referential” relation, indicating “land with crops planted”. The discussion here has mainly adopted a synchronic perspective.

It is very popular in South-East China and considered to have many health benefits.

As Simpson, the former chief editor of the OED, points out: all statistics of this nature are suspect! The OED deals with English throughout the world, whereas David Barnhart was more interested in American vocabulary.

In expressing the meaning of two, the 只 as in 两只 (two individuals), denotes an individual. It does not indicate the relationship between the two, but when 一双 (‘a couple of’) is used, a relationship between the two is implied.

Here we can see that the CED-3 and the GCCFD have adopted different policies in “borrowing” from the MCD. The GCCFD adheres closely to the MCD, although it may add some facts not yet covered by the MCD or omit certain things identified as outdated but not yet deleted from that dictionary. The CED-3 is bolder in treating things differently from the MCD, downgrading some sub-entry headwords to free combinations and putting them as illustrative examples.

Ah Q (the main character in Lu Xun’s The True Story of Ah Q (阿Q正传), who consoles himself by interpreting his defeats as moral victories) [cited from the CED-3]

December 9th Movement (a demonstration staged on 9 December 1935 by Beijing students under the leadership of the Chinese Communist Party, calling for resistance to Japanese aggression and national salvation) [cited from the CED-3]
References

A. Dictionaries

A Chinese-French Dictionary (ed. by the compilation team of A Chinese-French Dictionary, from the Department of Western Languages and Literature, Peking University and the Department of Western European Languages, the Second Beijing Foreign Languages Institute) (1991). Beijing: Commercial Press.


B. Other Literature


Guzhuang [G. Xu]. 1983. ‘An Exploration into a Style for Translating Academics: from a Case Translation of Ten Paragraphs on Renaissance in Russell’s History of
Western Philosophy.’ Foreign Languages 1: 1-6.


Criteria for Selecting Trisyllabic Words as Headwords in the Chinese-French Dictionary

Shuyan Wang
Center for Lexicographical Studies, Guangdong University of Foreign Studies
200110777@oamail.gdufs.edu.cn

Peng Zou
School of Foreign Languages, Zhongkai University of Foreign Studies
285582956@qq.com

Abstract

In the 2015 concluding report An Approach to Revising Chinese-French Dictionary -- Resequencing Entry Words, sponsored by Lexicographical Studies Center at Guangdong University of Foreign Studies, we put forward three criteria for adjusting trisyllabic words based on their disyllabic. The three criteria suggest to maintain, deprive or restore the use of trisyllabic words as headwords for separate entries. Considering the enormous size of this word category and its complicated intrinsic semantic relations as well as diversified grammatical features, this paper takes thirty trisyllabic words that fall in Yang Shujun’s “Nine Structure Categories” as examples to check whether the above three criteria can be widely applied and promoted in reality through a method of word prosody trichotomy (trisyllabic words are classified into three general patterns, namely [1+1+1], [2+1], and [1+2]).

Key words: Trisyllabic words, Chinese-French dictionary, Selection criteria

1. Introduction

In a publication in the early 1960s, Lǚ Shuxiang elaborated on ‘trisyllabic segments’, and since then scholars in China have launched many studies on the topic. Trisyllabic segments are one of the prosodic segments in the modern Chinese language, which require further studies from different perspectives (Jiang Huihui, 2011).

According to Zhou Jian (2003), the Modern Chinese Dictionary (revised edition, published in 1996) collected a total of 4,910 trisyllabic words, an amount approximate to the 5,000 entries established in the Modern Chinese Dictionary of Trisyllabic Words (Wang Yankun, 2005). The word list compiled by the Chinese State Language Commission, which was later used as a corpus by Yang Shujun for academic research, however, witnessed a total collection of 10,963 trisyllabic words (Yang Shujun, 2008). These numbers demonstrate the fact that trisyllabic words constitute an important part of the modern Chinese lexical system.

Based on a the disyllabic roots in An Approach to Revising Chinese-French...
Dictionary --- Resequencing Entry Words, the concluding report for the 2015 research cultivation project sponsored by Center for Lexicographical Studies at Guangdong University of Foreign Studies, we put forward three criteria for adjusting trisyllabic words. The three criteria suggest to either maintain, deprive or restore the right of trisyllabic words to be headwords for separate entries. Considering the vastness of this word category, and in view of its complicated intrinsic semantic relations and diversified grammatical features, this paper takes 30 trisyllabic words that fall under Yang Shujun’s “Nine Structure Categories” as an example to check whether the above three criteria can be applied in practical operations and promoted for general use by way of word prosody trichotomy. We classified trisyllabic into three general patterns, namely [1+1+1], [2+1], and [1+2]. In short, we are looking for factors that influence the position of trisyllabic words in dictionaries, whether headwords, subordinate entries, or even examples within an entry.

2. The General Pattern of [1+1+1]

The general pattern of [1+1+1] we refer to hereunder comprises two types of three-character words: the foremost of Yang Shujun’s Nine Structure Categories, simple trisyllabic words (like 安琪儿, 阿拉伯, 阿波罗), and the third type, words combined by three parallel monosyllabic morphemes (for example, 党政军, 德智体, 传帮带). With the above categorizations, we clearly denoted evidence that simple words and compound parallel-structure words both belong to the same category, namely an infinite combination of individual characters, as can be represented by the general pattern of [1+1+1].

According to the statistics given by Yang Shujun, trisyllabic words that fall in such pattern are rare, with 63 simple and 39 compound parallel-structure words. The 102 words make up a share of 0.94% (a sum of the boldface numbers listed below).

| Table 1 Statistics on Trisyllabic Words by Yang |
|-----------------|-------|------|----|
| structure       | 0     | 0+1/1+0 | 1+1+1 |
| number          | 63    | 259    | 39  |
| ratio           | 0.58  | 2.36   | 0.36 |

1To avoid confusion with the method used by Yang Shujun for syllabic structures of trisyllabic words, we use brackets [] to differentiate and denote a more inclusive general pattern. Here in the paper, word structures without brackets (like 2+1) are recognized as Yang’s method for categorization, which is a relatively narrow and multi-category pattern.
Here, we believe and make the suggestion that all transliterated words have the right to be collected as headwords for entries. Committed to both theoretical research and practical compilation of bilingual dictionaries, the Center for Lexicographical Studies devotes itself to bridging the gap between different linguistic symbols and realizing the equivalence between two languages. Hence, we’re more interested to see from the French perspective how editors of the *Chinese-French Dictionary* defined the word 天使. Through observation, we came to realize that they failed to demonstrate the relevant derivatives of this loanword. In comparison, the masculine noun *ange* is clustered with numerous derivatives, for example, *angelet* or *angelot* (minor noun, denoting little angles in religious art), *angélique* (adjective), *angéliquement* (adverb), *angélisme* (noun), *archange* (noun), and *archangélique* (noun). These words, in terms of morphology, are all derivatives of *ange*.

The case of 阿拉伯, however, is different from that of 安琪儿. The French equivalent *arabe* is defined by a supplement series of noun, verb, and adjective derivatives of the same word family. For example: *arabesque*, *arabique*, *arabisant*, *arabisation*, *arabiser*, *arabisme*, *arabiste*, *arabité*, and *arabophone*. It is also used as a prefix, like *arabo-islamique*, and to form certain proper nouns. Common examples include *Arabe* (Ligue), *Arabes Unis* (Emirats), *Arabe Unie* (République), *Arabie, Arabie Saoudite ou Séoudite*, *Arabique* (golfe), and *Arabique* (mer).

Compared with the above two words, 阿波罗 has less derivative forms. The only two examples are *apollon* and *apollinien, ne*.

In reference to the above analysis and based on the fact that transliterated words can better reflect specific information in Western cultures, we therefore infer that making the above mentioned trisyllabic words headwords to establish entries in Chinese-French bilingual dictionaries would not only acquaint users with more French vocabulary, but also assist them in understanding Western cultures, to better facilitate cross-cultural communication.

The Chinese-French Dictionary hasn’t yet collected 党政军, 德智体, and 传帮带, but rather used the second as an example under the headword 德. We suggest that such words can also be collected to make headwords for separate entries, as is the case for simple trisyllabic words.

Briefly, we believe it is a sound option to maintain or collect trisyllabic units in
the general pattern of [1+1+1] as headwords, since such words enjoy a distinctive morphological structure and are limited in number.

3. The General Pattern of [2+1]

Our general pattern of [2+1] also includes multiple elements. We included Yang’s fifth category narrow 2+1 structures, trisyllabic words formed by a disyllabic word and a monosyllabic morpheme, as well as the subsection of the second category, the 0+1 structure, trisyllabic words formed by a disyllabic simple word and a monosyllabic morpheme. We also included Yang’s seventh category, the (1+1)+1 structure, trisyllabic words formed by two temporarily matched morphemes and the AAB form nineth category, a monosyllabic morpheme and (1+1)+1 structure.

According to the above categorization standard, the general pattern of [2+1] shall include 4 subcategories, namely, simple trisyllabic words (e.g. 侏罗纪, 爵士乐, 傈僳族), compound trisyllabic words (e.g. 男子汉, 扩大化, 脑袋瓜), temporarily matched trisyllabic words (e.g. 孔方兄, 纠风办, 备忘录), and reduplicated trisyllabic words (e.g. 飘飘然, 欣欣然, 蒙蒙亮).

Based on the statistical analyses, Yang pointed out that the trisyllabic words formed by a disyllabic word and a monosyllabic morpheme serve as the dominant structure in modern Chinese (6,105 of all trisyllabic words), so we only described the second category as mentioned.

In the dictionary, we observed that all of the compound trisyllabic words were collected as headwords, as is the case for 男子, 扩大, and 脑袋. See these examples below:

【男子】

*homme m (adulte)气概 caractère mâle; virilité; masculinité ||特权 privilège de masculinité ||团体育 épreuve masculine par équipes ||主义 machisme m ||双打[sports] (épreuve) double m messieurs
double m messieurs |||×打[sports] (épreuve) double m messieurs

【男子汉】

（Q. 个、位、条、名）homme m de caractère; gaillard m ➤拿出点儿~的气概来！Conduis-toi en homme. ||别哭，要像个~啊！Ne pleurs pas ! Sois un homme ! ||她干活赛过~。Elle surpasse les hommes au travail.

Also, after comparing the above twoFrench equivalents for 男子 and 男子汉, we
believe that the latter can be maintained as a headword, whereas it should also be established as an inner entry under the former, to show the close semantic relation between the two.

【扩大】
agrandir v.t; amplifier v.t; élargir v.t; étendre v.t; accroître v.t ~ 耕地面积 étendre la superficie des terres cultivées || 势力范围 étendre la sphère d’influence || 影响 étendre son influence || 视野 élargir son horizon || 知识面 accroître le champ de ses connaissances || 出口 accroître les exportations || 企业自主权 élargir l’autonomie de l’entreprise || 理事会~会议 séance élargie du conseil ■~ 器 amplificateur m

【扩大化】
donner une ampleur démesurée [exagérée]; élargir excessivement l’étendue; extension f outrancière || 阶级斗争 extension outrancière de la lutte de classes || 将矛盾~ étendre excessivement le champ des contradictions

In the word-formation system of the modern Chinese language, the character 化 serves as a functional suffix for semantic analogy and extension, which gives the trisyllabic words formed by it a priority to be treated as subordinate derivatives. Therefore it is better to make trisyllabic words derived from disyllabic bases subordinate entries and then choose appropriate French equivalents that match their disyllabic bases. A comprehensive analysis shows that 扩大化 should better be treated as a subordinate entry under 扩大, rather than as a separate headword.

【脑袋】
① (Q. 个) tê te f ~ 圆 tê te ronde || 掉 perdre la tê te || 挨了一枪 recevoir une balle dans la tê te ■~ 头 ② esprit m; cerveau m; tê te f 我的~ 不大好使。J’ai l’esprit plutôt lent [lourd]. || 我不知道他~ 里想的什么。Je ne sais pas ce qui lui passe par la tê te. ■~ 脑筋

【脑袋瓜儿】
(=脑袋瓜子) ■~ 脑袋

The last example leads us to say that trisyllabic words derived simply from disyllabic words with no distinct French equivalents may well be crossed out, and shall not be used as headwords for entries. These trisyllabic words can be allocated under the entry
of their disyllabic bases, as they are simply synonyms.

In short, if certain trisyllabic words that fall in the general pattern of [2+1] can be established as headwords, we are of the opinion that others – often with a root and a suffix – have a better chance of finding semantically the same French equivalents, and should be made into subordinate entries.

4. The General Pattern of [1+2]

The general pattern of [1+2] is composed chiefly of the 1+2 structure raised by Yang, or the fourth category: trisyllabic words formed by a monosyllabic morpheme and a pre-existing compound disyllabic word. This pattern also comprises the 1+0 structure (a subcategory of the second category: trisyllabic words formed by a disyllabic simple word and a monosyllabic morpheme), the 1+(1+1) structure (the sixth category: trisyllabic words constituted by a monosyllabic morpheme and two temporarily matched morphemes), and the 1+(1+1)' structure (the eighth category: ABB type). In other words, the general pattern of [1+2] and [2+1] are relatively parallel in terms of word structure, as trisyllabic words fall in the two patterns are all formed with a compound two-character word base and a separate third word, with the only difference coming in the combination sequence.

We may therefore use the same set of terms to subdivide the above four types of lexical units: simple words (e.g. 泥菩萨, 老糊涂, 胡萝卜), compound words (e.g. 把兄弟, 包打听, 唱反调), temporarily matched words (e.g. 吃不消, 动不动, 凉白开), and reduplicated words (e.g. 急匆匆, 假惺惺, 静悄悄).

Similarly, we decided on the second subcategory, composed of 2,084 trisyllabic words. According to Yang’s statistical data, it’s a word-prolific trisyllabic structure. That is to say, we only explain 把兄弟, 包打听, and 唱反调 in our discussion.

As far as the three compound words are concerned, 兄弟 and 把兄弟 are both collected as headwords, as is the case for 打听 and 包打听. However, 唱反调 is relatively complicated: in addition to being collected as a headword, the trisyllabic word is also used as an example for the disyllabic word 反调.

【兄弟】

1frère m aîné et frère (s) m. pl cadet(s); frères m. pl  2<p. ext.> celui que l’on considère comme frère ; frère m  3<fig.> frères m. pl ; soeurs f. pl  ～国家 pays frères  || ～民族 nationalités soeurs
【兄弟】

<lang. p. > ① petit frère m ; frère m. cadet  ② [s’adressant à un homme moins âgés que soi-même, avec une nuance affecteuse]mon frère  ③ [term de modestie]moi pron (homme parlant de soi-même)

【把兄弟】

frères m. pl par serment ; frères m. pl jurés
According to their french equivalents, it’s better to make 把兄弟 a subordinate entry under the second sense of the former entry 兄弟. See examples above.

【打听】

s’inform er (de, sur) v. pr ; se renseigner (sur) v. pr ; s’enquérir (de) v. pr ; aller aux renseignements ; aller aux informations ▷ 向某人～消息 ～事实真相 s’enquérir de la vérité d’un fait

【包打听】

<dial.> ① agent m de police secrète ; policier m enquêteur ; détective ●≈ 包探 ② personne f chargée de recueillir des renseignements ; personne f à l’affût des nouvelles
The above two words can both be maintained as headwords because of their different parts of speech.

【反调】

ton m différent ; point de vue m contraire ; opinion f opposé ▷ 唱～ prendre un ton différent ; émettre une opinion contraire

【唱反调】

tenir des propos diamétralement opposés (à) ; aller à l’encontre de

This last pair, together with 高调 (high profile), 唱高调 (being bombastic, not realistic), 老调 (platitude), and 唱老调 (harp on the same string) shall be revised
after improving their French equivalents (completely different), as the trisyllabic words all fall in the verb-object construction. They can be, with a higher priority, made into examples for their disyllabic counterparts or used as subordinate entries for an extension of semantic meaning.

As we see it, the compound word 把兄弟 can better be treated as a subordinate entry, 包打听, whose meaning is very different from its disyllabic word base. Because they belong to different parts of speech, we suggest to keep them as headwords. 唱反调, in our opinion, should be demoted as an example for the headword 反调.

5. Conclusion

In summary, to be a headword in the dictionary, we believe that a trisyllabic word should meet either of the following two conditions: first, no relevant disyllabic word has been collected, and second, the trisyllabic word bears little or no connection with the relevant disyllabic root, given the latter has already been collected as a headword. In this paper, trisyllabic words that satisfy these requirements include: all trisyllabic words of the general [1+1+1] pattern, and some of trisyllabic simple words and some of temporarily matched trisyllabic words that fall in the general pattern of [2+1] and [1+2]. Trisyllabic words that have similar equivalents to their disyllabic roots and carry metaphoric meanings should better be made into inner entries. For example, 男子汉, which belongs to the general pattern of [2+1], can be made into an inner entry for 男子, the disyllabic root for which is collected as a separate entry.

The reason for making a trisyllabic word a subordinate entry is that the word is formed by a root and an affix (although its French equivalent has no such morphological structure). For example, 扩大化 falls in the general [2+1] pattern. The difference between an inner entry and a subordinate entry comes in that subordinate entries are deprived of the right to become headwords for separate entries, although they may be elaborated with examples, whereas inner entries can be made into separate and independent entries for a second time. Of all the three general patterns, the latter two include four subcategories that share same structures, and are therefore comparable. A major difference between trisyllabic words of these two patterns is that, when deprived the right to make separate entries, trisyllabic words under the general [2+1] pattern are more suitable to be treated as subordinate entries, while those under the general [1+2] pattern are better used in examples.
References


The Necessity of Compiling a Learners’ German-English-Chinese Etymological Dictionary

Changliang Qu
Dalian University of Foreign Languages
choryo55@yahoo.com

Abstract

Although historic phonology sounds highly theoretical, its principles are actually applicable in current language teaching practice in China. While students who major in English are expected to choose a second foreign language as a required course, Western European languages like German, French and Spanish are usually their most frequent choice due to their formal similarity and historical proximity to English. But unlike French that has exported a large vocabulary to English, the cognates between German and English are more typically reflexes from a shared ancestral Germanic language than results of direct borrowing. As a result, in English vocabulary it is often more difficult to recognize a German cognate than a French one, because phonological and semantic evolution has concealed much of their formal similarity. Quite often, learners do not automatically benefit more from the fact that German and English belong to the same branch of the Germanic language family. Therefore, students who begin to learn German after they become advanced learners of English may need a new type of etymological dictionary that facilitate their learning. This dictionary is not merely designed for the “élite” students who will deepen in theoretical linguistics, but for a larger group of learners who intend to grasp more effectively the vocabulary of a second foreign language. The present paper aims to explore how historical phonology is applied in preparing such a dictionary and what etymological information should be presented in this dictionary.

Keywords: cognates, etymological dictionary, historical phonology

1. Introduction

Etymological information, although usually forming an inseparable part in the entries of a dictionary for native speakers, is often considered unnecessary in a dictionary specially designed for learners of foreign languages. This partly explains the reason why etymological information is absent in many successful dictionaries of the latter type, like Oxford Advanced Learners’ Dictionary of Current English and Longman Dictionary of Contemporary English. However, when learners are trying to learn a third language historically related to their second language that they have largely
grasped, the need for such information manifestly arises for its role in vocabulary comparison between the second and the third language.

In China nearly all students who major in English language and literature are learners of a third language, since “the second foreign language” is a required course for both their undergraduate and graduate studies. Traditionally this course intends to offer them a better understanding of language structure, while nowadays grasping one more foreign language often becomes a skill that makes them more competitive in the job market. In practice, it is either French, or German, or Spanish that the students frequently choose to be their “second foreign language”. An obvious reason for this choice is these languages’ formal similarity and historical proximity to English, the language that they have grasp enough to produce positive transfer.

Based on historical linguistics, a comparison between German and English vocabulary is helpful for the students who have largely grasped English. But unlike French that has directly exported a large vocabulary to English, the cognates between German and English are more typically reflexes from a shared ancestral Germanic language than results of direct borrowing. Consequently in English vocabulary, it is often more difficult to recognize a German cognate than a French one, because phonological and semantic evolution has concealed much of their formal similarity. Quite often, learners do not automatically benefit more from the fact that German and English belong to the same branch of the Germanic language family. Therefore, the students who begin to learn German after they have become advanced learners of English may need a new type of etymological dictionary that facilitates their German learning.

Unfortunately, the etymological information is missing in the most frequently used learners’ dictionaries in China. It is given in neither the locally compiled Deutsch-Chinesisches Wörterbuch, nor the imported Langenscheidts Großwörterbuch Deutsch als Fremdsprache. And traditional classics like Kluge’s An Etymological Dictionary of the German Languages now look fairly intimidating to the average students. Therefore, it becomes necessary to compile a learners’ German-English-Chinese etymological dictionary which is not merely designed for the “élite” students who will deepen in theoretical linguistics, but for a larger group of learners who intend to grasp more effectively the vocabulary of a second foreign language. In this concise and practical dictionary, the comparison between the German and the English vocabulary should be conducted both phonologically and semantically.

2. Phonological Evolution as the Theoretic Basis for the Proposed Dictionary

2.1 The German side of the phonological evolution

An obvious necessity for compiling such a dictionary lies in the fact that the processes of phonological evolution on both the German side and the English side have long concealed the resemblance of the cognates. On the German side, compiling such a dictionary involves a revived interest in the “High German consonant shift”, the
phonological law familiarly known as the “second Germanic consonant shift” when compared to the Grimm’s Law (the first Germanic consonant shift). This shift is highly visible in any basic German wordlist.

Classical works on historical linguistics typically describe the two or three core stages, sometimes plus a few other stages, of the evolutionary process that made the High German consonantal system distinguishable from those of the other Germanic languages and dialects. The earlier of the core stages, in which the three Germanic voiceless plosives /p, t, k/ became their correspondent fricatives /f, s, x/ or affricates /pf, ts/, is prevalent, although the differences in their vowels often conceal their resemblance. When the etymological information is available, the vague correspondence between the English-German cognates looks more straightforward. For example:

schlafen (< OHG. slâffan) – sleep (< OE. slæpan)
essen (< OHG. eʒʒan) – eat (< OE. etan)
suchen (< OHG. suohhan) – seek (< OE. sēcan)
Pfund (< OHG. pfunt) – pound (< OE. pund)
Herz (< OHG. herza) – heart (< OE. heorte)

Equally highly visible is a later stage, in which the three voiced plosives /b, d, g/ became voiceless in High German. This evolutionary mostly occurs at the coda position, but orthographically the pre-evolution spellings of the final consonants have remained, making this rule widely known as “Auslautverhärtung” (final-obstruent devoicing) in almost all the beginners’ textbooks. Probably the only case of this stage that needs explanations is the initial German /tl/ corresponding to the English initial /d/, e.g. tanzen (dance), Tür (door), trinken (drink) etc., since the historical devoicing did not happen to the initial German /b/ and /g/.

Old High German experienced other consonantal changes after the High German consonant shift. In the German-English vocabulary comparison, these more recent changes are reflected in the correspondence between German /b/ and English /f, v/, as well as that between German /d/ and English /θ, ð/:

Leben (< OHG. ) – life (< OE. )
oben (< OHG. obana) – over (< OE. ufân)
durch (< OHG. duruh) – through (< OE. þurh)
da (< OHG. dâr) – there (< OE. þær)

In all these examples, consonantal shifts to some extent have hidden the historical closeness of the cognates. But what makes them even more unrecognizable is the vocalic differences. In classical works on German or Germanic historical phonology, the evolution of the vowels from Old High German to Modern German is a profoundly complicated and thoroughly discussed issue (e.g. Helfenstein 1870, Wright 1907 etc.). However, a practical dictionary for learners may well circumvent this sophistication. In most of the cases, the vocalic evolutionary correspondences are not easy to explain
smoothly. A lot of the cognates had already contained distinct vowels in the age of Old High German / Old English. In this case a description (rather than an explanation) is enough. For example:

gruß (< OHG. gruoʒ) – great (< OE. grētan)
hören (< OHG. hôren) – hear (< OE. hýran)
mehr (< OHG. mēro) – more (< OE. māro)
tot (< OHG. töt) – dead (< OE. deād)

2.2 The English side of the phonological evolution

The consonantal shifts took place not only on the German side, but on the English side. One of the most crucial change is the /g/ > /j/, w/ shift in Early Modern English that hide much of the resemblance of the German-English cognates in both pronunciation and spelling.

This shift is a cause for the resurgence of a new English diphthong system after the Old English diphthongs (ea, eo, êa, êo, eō etc.) had all disappeared during the transition to the Middle English (Jespersen 1909: 94). The new English diphthongs are combinations of vowels and /j/ or /w/ that shifted from /g/. These new diphthongs makes the English words look quite different from their German cognates that experienced no such change. In other words, these German cognates are much less recognizable. For example:

OE. æg, eg, āg > /æ(ː)i/ > /ei/:
day (< OE. dæg) – Tag (< OHG. tag)
sail (< OE. segel) – Segel (< OHG. segal)
OE. āw, āg > /ɔːu/ > /au/:
sow (< OE. sāwan) – sæen (< OHG. sāen)
dough (<OE. dāg) – Teig (< OHG. teig)
OE. aw, ag > /au/:
thaw (< OE. þawian) – tauen (< MHG. touwen < OHG. douwen)
gnaw (< OE. gnagan) – nagen (< OHG. gnagan)

Another wide-spread consonantal shift on the English side is the loss of /ç/ and /x/. The distribution of these two fricatives in Old and Middle English used to be the same as in German today: /ç/ appears after a front vowel and /x/ after a back vowel. With the loss of the two fricatives in Early Modern English, the vowels before them experienced compensatory lengthening, making the Modern English pronunciation distinctive from that of their German cognates. However, since the now silent English gh is retained orthographically, the German cognates are not very difficult to recognize:

light (< OE. lēoht) – Licht (< OHG. lioht)
right (< OE. riht) – recht (< OHG. recht)
daughter (< OE. dohter) – Tochter (< OHG. tohter)
high (＜OE. hēah) – hoch (＜OHG. hôh)

In addition, compared with the consonantal system, the English vocalic shifts happened more drastically but more regularly. In the stage of Early Modern English, the Great Vowel Shift gradually occurred as a systematic set of changes that deformed the whole pattern of long vowels. But since this change took place after instead of before the orthographic customs became stabilized, it did not let the German-English cognates look highly different from each other.

3. Semantic Information in the Word Entries

3.1 Faux-amis

A vocabulary comparison between any two or more cognate languages has to face the problems caused by faux-amis, i.e. cognate words in these languages that look or sound highly similar, but differ or only partially overlap in meaning. Famous German-English cases include:

German *aktuell*, which means not only “actual” but “current”, “relevant” or “topical”;

German *bekommen*, which means “get” or “receive”, and never “become”;

German *man*, an indefinite pronoun that roughly equals to English “one”, is never used as a noun meaning a male person. (The latter should be spelled as “Mann”.)

Since all these faux-amis are historically related to their counterparts in English, they need a special precaution mark in the proposed learners’ etymological dictionary.

3.2 Lexeme loss in Modern English

When a German word looks extremely unfamiliar for those who have grasped much English, it is possible that the lexeme is only found in High German, or that it may be found in several other Germanic languages but not in English. The former case is represented by words like *jetzt*, while the latter case is represented by *aber* (＜OHG. abur), which also appeared in Gothic as *afar* but not recorded in Old English or Old Norse (Walshe 1951: 2), though Kluge (1891: 2) reminded us of its relations to the Old English noun *eafora* “descendant”. Such information is academically important, but not crucial in a learners’ etymological dictionary.

What makes more sense is the German words that are related to some lost Old English lexemes replaced in the history either by French loans or by some other native lexemes. For example, the following German words, though unfamiliar in Modern English, were once active in Old English:

Beispiel (＜OHG. bîspel) is a cognate of OE. *spell* (meaning “tale”, one of the two components in “gospel”). In Middle English, it was replaced by the Old French word *example*;

*drucken* (＜OHG. drucchen) is a cognate of OE. *þryccan* (meaning “oppress”).
This lexeme also disappeared from the later English vocabulary while press, oppress and print are all from French;

Heirat (< OHG. hîrât) is a cognate of OE. hîred (meaning “household”). With the disuse of this lexeme in later English, both the newly introduced words family and marriage came from French.

Besides the above cases that Baugh and Cable (1993: 164) call “French words that poured into English”, there are some other cases that the German-English cognate words were replaced by another Anglo-Saxon lexeme. For example:

heiß en “be called” (< OHG. heiʒan) is a cognate of OE. hātan, now replaced by another native word call;

heute “today” (< OHG. hiutu) is a cognate of OE. heðæg, now replaced by today.

A few Old English words still have reflexes in Modern English, but have turned into marginal words that are no longer as active as before. For example:

riechen “smell” (< OHG. riohhan) is a cognate of OE. rēocan. Its descendant reek still exists in English today, but has fallen into the vocabulary of low frequency.

schmecken “taste” (< OHG. smecchen) is a similar case. Although it is a commonly used word in German, its cognate smack (< OE. smæccan) sounds more literary than spoken.

All the above information deserves a serious place in the proposed learners’ etymological dictionary.

3.3 The meaning shifts in German-English cognates

When cognate languages evolve on separate paths, their cognate words may also diverge semantically, causing constant semantic shifts that sometimes make them unrecognizable.

Campbell (2008: 266-267) visualizes the process of meaning shift with a model that consists three stages: When the lowercased letter a stands for a word, its meaning may shift from “A” via “A” + “B” to merely “B”.

If the shift happened in one language but did not happen in a language related to it, the meaning contrast, sometimes aided by phonological shift, can be strong enough to make the cognates unrecognizable. Furthermore, the cognates in different languages may experience different metaphors. Among the most frequently used German words, the following words have witnessed this type of semantic evolution. Unsurprisingly they seldom remind the learners of their English cognates:

klein “small” (< OHG. kleini “elegant”) – clean (< OE. clēne)
laufen “run” (< OHG. hlauffan) – leap (< OE. hlēapan “run, leap”)
Teil “part” (< OHG. teil) – deal (< OE. dēl)
Tisch “table” (< OHG. tisc) – dish (< OE. disc)
Zug “train” (< OHG. zug) – tug (< OE. tyge)

Some meaning shifts are not only semantic but morphological. For example, German
verb *fahren* (< OHG. *faran*) “ride” is a cognate of Old English verb *faran* “ride” which is no longer used in Modern English. However, *fare*, a noun derived from this verb is still commonly used today, with its meaning “the cost of a ride”. Similarly, the German adjective *krank* “ill” (< MHG. *kranc* “weak”) is a cognate of Old English *cranc* “feeble”, an adjective that Kluge (1891: 191) labeled as “occur[ring] rarely”. But *cringan* “writhe”, the verb form of this lexeme, was much more common, evolving into *cringe* in Modern English. Without an appropriate etymological dictionary, it is almost impossible for learners today to expect that German *krank* and English *cringe* are cognates.

4. Conclusion and Sample Dictionary Entries

To sum up this essay, we can take a look at how some of the above mentioned cognates will be processed as the word entries in this proposed dictionary. Since this dictionary is mainly designed for the Chinese learners who do not have many opportunities to be exposed to the Western-styled dictionaries of etymology, it should practically be a German-English-Chinese trilingual dictionary. The information in each entry is expected to be concise so that the dictionary will be easy to browse and carry.

Entry Category 1#, a cognate with a mere phonological shift:

**schlafen** (< OHG. *slâffan*) *vi*. schläft, schlief, hat geschlafen || sleep (< OE. *slêpan*) *vi*. 睡觉

Entry Category 2#, words with the risks of faux-amis:

**bekommen** (< OHG. *bi + chuēman* “come”) *vi*. bekam, hat bekommen || get, receive 得到，获得 ▲ Never “become”.

Entry Category 3#, the loss of an English lexeme:

**Heirat** (< OHG. *hîrât*) *f.*, -en || marriage *n*. 婚事 □ cog. OE. *hīred* † “household”.

Entry Category 4#, with meaning shift:

**klein** (< OHG. *kleini* “elegant”) *adj.* || small *adj*. 小的 □ cog. OE. *clēne* > clean.

Entry Category 5#, with meaning shift together with morphological issues:

**krank** (< MHG. *kranc* “weak”) *adj.* || ill *adj*. 患病的 □ cog. OE. Cranc † “feeble”. ► OE. *cringan* “writhe, twist in pain” > cringe.
Acknowledgements

This work is part of the Project “A Study on Otto Jespersen’s Ideas of Phonological Evolution (1886-1941)”, supported by National Social Science Foundation of China (Grant Number: 16BYY007).

References


The Treatment of Polysemous Entries in Spanish-Chinese Bilingual Dictionaries

Yuqing Deng
Universitat Autònoma de Barcelona
Yuqing.deng@qq.com

Abstract

This paper discusses the treatment of polysemous entries in Spanish-Chinese bilingual dictionaries, represented by six dictionaries of this type consulted in their last edition published in the twenty-first century. It examines first the entry structure and structural indicators employed in these bilingual dictionaries. Then, it analyses the semantic subdivision of a sample polysemous word in these dictionaries in comparison with its meaning structure as presented in a Spanish monolingual dictionary. After that, the equivalents of the sample word are examined from the perspective of equivalence typology. Finally, the devices used for meaning discrimination are also discussed with the sample entries. It is found that all these Spanish-Chinese bilingual dictionaries examined present a hierarchical and equivalent-oriented structure in polysemous entries, predominated by translational equivalents. As for the meaning discrimination of the equivalents provided, devices such as co-text, context, illustrative examples and usage labels are more frequently used in these bilingual dictionaries. During the discussion in each aspect, comments and suggestions are also formulated. It is suggested that the structural indicators should be used carefully and coherently to help construct the entry structure in the dictionary as well as the mental lexicon of the dictionary user. In the selection of equivalents for polysemous entry words, factors like efficiency, generative power and cultural acceptability should be taken into consideration, and it is suggested to cover the semantic space of a polysemous entry word to the utmost extent with the least number of equivalents by offering equivalents that are applicable to a broader range of contexts in the target language and with greater descriptive power. Also, various devices should be made good use of to indicate the associations between the equivalents and discriminate their meanings.

Keywords: Polysemy, bilingual dictionary, Spanish-Chinese
1. Introduction

The redaction of polysemous entries in a bilingual dictionary is without doubt a demanding task, in which a series of problems should be considered. An analysis of the treatment of these entries may help to reveal the decisions made by the lexicographers when dealing with those problems and indicate the future path for the improvement of bilingual dictionaries. This article intends to analyse the treatment of polysemous entries in Spanish-Chinese bilingual dictionaries from several aspects. First, it examines the entry structure and structural indicators employed in six bilingual dictionaries selected for the analysis, which help to obtain an accurate interpretation of the semantic subdivision of polysemous entry words and the semantic relations between the equivalents provided. Then, the semantic subdivision of a sample word in these bilingual dictionaries is analysed in comparison with its meaning structure as presented in a Spanish monolingual dictionary, thus revealing whether the entry structure adopted by bilingual dictionaries is equivalent-oriented or meaning-oriented (Jarošová 2000). After that, the equivalents of the sample word are examined from the perspective of equivalence typology (Zgusta 1971: 319-322; 1978; 1984). Finally, the devices used for meaning discrimination are also illustrated and discussed with the sample entries. During the discussion in each aspect, comments and suggestions are formulated as to how to improve the treatment of polysemous entries in bilingual dictionaries.

For the analysis carried out in this article, six Spanish-Chinese bilingual dictionaries are selected as representatives: *Nuevo diccionario español-chino* published by The Comercial Press (2003, referred to as NDEC-BJ in this article), *Nuevo diccionario español-chino* by Shanghai Translation Publishing House (2012, referred to as NDEC-SH), *Diccionario español-chino chino-español* (2011, referred to as DECC), *Diccionario manual español-chino* (2001, referred to as DMAEC), *Diccionario moderno español-chino* (2010, referred to as DMoEC) and *Conciso español-chino chino-español diccionario* (2011, referred to as CECCED). In order to reflect the latest techniques applied to the treatment of polysemous entries, all these dictionaries are consulted in their last edition published in the twenty-first century. For the comparison between the semantic subdivision of an entry word in bilingual dictionaries and its meaning structure in the source language, the Spanish monolingual dictionary *Diccionario de uso del español de América y España* (hereinafter referred to as DUAED) is consulted; and for the clarification of meaning of the equivalents provided in the target language, *The Contemporary Chinese Dictionary [Chinese-English Edition]* (hereinafter referred to as CCD) is also consulted.

2. Hierarchical structure and structural indicators

A comparison of the “Guide” to the six Spanish-Chinese dictionaries (or the Spanish-Chinese section of bidirectional bilingual dictionaries) selected reveals that
there is a series of common standards followed by all of them in their microstructure, although considerable differences are also detected. As for polysemous lexical units, the entry structure actually shows a higher degree of similarity between these dictionaries than what has been stated in the guide.

In all the six dictionaries, a system of structural indicators is employed to exhibit not only the division of semantic space of an polysemous entry word, but also the degree of semantic relatedness among its equivalents in the target language. Five of these dictionaries, namely, *NDEC-BJ, NDEC-SH, CECCED, DECCE* and *DMaEC*, use Arabic numerals to lead meaning subsections— or translation meanings, as are called by Ďurovič *(apud Jarošová 2000: 14)*— of the entry word, in which equivalents that are close synonyms are usually separated by a comma, while those that are near synonyms or with a semantic relation relatively weak, usually by a semicolon. For example, one of the translation meanings of the Spanish polysemous word *ultimo* is presented as follows in *NDEC-SH*:

**último, ma adj.** 1. 最后的，末了的：末尾的 [...] 

Therefore, the presentation of polysemous entries in the five bilingual dictionaries shows a hierarchical structure of at least three levels\(^1\), which can be represented by the following diagram:

![Hierarchical structure of polysemous entries](image)

**Figure 1** Hierarchical structure of polysemous entries in Spanish-Chinese bilingual dictionaries (equiv. = equivalent)

In the case of *DMoEC*, he only exceptional of the six bilingual dictionaries examined, there are two levels of semantic subdivision in polysemous entries. As stated in its guide (p. 9), a semicolon is used to separate translation meanings, in which the corresponding equivalents are delimited by a coma. Therefore, it can be interpreted that in *DMoEC* the level 2 of the hierarchy mentioned has been omitted.

Nevertheless, in spite of the general agreement on adopting an hierarchical structure for polysemous entries in practice, not all the dictionaries explain this structure and the function of the structural indicators clearly enough in their guide of

---

\(^1\) To distinguish the first two levels of the semantic subdivision, we resort to the term used by Ďurovič for level 1, that is, *translation meaning*, and reserve the term *subsection* for level 2.
use. *NDEC-BJ* (p. II) and *CECCED* (p. 1) only mention the use of Arabic numerals for introducing translation meanings, leaving the interpretation of the use of semicolons and commas and the perception of the relatedness among the equivalents to the user of the dictionary. It is suggested that the guide to the dictionary should provide a full explanation of all the devices used as structural indicators in polysemous entries, for they convey important information on the semantic subdivision of the entry word.

It is also reasonable to suggest that the structural indicators should be used carefully and coherently in the entries throughout a dictionary, especially a traditional one in paper format and predominated by linear texts. Without the aid of dendrograms, hypertexts and other devices, the structural indicators and the underlying hierarchy in polysemous entries are of great importance not only for the construction of entry structure in the dictionary, but also for the organization or categorization of concepts in the mental lexicon of the dictionary user.

### 3. Semantic subdivisions and equivalent-oriented structure

Since the lexical structure is not identical in different languages, the redaction of polysemous entries in a bilingual dictionary faces the problem of whether to preserve the meaning structure of the words in the source language as presented in a monolingual dictionary or to arrange the entry structure according to the semantic relations between the equivalents in the target language.

Among the six Spanish-Chinese bilingual dictionaries examined, only *NDEC-SH* mentions in its preface (p. 1) that monolingual Spanish dictionaries were consulted as reference during the compilation of the bilingual dictionary; and only *CECCED* states that its purpose is to provide the user with “a rapid and reliable translation”\(^2\) (pp. 1-2), which is in accordance with its title *Conciso español-chino chino-español diccionario*. Within the dictionary typology, a *concise dictionary* usually contains a limited number of lexical units with a minimal definition in the case of a monolingual dictionary, or with its equivalents in the target language in case of a bilingual one (Campos Souto and Pérez Pascual 2003: 76). The other four bilingual dictionaries, in contrast, do not state their decision in this regard explicitly. In order to identify the approach adopted by all these dictionaries, we take the Spanish polysemous verb *conocer* (“know”) for an example and examine its semantic subdivision in these bilingual dictionaries, in comparison with its meaning structure presented in the Spanish monolingual dictionary *DUAE*. The definitions of this verb provided by *DUAE* are as follows:

**conocer**  
1. Tener en la mente información sobre algo. (To have in mind information about something.)  
2. Tener información o conocimiento profundos y con experiencia directa sobre algo. (To have sufficient information or a deep

\(^2\) Text originally available in Chinese and in Spanish, translated in English by the author of this article.
knowledge of something with direct experience.)

[3] Tener experiencia interiorizada acerca de algo. (To have internalized experience of something.)

[4] Distinguir o identificar a una persona o cosa entre varias por una serie de características propias. (To distinguish or identify someone or something among several by a series of particular characteristics.)

[5] Tener trato o relación con una persona. (To be acquainted with or have contact with a person.)

[6] Ver o tener trato por primera vez con una persona. (To meet or come into contact with a person for the first time.)

[7] Tener una idea formada sobre el carácter o la manera de ser de una persona. (To have an established idea of the character or the way of being of someone.)

[8] Tener información directa de un lugar por haber estado en él. (To have direct information of a place for having been there.)

[9] *formal* Tener relaciones sexuales con una persona. (*formal* To have sexual relations with a person.)

By comparing the definitions of the verb in the monolingual dictionary with its translation meanings in Chinese provided by the bilingual dictionaries and making use of the illustrative examples for meaning discrimination (v. § 5), we manage to roughly make a match between them, and the results are presented in the following table, in which the definitions of *conocer* in *DUAE* are represented by the numerals:

**Table 1** Correspondence between the definitions of *conocer* in *DUAE* and its translation meanings in the Spanish-Chinese bilingual dictionaries

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>新西汉词典</td>
<td>新西汉词典</td>
<td>西汉西词典</td>
<td>简明西汉词典</td>
<td>现代西汉词典</td>
<td>精选西汉词典</td>
</tr>
<tr>
<td>[2] [7] [8]</td>
<td>2.懂，会</td>
<td>2.熟悉</td>
<td>2.熟悉</td>
<td>2.懂，会</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[4]</td>
<td>5.注意到，察觉到</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[4]</td>
<td>5.推测，预测，判断</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[3]</td>
<td></td>
<td>6.（凭经验）</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Several observations can be made from this table. First of all, concerning the number of translation meanings presented, there seem to be three levels of density that can be assigned to the entries for *conocer* in these bilingual dictionaries. *NDEC-BJ* and *NDEC-SH* show a highest level of density presenting up to nine translation meanings each; *DECCE* and *DMaEC* present a moderate level of density with seven meanings each; while *DMoEC* and *CECCED*, with the minimum level of density, only list three or four meanings each. Secondly, there seems to be a general agreement among the bilingual dictionaries on the selection of main equivalents for the verb, especially in the first three to four translation meanings. Finally but most interestingly, although the density of entry varies considerably, the match between the definitions in Spanish and the translation meanings in Chinese reveals that all these bilingual dictionaries manage to cover the meaning of the absolute majority of the senses of *conocer* registered by the monolingual dictionary. This implies that the density of a polysemous entry in a bilingual dictionary is not merely determined by its coverage of the lexical meaning of the entry word. Further analysis shows that the density is also influenced by the decisions made in the semantic subdivision and in the meaning discrimination.

As is observed in Table 1, in this case there is not a perfect one-to-one correspondence between the translation meanings of the word in the target language and its definitions in the source language. Taking the entry in *NDEC-SH* as an example, the correspondence is actually more complicated:
Figure 2 Correspondence between the translation meanings of the Spanish polysemous verb *conocer* in *NDEC-SH* and its definitions in *DUAE*

Reorganizing the content in Figure 2, the correspondence between the semantic subdivision of the selected verb in *NDEC-SH* and its meaning structure in *DUAE* can be illustrated more clearly in the following way:

Figure 3 Correspondence between the semantic subdivision of the Spanish polysemous verb *conocer* in *NDEC-SH* and its meaning structure in *DUAE*

The illustration shows that the senses of a polysemous word with the same equivalent, or with equivalents that are synonyms or near synonyms in the target language, can be joined together in the bilingual dictionary. For the polysemous verb *conocer*, the senses [1], [7] and [8] are joined in the same meaning subsection or translation meaning, and so is the case of the senses [5] and [6]. Such “compaction of information” is considered by Atkins (1996: 523) as a subtle distortion of the source language by the target language, “in order to make the bilingual dictionary better, allowing, for instance, a very brief entry”. On the other hand, equivalents for the same sense of an entry word
can be assigned to different translation meanings, specified by semantic restrictions (e.g. the senses [2] and [4] of conocer), syntactic restrictions (e.g. the senses [1] and [5] of conocer), etc. (v. § 4 and especially § 5), and consequently the semantic space occupied by the same sense of the entry word is also divided into different subsections in the bilingual dictionary. It is logical to suppose that the more sophisticated the subdivision is, the higher the density of entry is likely to be in the bilingual dictionary.

The comparison between the semantic subdivision of the entry word in NDEC-SH and the meaning structure in DUAE reveals that, although monolingual dictionaries have been consulted during the compilation of the bilingual dictionary, an equivalent-oriented structure rather than a meaning-oriented one is adopted in the polysemous entry. The other bilingual dictionaries examined vary in the coverage of the lexical meaning of the polysemous verb, but the semantic subdivisions remain roughly the same as that in NDEC-SH. In other words, all these six bilingual dictionaries adopt, at least in this case, an equivalent-oriented approach to construct the entry structure.

This approach, when adopted in polysemous entries, helps to avoid redundant repetitions of the same equivalents and, by allowing “looking at the source language through the prism of a target language” (Jarošová 2000: 19) it makes the dictionary much easier to use for the reception or decoding of the source language. However, as is commented by Atkins (1996: 523), it does prevent the user from gaining “a clear view of the potential of the SL item, which must be sought in a monolingual work”. She also admits that an ideal bilingual dictionary able to satisfy all needs is impossible in a printed work.

4. Translational equivalents and explanatory equivalents

According to Zgusta (1971: 319-322; 1978; 1984), two types of equivalents are distinguished in bilingual dictionaries. The translational equivalent is a lexical unit which can “be inserted into contexts of the target language” and “used in a fluent, good translation of whole sentences” (Zgusta 1971: 319), whereas the explanatory equivalent is usually not a fully stabilized lexical unit, but a “non-minimal” expansion which can “give more information about the lexical unit of the target language”. However, there are also equivalents that combine both advantages (Zgusta 1971: 319; 1978: 543).

Among the bilingual dictionaries examined, NDEC-BJ, NDEC-SH and DMAEC provide examples that illustrate the usage of the entry word in the source language together with their corresponding translation in the target language, which facilitate to a large degree the identification of the equivalent type in theses dictionaries. In the case of the polysemous verb conocer, we find that the majority of its equivalents have been successfully inserted into the translation of the examples, and the rest can either replace an equivalent used in the translation or be inserted into another context in the target language. That is to say, all the equivalents provided by these bilingual dictionaries for the Spanish verb conocer show what is called by Zgusta (1984: 147)
the translationality or insertability. However, some of the equivalents do not make a
translation “good” enough or natural enough in the target language, and actually show
a greater explanatory or descriptive power than their potential of being as immediate
translation.

One of these equivalents with great explanatory power is “与…相识” (yǔ… xiāng shí, ‘to be acquainted with…’) in the second translation meaning provided by
NDEC-SH. This phrase is less frequent in the target language than the other equivalent
for the same meaning, “认识” (rèn shí, here ‘to know [someone]’), and it is more
commonly used in written language or in a more formal style of spoken language. Yet
since “认识” is also polysemous in Chinese, the explanatory equivalent “与…相识”
not only provides the user with another expression for stylistic variation but also helps
to disambiguate the polysemous word in the target language (v. § 5.1).

Another example is the case of the equivalents “有过性经验” (yǒu guò xìng jīng yàn, ‘to have had sexual experience’) in NDEC-BJ and “发生性关系” (fā shēng xìng guān xi, ‘to have sexual relations’) in NDEC-SH and DMaEC. Although in
NDEC-SH the latter equivalent is directly inserted into the translation for the sentence
Hasta el día de su boda no conoció varón, as “婚礼之前她未与男人发生过性关系”
(‘Until the day of her wedding she had no sexual experience with a man’), the
translation seems to some extent straight and rigid in the common use of the target
language either from the cultural perspective or from the perspective of translation
strategies. The sentence in the target language makes a good explanation but not a
good enough translation. A better choice would be to replace the explanatory
equivalent used in the translation with another expression, such as “碰” (pèng, literally
‘to touch’). Then the translation would be “婚礼之前她没碰过男人” (‘Until the day
of her wedding she had never touched a man’), which not only is more acceptable in
the target language but also preserves the euphemism in the source language.
Meanwhile, it may inspire the user of the dictionary to imagine other possible
translations, thus enhancing the generative power of the bilingual dictionary.

The group of equivalents “分辨” (fēn biàn, ‘to distinguish’), “辨别” (biàn bié, ‘to
distinguish’), “识别” (shí bié, ‘to identify; to distinguish’) and “区别” (qū bié, ‘to
distinguish; to differentiate’) provided by NDEC-BJ and partially shared by the other
bilingual dictionaries except DMoEC, which has not included this meaning, presents a
more delicate issue. Although all these equivalents can be inserted into a context in the
target language, only one of them —“识别” (shí bié, ‘to identify; to distinguish’)– is
actually used for once in the translation of the examples for the entry word. For the rest of the translation under this meaning, the expressions used are “认” (rèn, ‘to recognize; to identify’), “识” (shí, ‘to identify; to distinguish’), “挑” (tiāo, ‘to select’) and “知道” (zhī dào, in the sense of ‘to know, to tell’), among which “认” (rèn, ‘to recognize; to identify’) is counted four times. As can be seen, the translation in Chinese of the Spanish verb conocer in this sense corresponds to the sememes “to recognize”, “to identify” and “to distinguish” (or others that comprise or imply such actions, as are the cases of “挑” and “知道”), mostly shared by the four equivalents provided by the bilingual dictionaries. However, the expressions used in the translation can not be replaced by these equivalents if a good translation is pursued. Therefore, these equivalents are actually more explanatory than translational.

Besides, being “分辨”, “辨别”, “识别” and “区别” explanatory equivalents and synonyms, there is no need to indicate too many of them in a bilingual dictionary. It is reasonable that NDEC-SH, DECC, DMaEC and CECCED only select two of these equivalents each. It is worth noticing that all these four dictionaries exclude “区别” and include “识别” together with one of the other two equivalents. According to the dictionary of the target language CCD, “区别” seems to emphasize on the concept of differentiation, which is relatively far from the lexical meaning of the Spanish verb conocer in this sense; whereas “识别” comprises the sememes ‘to identify’ as well as ‘to distinguish’. Since “[t]he lexicographer’s task is to indicate the most general translational equivalents which have a broad range of application” (Zgusta 1978: 545) and “equivalents that are both translational and explanatory are the bilingual lexicographer’s best friends” (Zgusta 1984: 148), it is suggested to provide, in this case, other equivalents such as “认出” (rèn chū, ‘to recognize’) or “辨认” (biànrèn, ‘to recognize; to identify’) together with “识别”. Then the relevant sememes of the entry word in this sense would be covered with the least number of equivalents, which meanwhile would be applicable to a broader range of contexts in the target language. Additionally, the characters “认” and “识” in these equivalents are shared by the equivalent “认识” (rèn shí) in another translation meaning, which could imply the associations between the two meanings, thus facilitating the perception of the lexical meaning as a whole for the user of the dictionary.
5. Meaning discrimination of the equivalents

When there are several equivalents in the target language corresponding to a lexical unit in the source language, Werner (1982: 289) distinguishes two possibilities: (i) the equivalents in the target language correspond to the word in the source language according to its different sememes, or (ii) in the target language there are certain rules (collocations, syntactic restrictions, dependence on factors of a specific communicative situation, etc.) or frequency of use that affect the distribution of the equivalents in the translation. However, “[i]t is very hard for a bilingual dictionary user to tell if a word in Language A means the same as an unknown word in Language B, far less whether they diverge in style, register, collocational potential etc.” (Atkins 1996: 522). Therefore, meaning discrimination is especially important, if not necessary, in the treatment of polysemous entries in a bilingual dictionary.

According to Zgusta (1978: 540-541), the applicational restriction of an entry word can be stated in a gloss, which is more explicit, or in the form of one or more examples, in which the information is more immediate. Atkins (1996: 523) also proposes that “[t]he ideal dictionary should offer the skilled user the chance to make his or her own judgement on equivalences, by scanning examples of the TL items (grouped according to meaning) in various types of context, as well as - for contrastive checking purposes - examples of the relevant meaning of the SL item in a wide variety of contexts”. James Mafela (2005: 283-284) complements that “[t]o achieve meaning discrimination, use can be made of different devices such as providing short definitions of equivalents, indicating parts of speech, adding etymology and usage labels, and giving context words and phrases, and illustrative examples”.

In the six bilingual dictionaries examined, devices such as co-text, context, illustrative examples and usage labels are more frequently used for meaning discrimination although there are different levels of complexity and elaboration in this regard. By comparing the guide and the content of these bilingual dictionaries we find that NDEC-BJ, NDEC-SH and DMaEC provide a considerable amount of examples in the source language with translation in the target language and their system of usage labels is more complex and more clearly explained; whereas in DECCE, DMoEC and CECCED, exemplification is really rare and the system of usage labels is more simplified and only briefly introduced in the guide. Here we keep using the case of the polysemous verb conocer for the analysis in this section and focus on the use of co-text, context and usage labels for the meaning discrimination.

5.1 Semantic restrictions indicated by the co-text

According to CCD, the equivalent “认识” (rèn shì) is also polysemous in the target language and can be used as a verb (‘to know’) or as a noun (‘understanding or knowledge’). The indication of other verbs as equivalents in the same translation meaning of the entry word –that is, as synonyms or near synonyms of “认识”– in
NDEC-BJ, DECCE, DMAEC, DMOEC and CECCED designates the part of speech of this equivalent and disambiguate its meaning to some extent. In NDEC-SH, the presence of the explanatory equivalent “与…相识” (yǔ… xiāng shí, ‘to be acquainted with…’) helps to further narrow down its meaning by implying the semantic restriction that its two arguments should both be human. The translation of the examples and the context given in the target language in NDEC-BJ, NDEC-SH and DMAEC also help to discriminate the meaning.

The equivalent “了解” (liǎo jiě, ‘to know, to understand’, ‘to find out’) is also polysemous and disambiguated by the indication of other equivalents and the examples in NDEC-BJ, NDEC-SH and DMAEC. However, the variation in the combination and hierarchy of the four equivalents “知道”, “了解”, “熟悉”, “认识” among the six bilingual dictionaries demonstrates that they are not clearly discriminated. According to CCD, “知道” (zhī dào, ‘to know, to be aware of’), “了解” (liǎo jiě, ‘to know, to understand’), “熟悉” (shú xī, ‘to know well, to be familiar with’) and “认识” (rèn shì, ‘to know, to recognize [something]; to be acquainted with [someone]’) are close synonyms and are mutually replaceable in many contexts in the target language, but there are also subtle nuances. Rather than leave the perception of the relations and differences among them to the competence of the user of the dictionary, it is suggested to reorganize the combination and hierarchy of these four equivalents and a possible solution would be: 1. 知道, 了解, 熟悉. 2. 认识, 与…相识. Moreover, the fact that the complement of these equivalents can be a person, a subject or a place should also be indicated explicitly either in the form of a gloss or in the examples.

5.2 Semantic restrictions indicated by the context

As is shown in § 3, the equivalents corresponding to the sense [2] of the polysemous verb conocer are divided into two different translation meanings: the equivalents “懂” and “会” on the one side and “审理”, “处理” and “解决” on the other. Being isolated characters, “懂” (dǒng, ‘to know, to understand’) and “会” (huì, ‘to know how to do [something] or to be able to do [something]’) are rather translational equivalents with very limited descriptive power, and have to be complemented by examples and contexts. However, a semantic restriction can be indicated by adding another character “行” (háng, ‘profession, trade’) to “懂” as its complement, thus forming a word “懂（行）” in the target language meaning ‘to know the business; to know the ropes’. Since
“行” is also polysemous and susceptible of other confusions, it can be extended into a phrase “某一行业或领域” (‘a certain profession or field’). Other expressions with greater descriptive power such as “内行” (nèi háng, ‘to be knowledgeable about or experienced in an issue or certain work’) and “精通” (jīng tōng, ‘to be proficient in [learning, skill, or profession]’) can be either provided as equivalents in the translation meaning or used in the translation of examples, as in NDEC-SH and NDEC-BJ.

When the context of the verb conoscere is specified, for example, when talking about a legal case, another equivalent of this sense can be derived: “审理” (shěn lǐ, ‘to try [a case]’). NDEC-SH, DECCE and DMAEC add different context words to indicate the semantic restriction of this equivalent. In NDEC-SH, a gloss is added in parenthesis: “(有权限) 审理” (‘(to have the authority) to try [a case]’), with the emphasis on the sememe of ‘to be able to do’. Whereas in DECCE the semantic restriction is indicated on one argument: “审理 (案件)” (‘to try (a case)’); and in DMAEC on both arguments: “(法官) 审理 (案件)” (‘(a judge) to try (a case)’). By comparison, the equivalents provided by NDEC-BJ for the entry word in a similar sense, “处理” (chǔ lǐ, ‘to solve’) and “解决” (jiě jué, ‘to solve, to settle’), can be applied to a much broader range of contexts in the target language and their complement can be a legal case or an issue of a general type. Therefore, it might be an better choice to present these two words as equivalents in the translation meaning and provide the more specific use of “审理” (shěn lǐ, ‘to try [a case]’) in the translation of an example or some examples. By this means, the equivalents could cover a broader range of application of the entry word, its lexical meaning would be organized in a more logical way and the generative power of the bilingual dictionary could also be enhanced.

The case of the sense [4] of the polysemous verb is more obvious. The first group of equivalents comprises “分辨”, “辨别”, “识别” and “区别”, previously discussed in § 4, for which we have suggested the alternative equivalents “认出” (rèn chū, ‘to recognize’) or “辨认” (biàn rèn, ‘to recognize; to identify’) together with “识别” (shí bié, ‘to identify; to distinguish’). The second group consists of “推测” (tuī cè, ‘to infer’), “预测” (yù cè, ‘to predict’) and “判断” (pàn duàn, ‘to judge’) provided by
and the third group is of “注意到” (zhù yì dào, ‘to notice’) and “察觉到” (chá jué dào, ‘to perceive’) offered by NDEC-SH. With the aid of the examples given in the dictionaries, it is not difficult to tell that the first group emphasizes on the identification of a person or a thing by its own characteristics or by the differences with others; and that the second group focuses on the mental action based on the observation or perception of something; whereas the third group stresses the perception of something or some change. Since the three groups of equivalents differ in the focus and are applied in different contexts, it is suggested to formulate three separate but consecutive translation meanings for the sense [4] of the entry word.

5.3 Syntactic restrictions indicated by usage labels

As is observed in these bilingual dictionaries, NDEC-BJ, NDEC-SH, DECC and DMaEC all indicate separately the pronominal use of the Spanish polysemous verb conocer with a usage label leading the equivalents.

According to Battaner and Torner (2008: 209), the alternation of pronominal and non-pronominal forms of the same verb affects the syntactic construction but not the meaning of the verb. The equivalents “自知” (zì zhī, ‘to know oneself’) and “有自知之明” (yǒu zì zhī zhī míng, ‘to have the knowledge of oneself’) are actually expansions of the equivalent “知道” (zhī dào, ‘to know, to be aware of’) by adding the complement oneself, and in the same way “了解自己” (liǎo jiē zì jǐ, ‘to know oneself’) is an expansion of the equivalent “了解” (liǎo jiě, ‘to know, to understand’). These three expressions are equivalents for the reflexive use of the verb conocer in the sense [1]. Similarly, the expression “互相认识” (hù xiāng rèn shì, ‘to be acquainted with each other’) is the expansion of the equivalent “认识” (rèn shì, ‘to know [someone]; to be acquainted with’), exhibiting the meaning of ‘reciprocal’ of the pronominal use of the verb in the sense [5]. Therefore, although separated from other translation meanings, the equivalents marked by the usage label for pronominal use do not create a new translation meaning, but rather specify a certain meaning that already exists by revealing the information embedded in the syntactic construction.

6. Conclusion

In this article we have discussed several aspects of the treatment of polysemeous entries in Spanish-Chinese bilingual dictionaries: the entry structure, the semantic subdivision
of the entry word, the type of equivalents provided and the devices used for meaning discrimination.

It is found that the Spanish-Chinese bilingual dictionaries examined present a series of common features in their microstructure despite certain discrepancies. First, all these dictionaries adopt a hierarchical structure when organizing the equivalents in polysemous entries with a system of structural indicators that exhibit the semantic subdivision and the degree of semantic relatedness among the equivalents. Second, the structure of polysemous entries in these bilingual dictionaries is oriented by the semantic relations between the equivalents in the target language rather than by the meaning structure of the entry words in the source language. Third, most of the equivalents provided by these dictionaries can be inserted immediately into contexts of the target language, but some equivalents show a greater descriptive power. Finally, for the meaning discrimination of the equivalents, devices such as co-text, context, illustrative examples and usage labels are more frequently used in these bilingual dictionaries.

In all these respects, the six bilingual dictionaries examined present different levels of complexity and elaboration. On the whole, the treatment of polysemous entries in *NDEC-BJ*, *NDEC-SH* and *DMaEC* is evaluated better, with an entry structure clearly constructed, a relatively more elaborate semantic subdivision and an effective use of illustrative examples and other devices for meaning discrimination. The guides to these three dictionaries also provide a more detailed explanation of the main features of the respective dictionaries.

For future bilingual dictionaries, it is suggested that the structural indicators in polysemous entries should be used carefully and coherently in order to help construct the entry structure in the dictionary as well as the mental lexicon of the dictionary user. In the selection of equivalents for polysemous entry words, factors like efficiency, generative power and cultural acceptability should also be taken into consideration besides the correspondence between the equivalent and the entry word. Therefore, it is suggested to cover the semantic space of a polysemous entry word to the utmost extent with the least number of equivalents, which could be achieved by providing equivalents that are applicable to a broader range of contexts in the target language and with greater descriptive power. Moreover, various devices such as co-text, context, illustrative examples, usage labels and shared characters—in the case of equivalents in Chinese—should be made good use of to indicate the associations between the equivalents and discriminate their meanings.

**Acknowledgments**

The author would like to express her deep gratitude to Professor Gloria Clavería Nadal of Universitat Autònoma de Barcelona for her scientific guidance that greatly assisted the research and for her valuable comments on this article.
References

Dictionaries


Other literature


Signposting Headwords for Increased User Friendliness in Bilingual Dictionaries

Mats-Peter Sundström
European Parliament, Swedish translation unit, Luxembourg
matspeter.sundstrom@europarl.europa.eu

Abstract

This paper proposes the use of computer technology to enhance user-friendliness of bilingual dictionaries. More specifically, graphic symbols maybe apposed next to headwords (“signposting”) to alert the user to possible problems relating to the equivalents, complete with links to additional information. These symbols may assume the shape of a warning triangle and refer to, inter alia, five possible complications. Under the additional information links supplementary information could then be provided to the user about such usage-related restriction and lists of circumlocutory target language constructions presented, when appropriate, always keeping in mind the non-exhaustive character of such lists. Signposting of the above kind may reasonably be presumed to play a key role notably in bilingual dictionaries between structurally and culturally very distant target languages, e.g. English and Asian languages and vice versa, above all when these dictionaries are used for the production of target language texts.

Keywords: Graphic symbols, text production aids

1. Introduction

A bilingual dictionary, or a BD for short, is sometimes considered secondary to a monolingual dictionary, mainly because of what is perceived as the simplistic nature of its definitions. After all, in a BD the definitions of the entrywords consist of the equivalents themselves. As is well known, however, there are cases where equivalents must be complemented by other devices, or even replaced by such devices where equivalents are lacking. Among these devices are found notably labels, or diasystematic markers, giving information, as pointed out by Svensén, either about the headwords themselves or about the equivalents (cf Svensén 2004: 380). Usually this information is related to aspects such as style, subject field or regional provenience. In the words of Otogetswe, “[a]lthough the general terminology [of dictionary labels] is not standardized, there is consensus of the necessity of dictionary labels” (Otogetswe 2019:299).

In addition to these labels come explanatory metalanguage and illustrative phrases, IP’s for short. Explanatory language fulfils a function notably to explain headwords with for cultural or social reasons have no equivalent in the target language. Thus, in a BD having Finnish for its
source language, headwords referring to items of Finnish material culture, or to institutions of government, society etc. in Finland might be expected to be followed by a succinct metalanguage comment in the target language (in this context using the target language by default imposes itself, as source language speakers could reasonably be expected to know the meaning of the headword concerned).

In an imaginary Finnish-English BD, an illustration to explanatory metalanguage could look like this: a Finnish headword kalakukko followed by an explanatory metalanguage in English: “a Finnish national dish consisting of fish baked into a crust of dough”.

With respect to IP’s, their function is rather similar to that of explanatory language. The main difference resides in the fact that the use explanatory phrases makes the BD suitable also for production purposes, and not only reception, as they serve to illustrate the way the contents covered by a headword lacking one-to-one equivalents in the target language may be expressed in that language.

2. Signposts as an additional means of information

This paper proposes yet another method to be included as part of the BD information strategies. For the purposes of this paper, the method has been called signposting. In practice, it implies using typographical devices referring to micro- or macrostructural features containing additional information about the entryword thus denoted. In terms of appearance, these signposts could come in almost any shape. One alternative consists of given them the form of a small triangle with an exclamation mark in the middle. Placed next to the entryword itself, it would alert the dictionary user at first glance to the fact that there is something special about the entryword or its equivalents meriting further attention. These signposts would increase user-friendliness, as in the words of Cowie, they would be "designed to save the user from having to read through complete definitions of every sense of a word (Cowie 2012: 288). Capitalising on the possibilities of modern technology these signposts could then direct the dictionary user to on-line links providing additional information about the entrywords. Alternatively, although in a more space-consuming manner, this additional information could be placed within a framed-in box close to the dictionary article, or even included as part of it, perhaps at the article’s end.

3. Why should signposts be needed at all?

In this context a question presents, or rather imposes itself. Why should signposts of this kind be needed at all? Surely labels, IP’s and explanatory metalanguage should be sufficiently strong vectors of information to tell BD user all he needs to know about the entrywords? – Experience shows that this is not necessarily the case. To begin with, labels and explanatory metalanguage both, by their very nature, have to be short and succinct and, as was pointed out by Overgaard Ptaszynski, “It is reasonable to assume that [dictionary] users who are not highly proficient in [the source and/or target language], and who lack experience in using dictionaries would be served better by relatively non-condensed data” (Overgaard Ptaszynski 2010:433). Also, IP’s do not give explicit answers but rather conclusions by way of inference.
As a further illustration, attention may be drawn to lexicographic anisomorphism on the level of individual words. In this context, most likely there are cases where neither IP’s or explanatory metalanguage, whether in isolation or in combination, can shed sufficient light on a given headword from the target language’s point of view.

Consideration may be given, either to the Finnish verb *jaksaa* or to the Swedish verb *orka*, which both, for all practical purposes, are basically similar in meaning. Apart from this similarity, they also have another feature in common. They refer to a universal aspect of human life, but nonetheless they have no one-to-one English equivalents. Admittedly, they both primarily mean *have the strength/the energy* to do something (either in a physical or mental sense), but their usage is far more extensive. Very often, particularly in spoken, colloquial language, they serve to express an element of annoyance or irritation, like in the Finnish *En jaksa kuulla tuota enää!* literally *I do not have the energy to hear any more of that!*

In an actual speech situation, however, such an English construction would appear unnatural. What the Finnish original means is simply *I’ve already heard my fill of that!* ... or even more simply *That’s enough!*

Owing to the character of the above-mentioned verbs it is unlikely that any amount of explanatory metalanguage in the BD article, or any number of IP:s could give sufficient information about the meanings and/or usages covered by a verb such as *jaksaa* or *orka*. At this stage the signposts could come in quite handy. Once having captured the users attention they could refer the user to an on-line link and/or a framed-in set of additional explanations as regards both the source language meaning (for reception purposes) and the possible ways of target language expression (for production purposes).

This being so, the user could largely avoid a strategy called contextual guessing by Yuzhen Chen, who went on the specify: "Useful as it is, contextual guessing might be a process prone to incomplete or wrong interference" (Chen 2012: 242).

Another example, yet again drawn from the Finnish language, is furnished by the noun *irtiotto*. At its simplest, it means an act of breaking away from something, and is used in connection with sports events to designate a runner’s making a burst, i.e. breaking away from the other runners and taking the lead. In actual usage, however, it has acquired an extended meaning, often in the sense of "getting away from it all”, i.e. breaking with boring daily routine. However, it may also be used to describe acts of (mostly) an adolescent’s rejection of parental values and/or authority, as in *Irtiotot ovat tarpeellinen prosessi, joka auttaa nuorta myöhemmin pärjäämään omillaan*. – In this context, an English translation just does not exists – or rather, a "once-for-all" translation does not exist. In fact, there are well-nigh infinite possibilities open. The main point, however, resides in the fact that the expression means something like *Breaking with parental values/discipline is a necessary process and helps the young people acquire independence later on.*

In this connection, a signpost could render most valuable service. Firstly, for productive purposes, a standardised phrase could be used to signal the non-availability of any equivalent proper. Secondly, a usage note could be added, to give information as to the range of context where the headword may be used. Thirdly and finally, a number of IP’s could be provided, offering illustrations to possible way of conveying the source language message in the target language. All this would hardly be possible to fit within the confines of the microstructure in a
BD, and thus, as a source of additional information, a signpost with a message attached might come in handy.

Another telling illustration to lexicographic anisomorphism where dictionary signposts could prove most helpful is found in the Finnish noun etsikoaika. Having the literal meaning of "times of searching", it originated as a biblical term, corresponding to the English "time of visitation", meaning, according to some theological interpretations, times when God calls us more powerfully than on other occasions. However, according to a Finnish monolingual online dictionary, the lexeme has spread into general language also, acquiring the sense of "a time when action is particularly called for or else, a good opportunity may be lost or dire consequences may result (Urbaaani Sanakirja 2017). As an English equivalent, window of opportunity may serve – on occasion. In a signposted comment, information could be added to the effect that the Finnish lexeme comes with more sinister connotations in case of non-action.

As a final illustration to this theme could be given the Finnish noun ruuhkavuodet. Literally it means "the rush years" or "the peak years" and could be used by way of extension to designate any years characterised by heavy workloads. In its original sense, however, it refers to the time in a family’s (or even more, a single provider’s) life when the children are small and where there is a constant struggle to balance the demands of working life and family life. Thus, its use is normally restricted to a given set of situations. Somewhat similar is the case with the English phrase the awkward years. While a language user not very familiar with the English language may easily construe them as referring to any number of years where someone finds himself in an awkward situation, in actual English usage they basically refer to the teen years, the time of adolescence.

In both the above cases, the dictionary entries could be supplied with signposts directing the reader to spots with further information (again, irrespective of whether provided in a boxed-in structure or by means of an online link). These information spots could further illustrate in what particular contexts these lexemes are used, along with explanations, if needed, as to why these years are either rush years or awkward years. Perhaps the Finnish ruuhkavuodet could also be furnished with an extralinguistic, but relevant comment to the effect that it was coined in the title of a debate book in 2003, Elä mä n ruuhkavuodet that is The busy years of life describing in great detail what being a working parent entails and subsequently becoming a stock expression in the Finnish language. Obviously, lexicographic anisomorphism is not confined to the language par Finnish and English. To give but one final example, reference may be made to the French jeunisme which in the European Union database IATE is given the English equivalent obsession with youth culture. The problematic aspect about the English equivalent is that it is a description rather than a lexeme. In this context, we might have needed a signpost referring the dictionary user to explanations, such as those found in the monolingual French Wikipedia, giving diaevaluative information to the effect that this French word is generally pejorative and occurs notably in connection with employment matters (referring to a preference for young employees in hiring policies) or questions relating to physical appearance (cosmetic surgery); alternatively also to a cult of youth in general (cf French Wikipedia).

In connection with this last example, the present author would like to distance himself somewhat from the opinion proposed by Overgaard-Ptaszynski to the effect that "[t]he need for diachronic and diastatic information does not arise in connection with text reception in L2 [ ...]. Consequently this information need not be signalled in dictionaries for text reception"
Surely such information may reasonably be presumed as relevant also in connection with text reception?

4. Particular occasions where signposting may be particularly helpful in BDs

In this paper attention will be given to three particular cases when signposts might be instrumental in giving the BD user. They are 1) focalising words, 2) stock phrases and 3) proverbs.

The term focalising words was coined by the present author to designate words which on the semantic level summarise "an often rather complex development, procedure, chain of events" (Sundström 2011: 72). For the purposes of this paper the following three may be chosen as illustrations: the French *litoralisation* and the English *gig economy* and *silo structure*. The French noun refers to the concentration of settlements and urban/economic activity in areas close to the seafront. As is easily seen, and as was also the case with the French *jeunisme*, it has no English equivalent and is well-nigh impossible to explain exhaustively by means of either explanatory metalanguage or IP’s.

Somewhat similar, from an non-native English speaker’s point of view, is the case with such English expressions as *gig economy* or *silo structure*. The former refers to an economy where "instead of a regular wage, workers get paid for the ‘gigs’ they do, such as a food delivery or a car journey", according to a definition presented by the BBC Business News,. The latter refers to a situation when, in an organisation, administration etc. work is done in isolation, without cooperation between the various departments or individuals involved in it. For these lexemes, I think it safe to say that most languages lack equivalents and an in-depth explanation given by a signposted reference might be most informative. This is true also of various set phrases. Consider the English *I stand corrected*. Used as it is to indicate the willingness to admit the appropriateness of a correction made or a reprimand received, it also indicates an element of submission to facts, without, however, humiliating oneself. Also it is stylistically cross-cutting and, from a diaevaluative point of view, neutral rather than connoted, as in itself, it implies no element of sarcasm. Altogether different is the case with the English phrase *Have it your own way!*. Informal in style and ironic in terms of connotations, it rather implies an element of resignation and disagreement, coupled with an almost fatigued desire not to pursue a dispute any further. All these aspects will be needed to enable a dictionary user to properly understand the phrases concerned and are probably best given under a signposted piece of information. A rather curious case, from a user pragmatic point of view, is found in the Finnish phrase *Ohi on!* While it certainly means *It is (all) over now*, its actual usage is mainly restricted to a joyful exclamation on the occasion of [some]one’s discharge after having completed compulsory military service.

Proverbs finally, often come with underlying meanings lost or misrepresented in translation, even though a very similar proverb seems to exist in the target language. An eminent illustration to this was given in 1998, in a parliamentary question posed by a Finnish Member of the European Parliament when he claimed that an answer furnished by the European Commission was nothing except *Hyvää päivää - kirvesvartta*, or, in word-for-word English translation, the obviously meaningless *Good day – axe handles!* Placed at the
entryword *kirvesvarsi* (‘axe handle’) a signpost could guide the dictionary user to an explanation informing him that this is an ironic exclamation used to indicate that an answer received was perfectly irrelevant to the question asked. This main piece of information could then be supplied by an explanation of the underlying folk tale, of a deaf man who was ashamed of admitting his disability and thus tried to figure out in advance what people would say when they saw him, so as to be able to engage in conversation with them. Sitting by the roadside and carving at an axe handle he presumed that the very first thing spoken to him would be a question as to what he was making and so, he would answer ”Axe handles!” Much to his bad luck, the first thing that was said was an actual fact a phrase of salutation, roughly equivalent to the English *Good morning!* and so, he returned the salute by shouting ”Axe handles!” Incidentally, the English official translation *if you want a straight answer, you can whistle for it* was not altogether appropriate, as it implies arrogance, and an unwillingness to answer, connotations absent from the Finnish proverbial phrase used. In this particularly case, an English translation such as *a red herring* might have been used to good effect.

5. **Connotational information provided through signposting**

Connotations are notoriously tricky and difficult to perceive for the foreign language learner. For starters, mention may be made of the English words *fairly* and *rather*. As explained by Stannard Allen, they are both used to denote a moderate degree but the former ”is used when the speaker or writer wishes to affirm some positive or pleasant idea; *rather* is used when the idea is negative or unpleasant”, followed by a statement to the effect that fairly is a step towards an ideal but *rather* is a step away from it (Stannard-Allen 1974:283). In its present form this information would be a graphic illustration to information apt for inclusion after a signpost.

Even more striking is the case with the Swedish indefinite pronoun *somliga*, meaning ‘some’. Although stylistically on a level with other indefinite pronouns of the Swedish language such as *några, en del and vissa*, which all essentially carry the same meaning, *somliga*, as was pointed out by the present author, is not connotationally neutral. Over time a development has taken place, resulting in *somliga* being used almost exclusively in contexts implying criticism or disapproval. Thus it would be adequate to use it in a Swedish translation of an English phrase such as *Some of the translations were perfectly worthless* while rather not in a construction like *Some of the translations were impeccable*. In an article discussing this particular word, the present author called for these restrictions to be explicitly presented in dictionaries involving Swedish as a target language (cf Sundström 2015: 43). As is well known, sometimes an opinion or emotion may be inherent in the very choice of lexematic expression. This obvious state of affairs gets somewhat obscured when in a source language we have only one lexeme whereas in the target language there are several; some of them neutral, some connoted. An illustration may be provided by the various options available for translating the English verb *repurpose* into German, i.e. *umfunktionieren* and *zweckentfremden*. If in a phrase

After the conquests the Great Mosque of Cordoba was *repurposed* as a cathedral, whereas the Hagia Sofia was *repurposed* as a mosque.
the translator into German uses the verb *umfunktionieren* a neutral statement of fact is made, while the use of *zweckentfremden* [literally: "purpose-alienate"] implies an element of disapproval or lamenting the loss of the original purpose. An obvious difference to signalise to unsuspecting users of an English-German BD for productive purposes! Connotations brought to mind may sometimes assume quite unexpected, and even facetious aspects, totally opposed as they might be. A few illustrations will follow: In English a *bear* is usually perceived as a grumpy, ill-tempered animal (and in an extended sense, drawn from the language of financial markets, a pessimist), whereas both the Swedish *björn* and the Finnish *karhu* come with associations of a very strong but essentially sympathetic denizen of the vast northern forests. In Swedish the owl, or *ugglan* is a symbol for wisdom, whereas in Finnish the same bird, called *pöllö*, symbolises … stupidity. A more modern example of diametrically opposite connotations is offered by the word pair *tar* and its Finnish equivalent *terva*. In Finnish *terva* carries associations of something wholesome and aromatic, pregnant as it is with associations of Finnish folk medicine. The English *tar* obviously comes with unpleasant connotations of some sticky, polluting substance (who ever, for instance, would market a hair shampoo to native speakers of English as a product containing genuine *tar*?). The underlying reason is perfectly extralinguistic: in English-speaking nations *tar* is primarily seen as a product resulting from the distillation of coal (or from smoking cigarettes!), whereas in Finland, tar used to be produced from wood.

Somewhat more complicated is the case with the word pair *evangelisk* vs. *evangelical*. In even today predominantly Lutheran nations like Finland, the major national church is designated as the Evangelical Lutheran Church of Finland, and this being so, the corresponding adjective, whether *evangelisk* in Swedish or *evankeelinen* in Finnish, comes with connotations of "staid", "established" or "part of society’s order of things". In modern English usage, however, *evangelical* has come to acquire connotations of "charismatic", "free-wheeling" or "ecstatic", reportedly causing problems of interpretation for the major missionary organisation of the aforementioned church, the Finnish Evangelical Lutheran Mission.

As the above examples show, signposting could serve the interests of BD users also for target language production purposes. Even more marked is this quality in examples of the kind presented in the following section.

6. Signposts to prevent linguistic faux pas with their attendant loss of face

"Certainly, one would like to avoid using a restricted word in an inappropriate context." This statement (Overgaard-Ptaszynski 2010: 437) may come as something close to a truism but its relevance appears in a new light once one realises the vast variation of acceptability versus unacceptability between different languages. To be sure dictionary labels, such as *vulg* for "vulgar" or *slang* give a partial solution to the problem. Nevertheless there are cases where more comprehensive explanations are called for. To give a somewhat trivial example, we may consider the English noun *toilet*. Although certainly not impolite, let alone vulgar, in English-speaking countries it is often considered, as it were, too straightforward, and thus, at least in social conversation better replaced with alternatives such as *lavatory, restroom, washroom* or, notably in American English, *bathroom*. For users of languages where a word similar to *toilet*
comes without social restrictions on its use (such as the Scandinavian languages), additional socio-pragmatic information of the corresponding English word would be helpful.

Even more needed is information designated to help someone wishing to produce utterances in the target language avoid potentially embarrassing and unintentional pitfalls. Consider the Finnish word elämänmuutos. Derived from the noun elämä, signifying "life" and muutos, signifying "change" it may refer to any event that produces a change in someone’s life, as in Muutto Australiakan eller täälleni suuri elämänmuutos which would literally translate into English as "Moving to Australia was a great change of life for my aunt". However, this English translation is manifestly inappropriate, since "change of life" in English is usually construed as a reference to menopause. Thus, a well-placed warning in the form of a dictionary signpost may prevent a source-language native speaker to produce an utterance where, for instance, someone about to emigrate to Australia is unsuspectingly wished a happy change of life.

Somewhat similar is the case as regards the use of the English adjective adult in various combinations, such as adult language or adult movies. Unsuspecting non-native English speakers could easily take the expressions as references to anything relating to the world or concerns of grown-up people, taking adult language to mean the language used by people of mature age, as opposed to child language or teenager slang, instead of a language characterised by profanity and vulgar jokes etc.

Where appropriate, in order to enable BD users non-native in the target language to avoid possible embarrassing mistakes, such signposts should feature quite prominently in the article microstructure. Consider the French verb avoir, i.e. "to have" or, alternatively the English noun rule (in the sense of 'prescription') with its French translation règle. Having a readily visible signpost apposed at each entryword could help avoid truly face-losing gaffes of a kind I even hesitate to mention here (for reasons of elementary discretion both the acting person and the location shall be omitted). A elderly gentleman working for a European Union institution was reprimanding a group of trainees for their alleged non-observance of applicable rules. In an effort to make his message truly go home, he added a comment to the effect that I myself have to have my rules and abide by them, followed immediately by a literal translation into French: Moi aussi, je dois avoir mes règles ... only to be interrupted by a storm of embarrassed giggles. Unawares and most unfortunately to him, the French expression seemingly conveying the inoffensive meaning of "having my rules" can actually mean only one thing. In clear then (and with a conventional nod of apology), when said of a female person, it refers to the fact of her having her monthly period.

Occasionally, signposts of the kind discussed may be used for the opposite purpose. By this is meant reassuring the dictionary user that it is perfectly appropriate to use in the target language a word which resembles a given source language word (either by looking like it or carrying the same basic meaning), but which in the source language has become saddled with undesirable connotations. Thus a Swedish-language user of a Swedish-English dictionary may need a reassurance to the effect that the English party is a perfectly appropriate word to describe a solemn social function, as in this Wikipedia definition: Formal receptions are parties that are designed to receive a large number of guests, often at prestigious venues (Wikipedia 2017). In Swedish, the word party, borrowed from English, primarily designates informal function, often characterized by "high living" or "life in the fast lane". Likewise, Finnish users of Finnish–English BD’s may need a confirmation to the effect that the English
noun egg comes with no negative overtones whatever. The reason it, the Finnish muna, although basically meaning egg has come to be infested with such vulgar side meanings (derived from the similarity of appearance of an egg with something not fit for polite conversation), that its use often has become impossible in everyday Finnish and is thus replaced by kanannuna. i.e. hen’s egg, although in Finnish cooking, eggs are by default taken to mean just hen’s eggs.

7. Signposts as guidance to native-like language production

In the previous parts of this presentation, attention has been paid mainly to the usefulness of BD signposts from the point of view of language reception (except in the section about signposts as an aid against linguistic faux pas). While it is certainly true, as was pointed out by the Aarhus University Centre for Lexicography that the user who has a reception-related problem should not be unduly disturbed by production-related information (cf Malmgren & Sködeberg 2013: 129) a signposted dictionary could certainly also serve production needs. Some examples will be given below:

1. Although both English and Swedish have a whole set of words to describe various styles of headgear (hat, cap and hood, to mention a few, corresponding, respectively to hatt, mössa and luva in Swedish), in English daily usage hat will often function as a generic term for all kinds of headwear, whereas in Swedish the lexical distinction is maintained.

2. An Einwohner in German definitely matches the English inhabitant, even from a purely etymological point of view [German ein + wohnen (to live) Latin in + habitare (to live)]. Even so, in English inhabitant is rather uncommon. Instead of there are some 550 000 inhabitants in Luxembourg we would be more likely to find something like the population of Luxembourg totals some 550 000 or similar.

3. The English verb prepare does mean “to carry out the preparatory work for something”, but is also very often used in the sense of put together or elaborate. In connection with European Union legislation, for instance, it is very often said that the European Commission “prepares” a draft directive or regulation, in the sense not only of gathering background material for a future legislative act, but also actually putting together writing the legislative proposal. This is a distinction worth noticing in an Swedish–English dictionary aimed at helping native speakers of Swedish produce texts in English.

4. While the Finnish adjective turhainen certainly corresponds to the English vainglorious a person approaching the English language with Finnish for his/her mother tongue would most probably appreciate that the English lexeme is not very usual. Thus, a phrase of colloquial Finnish such as Miksi ihmisen täytyy olla niin turhainen would be most unlikely to occur in spontaneously written English as Why does man have to be so vainglorious? (as a fifteen-year old Finnish schoolgirl once wrote in an English-language essay). Rather, something like Why do people have to be so obsessed with good looks would have sounded more appropriate.

5. To some extent, constructions could be included. Thus under the headword so in a German–English BD the phrase so ist es could be given, not only with the primary equivalent phrase of so it is, but also with examples of its wider use in constructions. A real-life example, worthy of
inclusion, would be the guidance for visitors at a wine restaurant in the German town of Trier. The German explanation opened with the words *So ist es bei uns in der Weinstube*. In English translation it was idiomatically reproduced as *Here is a guide to how things work at our Wine- Restaurant*. A word for word construction *So it is at our Wine-Restaurant* would have left the reader puzzled: surely there must be something missing, like *Smoking is illegal in German catering establishments and so it is [also] at our Wine-Restaurant."

8. Final words

Indeed, a translation may well, as highlighted by the eminent Chinese translator Wang Keping in his preface to his meritorious translation of Lao Zi’s “Dao De Jing” appear “as though a glass of fine wine has been mixed with water, reducing it to a less tasteful cocktail” (Wang 2008: 9). Almost in the same breadth Wang Keping went on to underline the value of translations to “make inter-cultural communication or dialogue possible to a certain extent” *(ibid.)* And by way of extension, this is obviously true of BDs also. As was pointed out by Gao: “bilingual dictionaries are always viewed as a powerful tool of linguistic and cultural exchange” (Gao 2013: 88) and to this end the opportunities offer by modern information technology should be pursued to the fullest, although even today perhaps, as Yamada pointed out in 2009 the majority of students may not be able to take full advantage of the potential it offer (Yamada 2009: 65).

Borrowing concepts from Kwary, we may say that signposted information of the kind proposed may help turn a BD from a satisfier of real needs to a satisfier of ultimate needs. Here real needs ”include the needs for information to answer [the user’s] lexicographical problems promptly and correctly, whereas ultimate needs designate needs occurring at the end of a learning process which includes dictionary consultation”, such as writing an good text in the target language. *(cf Kwary 2012: 32). This is something that in informal English could qualify for the description of a tall order. And needless to say, expectations should not be set inordinately high. After all: ”To ensure perfect text reception, a dictionary would have to contain data specifying all pragmatic features of every word included into it, in case some reader comes across an atypical use of a certain word in a particular text. *(Overgaard Ptaszynski 2010: 417), which is of course impossible.

To give but one final illustration to the merits of signposted information, recourse may be had to an statement by Lew.” There is no doubt that locating embedded elements in entries, especially long one, is among the hardest skills that user of paper dictionaries need, and it would be of great help if this skill were no linger essential” *(Lew 2013: 82). Most probably this vexing problem makes itself particularly felt in BD’s involving languages culturally and structurally very far from one another, like Western and Asian languages. If the device of signposting, as presented in this paper, even in an ever so small way, could serve to give the lie to the Kipling’s well-known about east is east and west is west and never the twain shall meet, at least in from a lexicographic point of view, then indeed the device discussed would have fulfilled a mission going far beyond a purely academic and theoretical interest.
References


Abstract

Communicative Language Teaching believes that language is best learnt through use in social context, that is, by communicating in it. The approach sets great store by development of communicative competence, which consists of grammatical competence, sociolinguistic competence, discourse competence and strategic competence (Canale & Swain, 1980), by creation of authentic language environments. In a typical scenario, the student is the center of language learning with greater autonomy, while the teacher becomes an organizer of resources instead of being the sole resource. The evolution of learners’ dictionaries over the past several decades in particular has been increasingly user-driven. So the connotations of Communicative Language Teaching should not be lost on the part of lexicographers.

The thesis aims at the identifying the measures adopted by lexicographers to better satisfy the changed needs, through a comparative study of LDOCE 4 and LDOCE 5. After comprehensive analysis of the differences between the two dictionaries in structure, entries and senses, with focus on entries under Letter A, it is found that new entries and senses have been added, “Collocations”, “Thesaurus” and “Register” boxes established, illustrations updated, while the original “Grammar” box and synonym and antonym information has been thoroughly collated, aside from sticking to corpus-based lexicography and the 2,000-word Longman Defining Vocabulary. In addition, supportive materials aimed at facilitating daily communication, including “Formality in spoken and written English”, “Abbreviations used in electronic messages” and “Emoticons used in electronic messages”, have been incorporated. The update can be significant for developing learners’ dictionaries. Specifically, it might be advisable to incorporate corpus-based new words and senses, and furnish sociolinguistic, stylistic and pragmatics information to help learners enhance their communicative competence by exposing them to language use in real-world scenarios.

Keywords: Communicative Language Teaching, connotations, comparative study, learners’ dictionary
1. Communicative Language Teaching

1.1 Introduction to Communicative Language Teaching

Communicative Language Teaching started from a theory of language as communication proposed by Hymes (1972: 269-293) who stated that the goal of language teaching was to develop as he referred to communicative competence. He defined the theory of communicative competence as the language knowledge a speaker needs to have in order to be communicatively competent in a speech community. That means that a person who acquires communicative competence acquires both knowledge and ability for language use.

In 1980, Canale and Swain published “Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing” in *Applied Linguistics*. In the thesis, the two French teachers from Canada systematically summarized the research findings of communicative approach to language teaching and concluded that communicative competence should be tripartite, including grammatical competence, which consists in the ability to use grammar rules and lexicon to produce and understand a message; sociolinguistic competence, which refers to the ability to use language depending on the social context involving time, place and social relationship; discourse competence, which means the ability to connect a number of ideas together appropriately in order to interpret and to exchange messages, and strategic competence, which lies in the ability for learners to use strategies in communication for dealing with lack of vocabulary and structures. The ultimate objective is to understand meanings and to be understood.

The competence structure theory was to become the major frame of reference for syllabuses subsequently developed for foreign language teaching.

1.2 The Fundamental Approach to Communicative Language Teaching

With the emergence of Communicative Language Teaching, the process of language learning has been viewed from a different perspective which involves using the language to communicate messages. That is, language learning is believed to occur through meaningful communication, with due emphasis on development of the three competences mentioned above under the guidance of grammatical rules. In other words, language in use should be taken into account, and grammar should also be indispensable, just as Richards (2006: 9) stated: “while grammatical competence was needed to produce grammatically correct sentences, attention shifted to the knowledge and skills needed to use grammar and other aspects of language appropriately for different communicative purposes, such as making requests, giving advice, making suggestions, describing wishes and needs, and so on”.

In Communicative Language Teaching, instruction becomes learner-centered rather than teacher-centered. Accordingly, the teacher assumes the dual roles of facilitator and monitor of the learning process, while the students assume greater autonomy, and become responsible for their learning in terms of both the content and the process. Authentic instruction materials, including excerpts from journals, videos,
radio broadcasts and etc. are adopted. However, another aspect of authentic language is grammatical rules, which have been found significant for improve the linguistic output of the students, and consequently the effect of their language learning.

2. Relevance of Communicative Language Teaching to Development of Learners’ Dictionaries

Historically, learners’ dictionaries have been developed on the basis of research findings of linguistics and language teaching theories. In the early stages, their compilers were almost invariably language educators or linguists. The New Method English Dictionary, the first monolingual dictionary to appear, was jointly compiled by Michael West and James Endicott, two specialists of vocabulary control theory. The pioneering learners’ dictionaries were mostly resultant from the compilers’ personal experience in language teaching. Oxford Advanced Learners’ Dictionary (OALD, 1942), a representative work of the second generation of learners’ dictionaries, was developed by A. S. Hornby the noted linguist and language educator. (Cowie 2000: 27-29)

The demand for and evolution of learners’ dictionaries over the past several decades in particular have been increasingly user-driven (Cowie 2000: 1). In other words, learners’ dictionaries are increasingly influenced by what users require or are thought to require. For instance, Longman Dictionary of Contemporary English (LDOCE, 1978) extensively drew on the latest findings in grammatical studies and semantics and adopted a series of innovative designs in sense interpretation, usage labeling and grammatical labeling, in response to the increasingly prominent user-driven trend in the demand for learners’ dictionaries.

3. Communicative Language Teaching and LDOCE

As mentioned above, Communicative Language Teaching places equal importance on authenticity of instruction materials and grammatical rules. In the following section, a comprehensive analysis will be made of the 4th and 5th edition of Longman Dictionary of Contemporary English (hereafter “LDOCE 4” and “LDOCE 5” respectively), in an attempt to study the measures adopted by lexicographers in drawing on the approach to Communicative Language Teaching.

3.1 Overall Comparison

<table>
<thead>
<tr>
<th></th>
<th>LDOCE 4</th>
<th>LDOCE 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Pages</td>
<td>2688</td>
<td>2949</td>
</tr>
<tr>
<td>Price</td>
<td>119 (RMB)</td>
<td>118 (RMB)</td>
</tr>
<tr>
<td></td>
<td>LDOCE 5</td>
<td>LDOCE 4</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td><strong>No. of Entries</strong></td>
<td>207,000</td>
<td>230,000 (with updated senses for certain entries)</td>
</tr>
<tr>
<td><strong>No. of Examples</strong></td>
<td>155,000</td>
<td>165,000</td>
</tr>
<tr>
<td><strong>No. of Collocations</strong></td>
<td>66,000, dispersed among senses</td>
<td>65,000, under specific collocation boxes</td>
</tr>
<tr>
<td><strong>Synonyms and Antonyms</strong></td>
<td>7,000</td>
<td>18,000</td>
</tr>
<tr>
<td><strong>Frequency label</strong></td>
<td>Y</td>
<td>Y, with frequency label for academic vocabulary added.</td>
</tr>
<tr>
<td><strong>Defining Vocabulary</strong></td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td><strong>Register and Pragmatic Information</strong></td>
<td>under specific senses</td>
<td>under specific senses, with over 400 additional boxes</td>
</tr>
<tr>
<td><strong>Illustration</strong></td>
<td>Y, including 16 pages of full-color illustration</td>
<td>Y, with extensive updates</td>
</tr>
<tr>
<td><strong>Appendix</strong></td>
<td>72 pages, including 16 pages of full-color illustration, 28 pages of Language Notes (on grammar, pragmatics, collocations), 28 pages of tables (on numbers, weights and measures, military ranks, word formation, irregular verbs, geographical names and Longman Defining Vocabulary list)</td>
<td>77 pages, including 35 pages of formality in spoken and written English, 19 pages of Longman Communication 3000 (3000 most frequent words in both spoken and written English), 9 pages of Longman Defining Vocabulary, 2 pages of numbers, 3 pages of weights and measures, 2 pages of military ranks, 2 pages of word formation, 3 pages of irregular verbs, 7 pages of geographic names, 3 pages of abbreviations and 2 pages of emoticons used in electronic messages</td>
</tr>
</tbody>
</table>

It can be seen from the table that LDOCE 5 sets greater store by satisfying learners’ needs for cultivation of communicative competence, specifically through comprehensive updating the and collating the original grammar box, establishing the collocation, pragmatics and register boxes, and furnishing the list of emoticons and abbreviations commonly used in electronic messages. LDOCE 5 has also taken the lead to prepare Longman Communication 3000, a list of 3000 most frequently used words in spoken and written English, based on its corpus of over 390 million words. In addition, it has included the abbreviations and emoticons commonly used in electronic messages to the benefit of the users.
3.2 Comparison of Entries under Alphabet A

Table 2 Comparison of Entries of Alphabet A

<table>
<thead>
<tr>
<th></th>
<th>LDOCE 4</th>
<th>LDOCE 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Pages</td>
<td>124</td>
<td>149</td>
</tr>
<tr>
<td>No. of Entries</td>
<td>2057</td>
<td>2090(with updated senses for certain entries)</td>
</tr>
<tr>
<td>Register Box</td>
<td>0</td>
<td>47</td>
</tr>
<tr>
<td>Collocation Box</td>
<td>12(without title, in bold font)</td>
<td>33</td>
</tr>
<tr>
<td>Synonym and Antonym Labels</td>
<td>34 and 12</td>
<td>354 and 102</td>
</tr>
<tr>
<td>Thesaurus box</td>
<td>7(under the title of “Word Focus”, list of entries without exemplification)</td>
<td>32</td>
</tr>
<tr>
<td>Frequency chart</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Grammar box</td>
<td>7(under the title of “Word Choice”)</td>
<td>12</td>
</tr>
<tr>
<td>Illustration</td>
<td>13</td>
<td>12</td>
</tr>
</tbody>
</table>

It can be seen from the above table that the number of entries increased under the Alphabet A is disproportional to that of pages increased. Upon closer scrutiny, it is found that the additional pages have been mostly devoted to language-use information boxes, including the register boxes, thesaurus boxes, grammar boxes and etc.

3.2.1 Additional Entries and Senses

In comparison to LDOCE 4, LDOCE 5 has an additional 33 entries in total, including A*, Abdication, ablative, Acai berry, LCAS, accelerant, ABTA, the Addams Family, Airmiles, the Alladin’s Cave, the Alamo, Alcatraz, Alice in Wonderland, the AMA, the American Dream, the America’s Cup, the American Way, the American League, America the Beautiful, ammonite, amniotic fluid, Angry Young Man, antebellum, the Artful Dodger, the ASA, the Ashes, the ASPCA, the Asperger’s syndrome, A to Z, Augean Stables, Australian Rules Football, Axis of Evil and Axminster. A total of 17, or over half of those new entries are about Western culture, history or literature, indicating greater emphasis placed on background knowledge to enhance learners’ communicative competence. Take the entry “Alamo” for example. Aside from the background information, it also includes explicit information for usage in daily communication.
Update in entries is also reflected in senses added. Take the entry “Adam” for example. In LDOCE4, the entry has only one sense, as in the idiom “not know someone from Adam.” In LDOCE5, the sense of “Adam” as a figure of the Bible has been added, making the entry complete and the meaning of the idiom self-explanatory.

3.2.2 Updated Pragmatics Information Boxes

1) Collocation Box

Collocation with function and content words is indispensable to correct and natural language output, in both written and oral forms. However, Nesselhauf (2005: 82) once observed to the effect that compared to native speakers, foreign learners of English tended to overuse certain common collocations while their overall use of collocation was limited. They were unconfident for lack of understanding for the restrictions of collocations on word use and were more prone to make mistakes. In other words, collocations are important for English learners but are a major obstacle. In LDOCE5, collocations have been thoroughly collated, categorized and highlighted through separate boxes, as shown in the entry “action”.

Al-a-mo, the /əˈlaʊmoʊ; ˈæləmoʊ/ a famous battle that took place in 1836, when the US and Mexico were fighting each other for the control of Texas. The Americans were eventually all killed, but their brave action encouraged others, and Texas later became part of the US. The phrase ‘Remember the Alamo!’ is used to encourage people to continue doing something very difficult. 阿拉莫战役（1836年美国和墨西哥为争夺得克萨斯控制权而进行的战役; 短语Remember the Alamo!用于鼓励人们在困境中坚持不懈）

Ad-am /ˈædəm; ˈædəm/ 1 in the Old Testament of the Bible, the first man created by God. God put him in the Garden of Eden, and created Eve, the first woman, from one of Adam’s ribs. He was the father of Cain, Abel, and Seth. 亚当 [《圣经·旧约全书》中的人物; 人类的始祖] 2 not know someone from Adam informal to not know someone at all [非正式]根本不认识某人
In contrast, LDOCE 4 seems to be much less systematic. The collocations are indiscriminately piled up, without consideration for the cohesion between collocations and exemplifications:
Degrees of formality matter. That is the reason for LDOCE5 to set great store by differentiating formalities for word groups. Aside from inscribing 35 pages of formality information in spoken and written English, it has also made conscious efforts in the text of the dictionary, as exemplified by the entries of “activate” and “adjust.”
3) Thesaurus Boxes

In English, there are innumerable cases of different words meaning essentially the same thing, but with subtle differences that tend to be lost on non-native speakers. In other words, synonyms have proven tricky to learners of English as a foreign language for their minute differences in meaning and usage. This is a considerably augmented aspect of LDOCE5, in which the differences are illustrated with pragmatics information aided by examples. Take the entry “admire” for example:
4) Grammar Boxes

As mentioned before, grammar is another important aspect of communicative language teaching with significance influence on the desired effects of learning. English grammar has been notoriously difficult for learners of English as foreign language and known to incur various mistakes. In many cases, even the native speakers are at a loss. In LDOCE5, grammar information is highlighted in separate boxes for quick reference, as exemplified in the grammar box under the entry “ago”:

**GRAMMAR 語法: ago, before, previously**

Use ago to say how much time has passed from the time something happened to the present time. Use before to say how much time passed from the time something happened to a time in the past. Use previously in the same way, but is more formal. Previously has the same meaning as before, but it is more formal.

*The meeting was a follow-up to one that had been held four days previously.*

Use the past tense, not the present perfect, with ago and before.

▲ Do not use a preposition (at, in, on etc) before a phrase with ago.  带ago的短语之前不用at, in, on等介词: They first met fifteen years ago (NOT said at/in fifteen years ago).

▲ Do not use “since” or “before” with ago or before.

▲ Do not use “since” or “before” with ago. ago不和since或before连用: I came to the US two months ago (NOT said since/before two months ago).

In contrast, grammar information is much fewer in LDOCE 4, and mostly found in “Word Choice” boxes:
5) Antonym and Synonym Labels

In LDOCE 5, the synonym and antonym labels have been comprehensively updated, specifically by adding relevant SYN and OPP labels where necessary and replacing the original “=” and “≠” in LDOCE 4 with SYN and OPP respectively.

For example, in LDOCE 5, the labels are as follows:

\[\text{a-b-o-m-i-n-a-t-e} /'abəmɪneɪt; ə'bamə/\ v \text{[T not in progressive 不用进行式] formal to hate something very much 正式] 嫉恨, 厌恶} \text{SYN abhor}\]

\[\text{a-b-s-e-n-c-e} /'æbsəns; ə'absəs/ \ n \text{[1] (CU) when you are not in the place where people expect you to be, or the time that you are away 不在, 缺席 in/ during sb's absence Ms Leighton will be in charge during my absence (=while I am away). 我不在时, 由来顿女士负责。} \text{[2] (singular 单数] the lack of something or the fact that it does not exist 缺乏; 没有 OPP presence [+of] a complete absence of any kind of planning 毫无计划 In the absence of any evidence, the police had to let Myers go. 警察没有证据, 只好把迈尔斯放了。} \text{SYN indigenous}\]

538
While in LDOCE 4, they are like this:

![Dictionary page example](image)

### 4. Conclusion

In 2008, Chen Guohua and Tian Bing pointed out in “Design Features of the Next Generation of English Learner’s Dictionaries” that the next generation of English learners’ dictionaries should feature well-controlled defining vocabulary, systematic and comprehensive information on collocation with function and content words, after systematically reviewing the evolution history of English learners’ dictionaries. From the above analysis it can be seen that LDOCE 5 has drawn on the approach of Communicative Language Teaching for innovation in pragmatics information labeling, while adhering to its defining vocabulary of 2000 sifted from its corpus. It is consistent with the design features of the next generation of learners’ dictionaries.

The features of LDOCE 5 are undoubtedly of reference value to development of other learners’ dictionaries. Specifically, it might be advisable to incorporate corpus-based new words and senses, furnish sociolinguistic, stylistic and pragmatics information, such as collocations, thesauruses, grammar and register, in a conveniently accessible manner, for quick reference by learners.

### References


Treatment of Academic N-grams in English Pedagogical Dictionaries: A Corpus-based Study

Peng Chen
PLA University of Foreign Languages
ccp2111abc@163.com

Abstract

One of the key issues concerning academic writing is the use of phraseological sequences (in particular n-grams). Despite an increasing focus on n-grams in dictionaries, compared with single words, n-grams only have a ‘secondary’ status in the dictionary entry. Motivated by the research gap, i.e. the extent to which academic n-grams are represented in English dictionaries is underexplored, this study set out to identify 4-grams largely underused by Chinese academic writers and examine their treatment in some latest English pedagogical electronic dictionaries (OALD9, LDOCE5, COBUILD6, CALD4, MED2, OLDAE, and OCD2). It was found that the coverage and presentation (including the lexicographic form, location, and visibility) of underused 4-grams vary from one dictionary to another. For example, LDOCE5 shows the strongest preference for extra examples that contain target 4-grams. While there is no specific area for extra examples in COBUILD6, MED2 and OCD2 due to their simplistic software design. In terms of the visibility of 4-grams, COBUILD6 and MED2 have the highest ratio of explicitness whereas LDOCE5 the lowest. Moreover, a comparison of concordances of a sampled 4-gram ‘with the exception of’ suggested that L1-English and L1-Chinese writers differ in both structural and functional respects. Some suggestions were thus put forward to help improve the treatment of n-grams in future EAP dictionaries. Limitations and directions for further research are also discussed.

Keywords: n-grams, pedagogical dictionaries, treatment, EAP

1. Introduction

In recent years, a growing amount of literature has been published on examining phraseological sequences (in particular n-grams or lexical bundles) in dictionaries. Overall, these studies have investigated the presentation of n-grams in bilingual dictionaries (e.g. Granger and Lefer 2012, 2016) as well as the microstructural features of dictionaries through n-gram methods (e.g. Kamiński 2016; Nam et al. 2016). Lexicographers’ interest in n-grams can also be seen in Granger and Paquot (2015), which is probably one of the first to explicitly set up a ‘Lexical bundles’ section in
dictionaries. So far, however, there is a general paucity of dictionaries featuring the use of n-grams in English for academic purposes (EAP) settings, and the extent to which academic n-grams are represented in English pedagogical dictionaries is still underexplored.

Drawing on self-built comparable corpora, this study therefore set out to filter out a list of 4-grams largely underused by L1-Chinese academic writers and examine their treatment in some latest English pedagogical dictionaries, with a view to improving the lexicographic coverage and presentation of academic n-grams and contributing to the construction of an EAP n-grams dictionary in the future. This paper is structured as follows: Section 2 describes the corpus data, the procedures for identifying underused 4-grams on the part of Chinese academic writers, and the dictionaries chosen for the study. Section 3 presents a statistical and qualitative account of the coverage and presentation of 4-grams in various dictionaries and suggests ways in which the lexicographic treatment of these units can be improved. Section 4 summarizes the main findings of the study and points out some limitations and directions for future research.

2. Methodology

2.1 Corpus data

The corpora for this study consisted of Applied Linguistics (AL) English-medium research articles (RAs) published between 2010 and 2016 by L1-English academic writers (AL-EN) and by L1-Chinese writers (AL-CH). The AL-EN corpus RAs were selected from seven accredited SSCI-indexed AL journals, and the AL-CH corpus RAs from seven non-SSCI-indexed AL journals (see Table 1). Measures were taken to ensure that all RAs in each corpus are written by native speakers from the U.S./Britain and mainland China respectively (cf. Pan et al. 2016: 63). The RA text file was first converted from the original PDF and further ‘processed’ to keep only the body part—irrelevant elements like the header, abstract, tables and figures, reference, etc. were removed.

Table 1 Corpus data by journal (2010-2016)

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Journal</th>
<th>Number of RAs</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL-EN</td>
<td>Applied Linguistics</td>
<td>20</td>
<td>135,784</td>
</tr>
</tbody>
</table>

1 RAs in both corpora only include original research articles and review articles; Book reviews, commentaries, reports, etc. were excluded.
2 Considerations for choosing non-SSCI rather than SSCI-indexed journals are that: 1) RAs in non-SSCI-indexed journals may reveal, to a greater extent, n-grams differences between L1-English and L1-Chinese writers, and 2) these non-SSCI journals also follow the academic convention of peer-review and mainly target Chinese and other Asian scholars.
3 The token numbers were obtained by using the ‘Word List’ function of AntConc (version 3.4.4; Anthony 2014).
2.2 Procedures

2.2.1 N-grams identification

After each text file of the corpora had been meticulously proofread, n-grams (3 ≤ n ≤ 7) were automatically extracted by using KfNgram (version 1.3.14; Fletcher 2012). Based on a close examination of all the n-grams lists, we decided to focus on 4-grams as they often subsume 3-grams (cf. Cortes 2004) and are indeed much more frequent than n-grams of greater lengths. Following the practice of previous research (Hyland 2008a, 2008b; Pérez-Llantada 2014), the study selected a conservative cut-off point of 20 times per million words and a dispersion threshold of at least 10% of the texts. In addition, 4-grams containing the hash symbol # (representing digit 0 to 9) were excluded to guarantee lexico-grammaticality of the results (Pérez-Llantada 2014: 86). This yielded a total of 152 4-grams in AL-EN and 142 in AL-CH. Specifically, there were 73 overlapping 4-grams, and 79 and 69 idiosyncratic 4-grams in AL-EN and AL-CH respectively (Table 2).

<table>
<thead>
<tr>
<th>Overlapping 4-grams</th>
<th>Idiosyncratic 4-grams (AL-EN)</th>
<th>Idiosyncratic 4-grams (AL-CH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>in the present study</td>
<td>in the process of significant difference between the\n</td>
<td>on the other hand</td>
</tr>
<tr>
<td>in terms of the</td>
<td>that is to say</td>
<td>it is possible that</td>
</tr>
<tr>
<td>at the same time</td>
<td>no significant difference between</td>
<td>at the time of</td>
</tr>
</tbody>
</table>

Table 2 Top 10 common-core and idiosyncratic 4-grams
at the end of it was found that in the United States
the results of the is one of the to be able to
on the basis of the participants in the in the same way
the end of the the students in the at the level of
as shown in table students were asked to the case of the
of the present study be attributed to the in the absence of

2.2.2 Identifying underused 4-grams

To examine statistical differences of 4-grams between L1-English and L1-Chinese writers, Rayson’s online calculator (http://ucrel.lancs.ac.uk/llwizard.html) was used to produce log-likelihood scores. Pérez-Llantada (2014) only tested log-likelihood values of overlapping bundles (L1 English-L2 English, L2 English-L1 Spanish), leaving aside those of idiosyncratic ones. However, our initial analysis showed that as regards a plenty of idiosyncratic 4-grams there was little statistical difference between writer groups. Thus in order to better identify, if not exhaustively, underused 4-grams by Chinese writers, log-likelihood scores of both overlapping and idiosyncratic 4-grams were calculated, resulting in a list of 8 overlapping and 52 idiosyncratic underused 4-grams (see Table 3 and 4).

Table 3 Underused overlapping 4-grams by sig. (L1-Chinese relative to L1-English)

<table>
<thead>
<tr>
<th>p &lt; .05 (n = 2)</th>
<th>p &lt; .01 (n = 1)</th>
<th>p &lt; .001 (n = 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>the fact that the can be found in in the case of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>as a result of the extent to which in the context of</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>p &lt; .05 (n = 11)</th>
<th>p &lt; .01 (n = 11)</th>
<th>p &lt; .001 (n = 30)</th>
</tr>
</thead>
<tbody>
<tr>
<td>the ways in which the time of the in the current study</td>
<td></td>
<td></td>
</tr>
<tr>
<td>extent to which the whether or not the for each of the</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the degree to which each of the it is possible that</td>
<td></td>
<td></td>
</tr>
<tr>
<td>in each of the by the fact that at the time of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>as well as a in an attempt to in the United States</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4 Sample of underused idiosyncratic 4-grams by sig. (L1-Chinese relative to L1-English)

The list of underused 4-grams was further scrutinized on the basis of the following criteria: semantic transparency, structural integrity, topic-nonspecificity and

---

4 Topic-nonspecificity requires that the selected 4-grams should not be specific to a certain research topic (e.g. ‘in the United States’ and ‘the common European framework’ are considered topic-specific and may be of little lexicographic interest).
lexicographic potential. All 4-grams unable to meet any one of the above criteria were weeded out. The final list thus consisted of 25 4-grams: ‘in the current study’, ‘it is possible that’, ‘at the time of’, ‘in the same way’, ‘in the absence of’, ‘for the purposes of’, ‘as a means of’, ‘in a variety of’, ‘over the course of’, ‘a wide variety of’, ‘studies have shown that’, ‘it is not clear’, ‘by the fact that’, ‘in an attempt to’, ‘it is likely that’, ‘in a way that’, ‘the degree to which’, ‘it is possible to’, ‘with the exception of’, ‘in the area of’, ‘in the case of’, ‘the extent to which’, ‘in the context of’, ‘it is important to’, and ‘as a result of’.

2.2.3 Analyzing underused 4-grams in English pedagogical dictionaries

Following Tarp’s (2011: 228) taxonomy of pedagogical dictionaries, this study used the following e-dictionaries (CD-ROM/DVD-ROM): the ‘Big Five’—OALD9 (2015), LDOCE5 (2009), COBUILD6 (2008), CALD4 (2013), MED2 (2007), an EAP learner’s dictionary—OLDAE6 (2014), and a collocations dictionary—OCD2 (2009). Choosing electronic rather than paper dictionaries is motivated by our practical concern: compared with paper dictionaries, e-dictionaries allow a much quicker search of word sequences and usually provide extra information (e.g. collocations, examples) in a collapsible panel (see Figure 1) or pop-up window.

The next step was to prepare lexicographic information of underused 4-grams in Excel. First, content words in each 4-gram are typed into the search box of each e-dictionary. Then, entry information (including that in a collapsible panel and pop-up window) of the content word (e.g. ‘possible’) was pasted into a Word file to facilitate a thorough search for the target 4-gram (e.g. ‘it is possible that’). Finally, the lexicographic data of 4-grams—including the entry text, frequency of occurrence by form and by location, etc.—were put down in detail in an Excel file.

5 Note that LDOCE and COBUILD have ceased to update their electronic versions, their updated paper versions being LDOCE6 (2014) and COBUILD8 (2014). All the e-dictionaries chosen were state-of-the-art ones by the time of writing.

6 We prefer to classify OLDAE in a category different from traditional learner’s dictionaries, as it is believed to be ‘the first [EAP] dictionary of its kind’ (Lea 2014: vi).
3. Results and discussion

3.1 Coverage

In this study, the *coverage ratio by dictionary* is defined as the number of ‘searchable’ 4-grams (i.e. the number of 4-grams that can be looked up in a dictionary) divided by the total number of 4-grams. As shown in Figure 2, dictionaries vary in their coverage of underused 4-grams. LDOCE5 and OLDAE cover the largest number of searchable 4-grams (88%), followed by OALD9 (68%), MED2 (48%), CALD4 (44%), and COBUILD6 (40%). OCD2, perhaps due to a focus more on collocations (2-grams) than on larger n-grams, has the lowest ratio (32%).
On the other hand, by calculating the **coverage ratio by 4-gram** (i.e. the number of correct look-ups of a given 4-gram divided by the total number of dictionaries), variation was found between different 4-grams (see Figure 3). The number of correct look-ups ranges from 7 (‘in the absence of’, ‘with the exception of’, ‘in the context of’) to 1 (‘in the current study’, ‘to the degree to which’, ‘in the area of’), showing extreme individual differences. Though at this stage it is still difficult to perceive any relationship between 4-grams and the number of look-ups, the maximum and minimum figures, to some extent, indicate variation in ‘popularity’ regarding these 4-grams. In other words, the degree to which different 4-grams are included in the entry of one of their component words are not the same.

![Figure 3 Coverage ratio by 4-gram](image)

### 3.2 Presentation

This part reports on some descriptive statistics of the treatment of underused 4-grams, focusing on 3 types of micro-structural features: the lexicographic form, location, and visibility.

#### 3.2.1 Lexicographic form

As Table 5 shows, for each of the chosen dictionaries, none of the underused 4-grams has been given the headword status, which supports Sinclair’s (2008: 408) claim that ‘[d]ictionaries are almost exclusively focused on the word’. Overall, 4-grams take the lexicographic forms of **decontextualized structures** (i.e. 4-grams per se) and **contextualized examples**. However, the high standard deviation (SD) of structures and examples (including other examples) clearly indicates variation in the dictionaries’ practical treatment of 4-grams. The exemplification rate (examples per entry) is generally above 1 (mean = 1.77), i.e. for every searchable entry there is at least, on
average, one example containing the target 4-gram. By contrast, the mean of structures per entry is much lower (mean = 0.54).

**Table 5** Descriptive statistics of lexicographic forms

<table>
<thead>
<tr>
<th>Dictionaries</th>
<th>Def.</th>
<th>St.</th>
<th>Ex./extra ex.</th>
<th>Other ex.</th>
<th>St./entry</th>
<th>Ex./entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>OALD9</td>
<td>0</td>
<td>2</td>
<td>31/17</td>
<td>4</td>
<td>0.1</td>
<td>1.8</td>
</tr>
<tr>
<td>LDOCE5</td>
<td>2</td>
<td>17</td>
<td>97/76</td>
<td>56</td>
<td>0.8</td>
<td>4.4</td>
</tr>
<tr>
<td>COBUILD6</td>
<td>4</td>
<td>0</td>
<td>11/0</td>
<td>4</td>
<td>0.0</td>
<td>1.1</td>
</tr>
<tr>
<td>CALD4</td>
<td>0</td>
<td>7</td>
<td>22/14</td>
<td>2</td>
<td>0.6</td>
<td>2.0</td>
</tr>
<tr>
<td>MED2</td>
<td>0</td>
<td>12</td>
<td>11/0</td>
<td>2</td>
<td>1.0</td>
<td>0.9</td>
</tr>
<tr>
<td>OLDAE</td>
<td>0</td>
<td>16</td>
<td>29/6</td>
<td>7</td>
<td>0.7</td>
<td>1.3</td>
</tr>
<tr>
<td>OCD2</td>
<td>0</td>
<td>5</td>
<td>7/0</td>
<td>7</td>
<td>0.6</td>
<td>0.9</td>
</tr>
<tr>
<td><strong>Mean (SD)</strong></td>
<td>0.85(1.57)</td>
<td>8.43(6.70)</td>
<td>29.71(31.10)</td>
<td>11.71(19.64)</td>
<td>0.54(0.36)</td>
<td>1.77(1.24)</td>
</tr>
</tbody>
</table>

Note: Def. = number of definition, St. = number of structures, Ex. = number of examples. Extra examples are those hidden in a pop-up window or collapsible panel. Other examples refer to those containing 4-gram variants (i.e. contraction: ‘it’s possible that’; or with inserted elements: ‘it is just possible that’).

From the data in Figure 4, it is apparent that the lexicographic form a 4-gram takes varies from one dictionary to another. LDOCE5, among others, shows a much stronger preference for extra examples than other dictionaries. All extra examples are placed in relevant boxes to the right of the entry window (e.g. ‘Collocations’, ‘Phrase Bank’, ‘Example Bank’; see Figure 5), taking up a significant majority of the total of lexicographic forms (76/116). Other dictionaries that offer extra examples include OALD9 (17/33), CALD4 (14/29) and OLDAE (6/45). In contrast, due to a much simpler software design, there is no specific area for extra examples in COBUILD6, MED2 and OCD2. In addition, COBUILD6 is the only dictionary devoid of 4-gram structures.
Figure 4 Stacked bar of lexicographic forms

Note: Non-extra examples refer to those that are explicitly displayed in an entry rather than hidden in a pop-up window or collapsible panel.

Figure 5 The main entry screen for the noun ‘variety’ (LDOCE5)
3.2.2 Location

As can be seen in Figure 6, dictionaries have a clear preference (mean = 59.7%) for treating 4-grams as part of a subentry or as ‘secondary headwords’ (Atkins and Rundell 2008: 253). In general, there are no more than 3 types of locations for 4-grams in a single dictionary. COBUILD6 is featured by a comparatively ‘monotonous’ style of placing all 4-grams in subentries. Due in part to their software design, MED2 and OCD2 similarly put 4-grams either in a subentry or in separate blocks at the end of an entry. LDOCE5 and CALD4, on the other hand, make good use of the pop-up window design and present the most varied types of locations.

![Figure 6 Percentage stacked bar of location](image)

3.2.3 Visibility

In this study visibility is mainly concerned with the way a 4-gram is visible—bold or plain, explicit or implicit—to a dictionary user. It can be seen from Table 6 that dictionaries vary in the ratio of lexicographic forms in bold type (i.e. bold%). MED2 has the highest ratio (12 of 23) while OCD2 the lowest (2 of 12). It has been suggested that ‘multiword expressions are usually in bold type’ (Atkins and Rundell 2008: 34), which, however, does not appear to be the case. Overall, the degree to which 4-grams are highlighted in bold is probably insufficient (mean = 31.9%). By contrast, the ratio of explicit lexicographic forms is much higher (mean = 69.9%). It can also be observed that dictionaries differ in their practical treatment of making explicit 4-grams.

Interestingly, by virtue of their simplistic software design, COBUILD6 and MED2 have the highest ratio of explicitness, while LDOCE5 the lowest. An important reason is that all extra examples have been placed in the Example Bank (cf. Figure 5). Previous research has indicated that the visibility of multiword units—if embedded in example sentences—may be undermined (e.g. Bentivogli and Pianta 2002). Therefore,
it is hoped that e-dictionary designers will pay more attention to the issue of visibility to help improve users’ look-up effectiveness as well as explicit/implicit learning of n-grams.

Table 6 Descriptive statistics of visibility

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Bold def.</th>
<th>Bold st.</th>
<th>Bold ex.</th>
<th>Plain</th>
<th>Bold%</th>
<th>Explicit</th>
<th>Implicit</th>
<th>Explicit%</th>
</tr>
</thead>
<tbody>
<tr>
<td>OALD9</td>
<td>0</td>
<td>2</td>
<td>13</td>
<td>18</td>
<td>45.5%</td>
<td>17</td>
<td>16</td>
<td>51.5%</td>
</tr>
<tr>
<td>LDOCE5</td>
<td>0</td>
<td>17</td>
<td>5</td>
<td>94</td>
<td>19.0%</td>
<td>38</td>
<td>78</td>
<td>32.8%</td>
</tr>
<tr>
<td>COBUILD6</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>26.7%</td>
<td>15</td>
<td>0</td>
<td>100.0%</td>
</tr>
<tr>
<td>CALD4</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>21</td>
<td>27.6%</td>
<td>12</td>
<td>17</td>
<td>41.4%</td>
</tr>
<tr>
<td>MED2</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>11</td>
<td>52.2%</td>
<td>23</td>
<td>0</td>
<td>100.0%</td>
</tr>
<tr>
<td>OLDAE</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>29</td>
<td>35.6%</td>
<td>36</td>
<td>9</td>
<td>80.0%</td>
</tr>
<tr>
<td>OCD2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>10</td>
<td>16.7%</td>
<td>10</td>
<td>2</td>
<td>83.3%</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>31.9%</td>
<td></td>
<td></td>
<td>69.9%</td>
</tr>
</tbody>
</table>

Note: By being explicit we mean that a lexicographic form is directly shown in the entry not in a pop-up window or collapsible panel. Our distinction between explicit and implicit forms rests on the premise that the latter requires additional look-up effort and time for a dictionary user.

3.3 Discussion

3.3.1 Findings

A number of findings merged from the above analysis. First, the coverage of underused 4-grams varies both by dictionary and by 4-gram per se. Second, the manner in which these 4-grams are presented—including the lexicographic form, location, and visibility—differ markedly from one dictionary to another. Take for example the 4-gram ‘with the exception of’ (see Table 7). A comparison of entries suggests a lack of consistency in lexicographical representation. Under the entry of ‘exception’, ‘with the exception of’ is treated in a variety of forms and listed in different phraseology sections like ‘Collocations’ (LDOCE5), ‘Idioms’ (OALD9, OLDAE, OCD2)7, and ‘Phrases’ (COBUILD6, MED2). Third, based on the descriptive statistics presented in the previous subsection, it can be seen that both the learner’s (the ‘Big Five’) and the collocations dictionary (OCD2), like the EAP dictionary (OLDAE), have indeed paid attention to the presentation of 4-grams. However, the illustrating examples they include are non-academic and in this aspect may fail to serve as an aid in academic writers’ production process.

7 It is interesting to note that OALD9, OLDAE and OCD2, published by the same publisher (Oxford University Press), adopt an identical editorial policy regarding the location of ‘with the exception of’.
<table>
<thead>
<tr>
<th>Source</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>OALD9</td>
<td><strong>exception</strong> noun</td>
<td>a person or thing that is not included in a general statement. <strong>LANGUAGE BANK</strong> [collapsible panel] <em>With the exception of</em> the title track, this album is a huge disappointment.</td>
</tr>
<tr>
<td>LDOCE5</td>
<td><strong>exception</strong> noun</td>
<td>something or someone that is not included in a general statement or does not follow a rule or pattern. <strong>COLLOCATIONS</strong> <em>with the exception of</em> We all laughed, with the exception of Maggie.</td>
</tr>
<tr>
<td>COBUILD6</td>
<td><strong>exception</strong></td>
<td>1. N-COUNT [oft with <em>the N of</em> n] An <strong>exception</strong> is a particular thing, person, or situation that is not included in a general statement, judgement, or rule. [+ of] <em>There were no floral offerings at the ceremony, with the exception of a single red rose.</em> 5. PHRASE You use <strong>with the exception of</strong> to introduce a thing or person that is not included in a general statement that you are making. <em>Yesterday was a day off for everybody, with the exception of Lawrence.</em></td>
</tr>
<tr>
<td>CALD4</td>
<td><strong>exception</strong> noun</td>
<td>B2 someone or something that is not included in a rule, group, or list or that does not behave in the expected way: <em>I like all kinds of films with the exception of (= but not) horror films.</em> <strong>Extra Examples</strong> [pop-up window] <em>With the exception of Jonathan Edwards, Britain had no likely medal winners in track and field.</em></td>
</tr>
<tr>
<td>MED2</td>
<td><strong>exception</strong> noun</td>
<td><strong>PHRASE(S)</strong> <em>with the exception of</em> used for saying that someone or something is not included in what you are saying</td>
</tr>
<tr>
<td>OLDAE</td>
<td><strong>exception</strong> noun</td>
<td>a person or thing that is not included in a general statement, or that does not follow a rule. <strong>NOTE BOX</strong> [collapsible panel expanded] <em>with the exception of...</em> <em>Tidal energy is predictable, offering an advantage over other forms of renewable</em></td>
</tr>
</tbody>
</table>
energy, with the exception of geothermal energy.

IDIOMS [collapsible panel expanded]

**with the exception of**
except; not including

*With the exception of the relatively small areas of upland rice, rice fields are flooded for part of the year.*

OCD2 exception noun

PREPOSITION

with the ~ of

*The whole of the island was flooded with the ~ of a small area in the north.*

IDIOMS [pop-up window]

**with the exception of**
except; not including

3.3.2 Suggestions for EAP dictionary making

An important way to better understand the production needs of Chinese writers is to scrutinize their academic writing by way of corpus queries. Again, take for instance ‘with the exception of’. A comparison of concordances shows considerable differences between L1-English and L1-Chinese writers. Structurally, the great majority of L1-English writers prefer to put this 4-gram at sentence- or clause-initial positions (22 of 26), whereas L1-Chinese writers are more likely to put it as embedded elements not preceded by a period or comma (8 of 12). Functionally, ‘with the exception of’ is clearly a ‘transition signal’ (Hyland 2008a: 49) denoting a contrastive tone (unusual or unexpected). Thus it would be redundant to put before it another contrastive adverb or conjunction like ‘however’ or ‘but’ (see Figure 7). These being considered, it is suggested that, concerning EAP dictionary making, efforts can be made in the following respects. First, lexicographic forms of n-grams—be they definitions, structures, or examples—might better be presented in bold and in explicit, easy-to-find locations. Second, academic n-grams can be incorporated in dictionaries as entry labels or as a long list in the reference section at the back matter. Third, usage notes explaining the appropriate usage (structure(s) and function(s)) of an n-gram can also be subsumed in the headword entry. Given that extensive research needs to be undertaken to explore the viability of treating n-grams as genuine headwords, the above suggestions are still conservative. Admittedly, there is a long way to go before the ‘ultimate [phraseology-driven] dictionary’ (Sinclair et al. 2004: xxiv) can be made.

the pre- & post-questionnaire results, with the exception of two items: Try to picture or visuali
t but partly support PDH and NPAHH with the exception of S-S type; the comprehension test s
tement. Subject is obligatory in English with the exception of imperatives, and it plays syntactic r
ce, applause, cheer and popularity, but with the exception of convenience they cannot be consid

---

8 Note that a number of EAP phraseology lists have already been built (e.g. Liu 2012; Martinez and Schmitt 2012; Wood and Appel 2014), though not for lexicographic purposes.
ter-interviewer variability. However, with the exception of Lorenzo-Dus and Meara’s (2005) s

Figure 7 Sample concordances of ‘with the exception of’ in AL-CH

4. Conclusion

This study has examined the lexicographic treatment of 4-grams underused by Chinese academic writers, integrating corpus techniques and dictionary entry data. It was found that dictionaries cover and present 4-grams in distinct ways. Concordances of a sampled 4-gram suggested that L1-English and L1-Chinese writers differ in its structural and functional aspects. Based on an inspection of corpus queries, some tentative suggestions were raised to improve the treatment of n-grams in future EAP dictionaries, with a view to meeting academic writers’ production needs.

There are two main limitations of this study. One limitation concerns the relatively limited size and (sub)disciplinary scope of the corpora. Adding RAs from a wide range of disciplines to increase the corpus size might be a solution. The other limitation concerns the narrow focus on 4-grams rather than on n-grams of varied lengths. Future studies in this area might consider examining: 1) both contiguous (n-grams) and non-contiguous sequences in academic discourse (e.g. p-frames, conegrams; cf. Römer 2010; Cunningham 2017); 2) the n-grams look-up behaviour of different groups of dictionary users (e.g. novice vs. professional academic writers); 3) dictionary users’ processing of n-grams through eye-tracking methodology (cf. Choi 2017).

Acknowledgments

This work was supported by the National Social Science Fund of China (16ZDA213).

References

Dictionaries


Other literature

Nam, K., H.-j. Song & J. Choi. 2016. A morpheme-based analysis of lexical bundles in


Which Words Should be Labelled Academic in an Indonesian Advanced Learner’s Dictionary?

Deny A. Kwary
Universitas Airlangga
kwary@fib.unair.ac.id

Almira F. Artha
Universitas Airlangga
airafidel@gmail.com

Abstract

Different researchers may have different opinion on what constitutes academic words. In the English language, Coxhead (2000) initiated the study with the formulation of Academic Word List (AWL). The AWL is used in the Oxford Advanced Learner’s Dictionary to label academic words. Paquot (2010) suggests a list called the Academic Keyword List (AKL). In the Indonesian language, so far there are only two studies on the Indonesian academic vocabulary, i.e. by Kwary (2016) and Rhubido (2017). Kwary (2016) used a small number of data and the results is only a short list of academic words. Rhubido (2017) used a bigger corpus and produced a longer list, following the method used by Coxhead (2000) to determine the AWL. The Indonesian Advanced Learner’s Dictionary which is still being made by the Indonesian National Language Agency needs to put a label for academic words. Consequently, it is necessary to conduct a comprehensive study to determine which words should be labelled academic. This study suggests that academic words should be selected using three criteria: keyness, frequency, and range. This study found 310 word types which are included in the Indonesian Academic Vocabulary. These words are suggested to be labelled academic in the Indonesian Advanced Learner’s Dictionary.

Keywords: Academic words, Learner dictionary, Indonesian language, Word type

1. Introduction

A number of researchers have mentioned the significant role of vocabulary in language teaching and learning (Carter & McCarthy 2013; Nation 2001; Schmitt 2010). Accordingly, studies on vocabulary has developed widely around the world, particularly those using the corpus linguistics approach. However, vocabulary studies using the corpus linguistics approach are still rarely conducted in Indonesia. In the case of academic vocabulary, initial studies have just been started by Kwary (2016) and Rhubido (2017). However, these studies still need further improvement in order to determine the right academic vocabulary, particularly for the purpose of labelling the words in an advanced learner’s dictionary.
Kwary (2016) writes a short article comparing the Indonesian general vocabulary with the Indonesian academic vocabulary. Kwary (2016) shows that the Indonesian vocabulary in general text is characteristically different from the vocabulary in academic text. The differences can be seen in terms of the words included in the high frequency words, the number of syllables in the high frequency words, and the number of loan words. Kwary (2016) concluded that in the top 100 words, there is a 41% difference between the general text vocabulary and the academic text vocabulary; the number of syllables of the words in the general text is smaller than that in the academic text; and, the academic words use 24% more loan words than the general words. Kwary’s study, however, only presented a small number of results, i.e. only the top 100 words, and the analysis was not quite comprehensive since it was only a short article publication.

Rhubido (2017) produced a longer list of Indonesian academic words. Rhubido (2017) adopted the method used by Coxhead (2000), who created the English Academic Word List, to formulate the Indonesian Academic Word List. Rhubido (2017) used a corpus taken from 200 Indonesian national journal articles comprising 672,737 running words. Rhubido created a word list containing 250 word families, following Coxhead (2000) who also used word families. Several scholars have criticized the use of word families in an academic word list (cf. Hyland and Tse 2007; Paquot 2007; Gardner and Davies 2014; Lei and Liu 2016). Gardner and Davies (2014), for example, mention that the use of word families in academic word list formulation causes problems on semantic aspects, and these word families do not consider grammatical parts of speech.

The needs for an Indonesian academic word list have increased in the past few years due to the government’s requirements for graduate students to write and publish academic articles before they graduate. In addition, the academic word list is also needed to label the academic words in the new dictionary called the Indonesian Advanced Learner’s Dictionary which is being created at Badan Bahasa (the National Language Agency). The academic word list will be beneficial for graduate students who are still new in writing academic articles and for Badan Bahasa to determine which words in the Indonesian Advanced Learner’s Dictionary should be labelled academic. Consequently, it is necessary to conduct a comprehensive study to formulate the Indonesian academic word list. This study aims to suggest that academic words should be selected using a frequency class ratio or keyword analysis by using Paquot (2010) method. This study also recommends that an academic word list for a dictionary labelling purpose should be based on word types, instead of word families or lemmas, and should not be excluded from high frequency words.

2. Method of the Study

In general, there are two methods that can be used to determine academic words, i.e. the AWL (Academic Word List) method used by Coxhead (2000) and the AKL (Academic Keyword List) method used by Paquot (2010). To formulate the AWL, Coxhead (2000)
compiled 414 academic texts comprising 3,513,330 tokens (running words), and used the following three criteria: (1) specialized occurrence, (2) range, and (3) frequency. The criterion specialized occurrence means that the academic word families do not belong to the high frequency words, or more precisely the General Service List (West 1953). This means that Coxhead first deleted all the GSL words from the list. Then, the range criterion is implemented by selecting the words that occur across the academic disciplines. Finally, the frequency criterion is used, where the words selected must occur at least 100 times in corpus. The final result comprises 570 word families, called the Academic Word List (AWL).

The AKL was created using two academic corpora namely Micro-Concord Corpus Collection B (hereafter: MC) and the Baby BNC Academic Corpus (hereafter: B-BNC). The MC comprises 33 book sections and the B-BNC is made up of 30 book sections and scientific journals. The total is 2,026,067 running words. Different from the AWL, the AKL was formulated by Paquot (2010) using the following criteria: (1) Keyness, (2) Range, and (3) Evenness of distribution. The keyness is a score assigned to a word based on the frequency of a word in the target corpus in comparison to that in a reference corpus. Therefore, in the AKL formulation, the GSL is not used to delete the words (the GSL was used in the AWL formulation). Paquot used a statistical test called log-likelihood to determine the keyness scores. Then, the range is also used to select the words. Lastly, the evenness of distribution criterion is used. This refers to a statistical coefficient of how evenly distributed a word is across successive sectors of the corpus. The final result comprises 930 lemmas, which are further categorized into nouns (355 word types), verbs (233 word types), adjectives (180 word types), adverbs (87 word types), and others (75 word types).

In formulating the Indonesian Academic Vocabulary, this current study uses the following three criteria: (1) Keyness, (2) Frequency, and (3) Range. The explanations on the implementation of these criteria in selecting the Indonesian Academic Vocabulary are presented in the fourth section of this paper.

3. The Corpora of the Study

Since the keyness is included as one of the criteria, this study needs to have two corpora: a reference corpus and a target corpus. In a keyword analysis, the target corpus should be compared with a larger or equal corpus called the reference corpus. In this research, we use the Indonesian corpus created at Leipzig University as the reference corpus. This corpus can be downloaded from the website: http://wortschatz.uni-leipzig.de/en/download. There corpus contains 4,531,786 tokens of Indonesian words taken from news, websites, and Wikipedia.

The target corpus used in this study was compiled by the authors. The corpus comprises the articles published in national journals which were accredited by the Indonesian Directorate General of Higher Education (DIKTI) in 2012-2016. Being an accredited journal by DIKTI means that the journal has shown a measurable quantity and quality in scientific community in Indonesia and it has followed the standards of
writing professions nationally and internationally (Ditlitabmas 2017). To ensure the coverage of various disciplines, the journal articles are classified into four, following the classification in Scopus, they are: (1) Health Sciences, (2) Life Sciences, (3) Physical Sciences and (4) Social Sciences.

The distribution of accredited national journals across those four disciplines is not equal, so the number of articles selected in one discipline can be different from that in another discipline. In the data of accredited journals, there are only 8 journal titles in Physical Sciences, 11 journal titles in Life Sciences, 12 journal titles in Health Sciences, and 43 journal titles in Social Sciences. In Physical Sciences, we downloaded 30 articles per journal title, in Health Sciences and Life Sciences we downloaded 20 articles per journal title, and in Social Sciences we downloaded 10 articles per journal title. The final data consist of 3,734,743 tokens (see Table 1).

**Table 1 The Distribution of the Corpus**

<table>
<thead>
<tr>
<th>No</th>
<th>Science Disciplines</th>
<th>Selected Journal Articles</th>
<th>Running words (tokens)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Social Sciences</td>
<td>430</td>
<td>1,995,369</td>
</tr>
<tr>
<td>2</td>
<td>Health Sciences</td>
<td>240</td>
<td>523,982</td>
</tr>
<tr>
<td>3</td>
<td>Physical Sciences</td>
<td>240</td>
<td>654,669</td>
</tr>
<tr>
<td>4</td>
<td>Life Sciences</td>
<td>220</td>
<td>560,723</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1130</td>
<td>3,734,743</td>
</tr>
</tbody>
</table>

4. The Formulation of the Indonesian Academic Vocabulary

As mentioned in Section 2, an academic word can be formulated using the AWL method (Coxhead 2000) or the AKL Method (Paquot 2010). The AWL uses the following three criteria: (1) specialized occurrence, (2) range, and (3) frequency. The AKL uses the following three criteria: (1) keyness, (2) range, and (3) evenness of distribution. In this study, we use three criteria adapted from those two methods. We do not use the specialized occurrence criterion because we believe that a high frequency word may also be considered an academic word if it has a key role in academic text. We also do not use the evenness of distribution criterion because it is very similar to the range criterion. Consequently, we use the following three criteria: (1) Keyness, (2) Frequency, and (3) Range.

To determine the keyness, the 3,734,743 tokens of the target corpus should be compared with the reference corpus, i.e. the 4,531,786 tokens from the Leipzig corpus. To calculate the keyness using the log-likelihood test, we use AntConc software (Anthony, 2015). This software automatically orders the words based on their keyness scores. In this first step, we chose 2000 word types with the highest keyness scores. In this second step, we use the frequency criterion. In this case, we only selected the words which have a minimum frequency of 1000. We also excluded abbreviations, such as ph, ppm, mg, and ml. In this step, we only have 320 word types left to be the
Indonesian academic word candidates.

For the last step, we use the range criterion. In this case, we first use the software called AntProfier (Anthony 2014) to determine the range of every word in our data. Since there are four disciplines on our data, we only chose the words with a minimum number of range is four. This long list of words is then compared with our 320 word types. We found that there are 310 word types out of the 320 word types which have a range of four. We therefore conclude that these 310 word types are the Indonesian Academic Vocabulary. The list of these words can be seen in the Appendix.

The Indonesian Academic Vocabulary does not only contain content words (i.e. nouns, verbs, adjectives, and adverbs), but also function words. The examples of function words in this Indonesian Academic Vocabulary are: pada (on), yaitu (that is), dengan (with), dapat (can), dan (and), dalam (in), and oleh (by). These findings are similar to those of Paquot’s. In Paquot’s (2010) AKL, there are 73 words out of 930 word types which are function words. The examples of function words in the AKL are, although, an, as, several, since, some, and whereas.

5. The Labelling of the Academic Words in the Indonesian Advanced Learner’s Dictionary

The words in the Indonesian Advanced Learner’s Dictionary which should be labelled academic are the word types, and these are not necessarily the headwords. A headword may contain several derived word types, but not all word types are academic words. In another case, the headword can be labelled academic, but the derived words are not necessarily labelled academic. Since the Indonesian Advanced Learner’s Dictionary is still in the compilation process, the screenshots used in this paper are taken from the Indonesian Comprehensive Dictionary, which is the larger version of the dictionary. For example, the headword konsumsi (consume) has the following derived word types: mengonsumsi, and pengonsumsian (See Figure 1). The word konsumsi is in the 310 Indonesian Academic Vocabulary, but the words mengonsumsi, and pengonsumsian are not in the Indonesian academic Vocabulary. Thus, the word konsumsi should be labeled ‘academic’ in the dictionary, while the words mengonsumsi and pengonsumsian should not be labelled academic.
In another case, the headword should not be labelled ‘academic’, but some of the derived word forms should be labelled ‘academic’. One of the examples is the headword tahu. This headword has several derived words, i.e. berpengetahuan, bertahu-tahu, bertahuan, bertahukan, ketahu-tahuan, mengetahui, pengetahuan, sepengetahuan, setahu, tahu-menahu, tahu-tahu. However, only the words mengetahui and pengetahuan are included in the 310 word types of Indonesian Academic Vocabulary. Therefore, only the words mengetahui and pengetahuan should be labeled ‘academic’ (see Figure 2).
6. Conclusion

Studies on vocabulary, particularly those using corpus linguistics approach have developed widely around the world. In Indonesia, however, the studies are still at the initial stage. This study uses the corpus linguistics approach to determine the Indonesian academic vocabulary. Using a corpus of 3,734,743 word tokens taken from 1,130 Indonesian journal articles, and using the following three criteria: keyness, frequency, and range, this study managed to formulate the Indonesian academic vocabulary. The list contains 310 word types which comprise both content words and function words. These words should be considered for placing the label ‘academic’ to the words in the Indonesian Advanced Learner’s Dictionary.

References

Appendix: The 310 word types of the Indonesian Academic Vocabulary

penelitian, pada, hasil, nilai, tabel, uji, terhadap, menunjukkan, analisis, dilakukan, variabel, tanaman, perlakuan, data, yaitu, berdasarkan, menggunakan, metode, digunakan, kadar, model, sampel, rata, gambar, sebesar, signifikan, jumlah, pengaruh, suhu, merupakan, faktor, berpengaruh, dengan, kelompok, perusahaan, diperoleh, dapat, daun, pakan, pengujian, dan, dibandingkan, tinggi, media, dalam, terdapat, bobot, konsentrasi, asam, penggunaan, memiliki, secara, proses, tingkat, positif, pertumbuhan, peningkatan, kontrol, bakteri, meningkatkan, protein, kinerja, jenis, responden, pembelajaran, antara, sosial, larutan, pengukuran, penurunan, rendah, aktivitas, dosis, kriteria, pasien, kandungan, regresi, koefisien, skor, risiko, persamaan, pendahuluan, kualitas, persentase, masyarakat, bahan, sebanyak, didapatkan, karakteristik, pemberian, perilaku, perbedaan, hipotesis, total, benih, sedangkan, sehingga, struktur, permukiman, sel, pola, bertujuan, tersebut, ukuran, fungsi, produk, pengamatan, parameter, rasio, nyata, dilihat, dimensi, penambahan, indikator, komunikasi, buah, pembahasan, kabupaten, masing, laju, statistik, sumber, sebagai, dihasilkan, menghasilkan, adanya, konsumsi, sistem, kondisi, keuangan, pengambilan, laba, serta, disebabkan, tipe, karyawan, berupa, perempuan, individu, oleh, produksi, negatif, aspek, daya, persepsi, lingkungan, bentuk, melalui, terdiri, manajemen, kajian, menyebabkan, dipengaruhi, kategori, lahan, meliputi, komponen, hubungan, populasi, teknik, antar, umur, konsumen, informasi, perubahan, selanjutnya, air, kunci, pendekatan, konsep, berbeda, mempengaruhi, sesuai, budaya, mengalami, usia, permukaan, berikut, bahwa, ruang, tertia, strategi, berbasis, kemampuan, bersifat, pengembangan, periode, ekonomi, kesimpulan, relatif, meningkat, sektor, tekanan, reaksi, penyakit, penilaian, politik, undang, terjadinya, ditemukan, maupun, publik, efek, pendapat, ikan, konteks, perkembangan, obat, status, kerja, laki, tahap, daerah, komunitas, jaw, tujuan, hukum, tanah, berkaitan, terkait, cenderung, standar, teori, menyatakan, jaringan, peneliti, permasalahan, pertanian, lokal, kegiatan, kawasan, laporan, diberikan, prinsip, warna, normal, pendidikan, menit, islam, aktif, pelaksanaan, mengandung, diketahui, adalah, utama, lebih, potensi, memberikan, mempunyai, studi, apabila, menentukan, semakin, keputusan, unsur, berat, wilayah, pasal, hutan, efektif, kebijakan, industri, panjang, memenuhi, peran, karakter, energi, angka, makna, lokasi, biaya, fisik, kesehatan, sebagaimana, organisasi, baik, umum, terjadi, ketiga, tertentu, pengetahuan, upaya, mampu, darah, kebutuhan, modal, membentuk, khususnya, luas, alat, diperlukan, mengetahui, materi, selama, tindakan, desa, bangunan, diharapkan, berasal, dimana, teknologi, akibat, dasar, sifat, hal, umumnya, hak, badan, jam, yakni.
Characteristics of Lexical Items Appealing on the Websites of Japanese Corporations

Toshikazu Ezure
Aomori Public University
tezure@b.nebuta.ac.jp

Abstract

In Japan, corporate governance and investor relations have gradually been recognized as significant for developing business globally. This idea is leading Japanese companies to disclose more information proactively in English. However, linguistic research on the use of English in this area has been lacking. Therefore, this paper compared material released on the web by Japanese major corporations with that by their U.S. counterparts. First, I suggested a hypothesis about some differences in the use of words and phrases found in the Japanese and U.S. corporate materials. Second, I collected material from official websites and count the content words. The material was collected from Toyota, Sony, Ford, and Apple, among others. Third, I made vocabulary lists of frequent lexical items. The research revealed the fact that there are several words appearing with unnaturally high frequency levels in the English-language material from the Japanese companies. They were repeatedly used without being replaced by pronouns. This use of lexical items maintains the exactness of materials, as if they were technical documents. On the other hand, it may in some readers create a too formal impression. And finally, I referred to descriptions of some EFL dictionaries to introduce tips for better usage for Japanese learners. These findings can have the following applications: 1) Japanese corporations can improve their future written English materials, using more natural and idiomatic English; 2) We can compile a list of important words and phrases for Japanese students of business; and 3) The list can provide pointers for classroom teaching, textbook editing, and the publishing of teaching materials.

Keywords: business English, lexical items, corpus, Japanese corporations

1. Introduction

Characteristics of lexical items have been analyzed according to different cultures. For example, Adler (1992: 29-30) provided an organizational chart in a company and some examples of miscommunication such as lack of awareness and unclear messages. Nishimura (2007: 177) analyzed Japanese verbs appearing in the Bulletin Board Systems, where two-thirds had formal forms. Mikhailov and Cooper (2016: 186) noted
that the Finnish verbs which meant “to enjoy” were overused in advertisements translated from Finnish into Russia. In Japan, corporate governance and investor relations have gradually been recognized as significant for developing business globally. This idea is leading Japanese companies to disclose more information proactively in English. However, linguistic research on the use of English in this area has not been sufficient. Therefore, this paper will compare material released on the web by Japanese major corporations with that by their U.S. counterparts.

2. Hypothesis and Methodology

I will suggest a hypothesis that there are some characteristics of lexical items which are highly appealing on the websites of Japanese corporations. Use of English words and phrases adopted on the Japanese websites are influenced by Japanese language and differentiate from those on English websites. Furthermore, plainer English words and phrases will be used more frequently on Japanese websites than the U.S. counterparts.

In order to test this hypothesis, I collected materials from the official websites of various industries. They include the official website of Japanese corporations such as Toyota from automobile industry, Sony from electronics, Tokyo-Mitsubishi UFJ from banking, UNIQLO from apparel, and Lion from commodities. The U.S. counterparts are Ford, Apple, JP Morgan Chase, Forever 21, and P&G. These corporations are selected because they are represented in Japan.

The lexical items on the websites are analyzed by a concordancer named Antconc windows version 3.4.4., since this software is easy for me to operate. I have used ‘the keyword function’ where comparative analysis between two concordances can be conducted. When the score of keyness of a word in a concordance is higher, the word is more frequently used than in another concordance. On the other hand, when the score of keyness is lower, the word is less frequently used.

3. Results

Table 1 shows the results of the most frequent and least frequent words of materials on the Japanese websites. Only content words are extracted from the raw data. Materials on the website of Japanese corporations include 2491 word tokens and 823 word types. On the other hand, materials on the website of the U.S. corporations include 3421 word tokens and 991 word types.

---

1 Antconc is a freeware corpus analysis toolkit that analyzes the frequency of words and collocations, as well as difference between two concordances. The software can be downloaded from http://www.laurenceanthony.net/software.html.
2 Keyness was measured without a lemma file, since some entry errors occurred.
Table 1 The most frequent and least frequent words of the Japanese websites

<table>
<thead>
<tr>
<th>Number</th>
<th>Frequency</th>
<th>Keyness</th>
<th>Word</th>
<th>Frequency</th>
<th>Keyness</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14</td>
<td>24.200</td>
<td>Lion</td>
<td>2</td>
<td>0.002</td>
<td>key</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>21.145</td>
<td>led (LED)</td>
<td>1</td>
<td>0.098</td>
<td>shared</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>20.743</td>
<td>flicker</td>
<td>1</td>
<td>0.098</td>
<td>reduce</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>20.743</td>
<td>image</td>
<td>1</td>
<td>0.098</td>
<td>real</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>20.743</td>
<td>sensor</td>
<td>1</td>
<td>0.098</td>
<td>providing</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>19.015</td>
<td>automotive</td>
<td>1</td>
<td>0.098</td>
<td>pieces</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>17.947</td>
<td>high</td>
<td>1</td>
<td>0.098</td>
<td>open</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>17.286</td>
<td>kaws (KAWS)</td>
<td>1</td>
<td>0.098</td>
<td>need</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>17.286</td>
<td>Sony</td>
<td>1</td>
<td>0.098</td>
<td>moving</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>13.829</td>
<td>bacteria</td>
<td>1</td>
<td>0.098</td>
<td>lower</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>13.829</td>
<td>cameras</td>
<td>1</td>
<td>0.098</td>
<td>like</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>13.829</td>
<td>gum</td>
<td>1</td>
<td>0.098</td>
<td>home</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
<td>13.829</td>
<td>hdr (HDR)</td>
<td>1</td>
<td>0.098</td>
<td>highly</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>13.829</td>
<td>performance</td>
<td>1</td>
<td>0.098</td>
<td>financial</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
<td>12.100</td>
<td>China</td>
<td>1</td>
<td>0.098</td>
<td>existing</td>
</tr>
<tr>
<td>16</td>
<td>7</td>
<td>12.100</td>
<td>mitigation</td>
<td>1</td>
<td>0.098</td>
<td>environment</td>
</tr>
<tr>
<td>17</td>
<td>7</td>
<td>12.100</td>
<td>product</td>
<td>1</td>
<td>0.098</td>
<td>effort</td>
</tr>
<tr>
<td>18</td>
<td>7</td>
<td>12.100</td>
<td>Snoopy</td>
<td>1</td>
<td>0.098</td>
<td>designed</td>
</tr>
<tr>
<td>19</td>
<td>10</td>
<td>11.678</td>
<td>quality</td>
<td>1</td>
<td>0.098</td>
<td>creation</td>
</tr>
<tr>
<td>20</td>
<td>6</td>
<td>10.372</td>
<td>Japan</td>
<td>1</td>
<td>0.098</td>
<td>commitment</td>
</tr>
</tbody>
</table>

The most frequent words include the names of corporations such as ‘Lion’ and ‘Sony.’ This means that the names are introduced as subjects, or possessive forms rather than ‘we’ or ‘our’ on the Japanese corporations’ website. For instance, the name appears in the sentence that ‘Lion will continue working to increase the use of plant-derived materials in its products.’

Similarly, proper nouns like ‘led (LED)’ and ‘kaws (KAWS)’ are favorably adopted on the Japanese corporations’ sites. Pronouns are excluded by the expressions such as ‘LED lights’ and ‘KAW’s latest collection’. Furthermore, technical terms including ‘automotive cameras’ and ‘flicker mitigation’ have higher frequency than English corporations’ websites. These words can indicate their specifications and functions exactly. Therefore, these words appear without using another words or pronouns.

I hypothesized that plain and normal English words would be used more frequently than unfamiliar words on the websites of Japanese corporations, since it is supposed that intuitions of nonnative speakers of English have produced unnatural

---

3 KAWS is a name of an artist.
phrases. Hence, I guessed even childish expressions might be introduced. Practically however, proper names and formal technical terms are repeatedly adopted for precise description of their products.

4. Discussion for Pedagogical Applications

Some style books encourage learners of English to write clearly and exactly as a primary goal. For example, Alred, Brusaw, & Oliu (2015: 538-9) mention that, “The goal of technical writing is to enable readers to use a technology or understand a process or concept.” They point out the writing style emphasizes exactness and clarity rather than elegance or allusiveness. It seems that writers of Japanese corporations’ websites follow this policy. For Japanese learners of English, it might be easily acceptable to write precisely in a technical writing. On the other hand, there are few repeated technical words and phrases on the website of the U.S. corporations. The topics of the websites are brief introductions of their products. They are not pure technical documents. Nevertheless, we really need to pursue the preciseness even we lose natural use of plain English. Instead of repeated use of ‘LED light’, use of pronouns and plain expressions should be encouraged in the material of Japanese corporations. Scott (2015: 235) points out that writers should avoid jargon-laden phrases that are overused in their industry. The focus should be on keeping balance between formal and plain English on the websites of Japanese corporations. With these points in mind, Japanese corporations can improve their future written English materials, using more natural and idiomatic English.

The result that formal words are frequently adopted can make students more conscious about the use of lexical items. Like ‘automotive’, several formal words are used in the websites of Japanese corporations. As a pedagogical purpose, when we encourage students to consult dictionaries, we make them notice a speech label and make use of them for vocabulary building. For example, Oxford Advanced Learner’s Dictionary (OALD9) adds the ‘formal’ label to the word ‘mitigation’. Similarly, Longman Dictionary of Contemporary English (LDOCE6) adds the same label.

From the viewpoint of the least frequent word list, materials are necessarily composed of words in simple English such as ‘development’ and ‘new’. In real English, we need to compile a vocabulary list that is a mixture of simple, but significant words and formal words for learners of business.

Also, the list inspires students to focus on collocations. All the cases of ‘mitigation’ in the list collocate with ‘flicker’. Comparing with the data of Corpus of Contemporary English (COCA), there is no example of flicker + mitigation. Instead, ‘mitigation’ frequently collocates with ‘compensatory’, ‘change’, and ‘wetland’. Teachers can use this list for generating students’ interest and expanding their repertoire.

Bargiela-Chiappani, Nickerson, & Planken (2013: 55) claim that rapport management, where the use of pronouns are contrastively observed between professionals and students, represents a positive way of viewing business as a collaborative and ultimately relational human endeavor.
5. Conclusion

This paper conducted research looking for characteristics of lexical items appealing on the websites of Japanese companies. The findings showed that several words such as the company’s name and technical terms were repeatedly used on such websites. Also, some formal words appear on the websites of Japanese corporations as if they were technical documents. These cases were not noticed in the websites of the U.S. companies. Unfortunately, the research in this paper is limited. Therefore more data from websites should be collected and I will conduct further and extensive investigation.

Acknowledgement

I am grateful to Dr. Rausch Anthony for carefully proofreading the manuscript.

References

A. Dictionaries, Corpus and software

Anthony, L. 2014. AntConc (3.4.4.) [Concordancer].

B. Other Literature

Ford. n.d. The City of Tomorrow: Ford begins to explore new opportunities, solutions, and experiments that could shape the city of tomorrow.


Toyota. n.d. TNGA: Toyota new global architecture.

UNIQLO. 2017. UNIQLO Set to Launch KAWS X PEANUTS UT Collection.
Illustrative Examples in Bilingual Learners’ Dictionaries

Naho Kawamoto
Tokyo University of Foreign Studies
kawamoto.naho.o0@tufs.ac.jp

Abstract

The quality of illustrative examples receives increasing attention, and research on dictionary examples has been done from several approaches: (1) extracting good example candidates from corpora (Kilgarriff et al., 2008; Kosem et al., 2011) and (2) investigation on a particular aspect of dictionary examples (Ishii and Minn, 2015; Notohara, 2015). However, research on learners’ use of illustrative examples is still scarce. Whether dictionary examples are intelligible to users, which is one of three criteria of good illustrative examples (Atkins & Rundell, 2008), largely depends on the proficiency levels of dictionary users. Therefore, how learners at different proficiency levels use illustrative examples should be observed to explore the criteria of ‘good’ illustrative examples adjusted to the proficiency levels of dictionary users. Two independent studies have been conducted in order to identify factors which contribute to determine the difficulty of accurately using illustrative examples in learners’ dictionaries: an observation study (Study 1) and an experimental study (Study 2). In Study 1, 12 Japanese EFL learners’ references to illustrative examples were closely observed, using a think-aloud protocol. In Study 2, on the other hand, an experimental study was conducted to further explore factors which determine the difficulty or ease of illustrative examples for learners at three different proficiency levels. 172 participants were asked to translate Japanese sentences into English with the help of illustrative examples, which was carefully controlled by the following four variables: (a) formal properties of illustrative examples (phrasal versus sentential examples), (b) the number of examples (single versus multiple examples), (c) distance between Japanese and English, and (d) availability of mediation in Japanese. The results of these two studies revealed several possible factors which determine the difficulty levels of illustrative examples, and difficulty levels of vocabulary, the number of examples, presence of Japanese translations, closeness of illustrative examples and the Japanese translations, and naturalness of Japanese translations were some of contributing factors.

Keywords: illustrative examples, learners’ dictionaries, dictionary use, difficulty level, L1-L2 translation
1. Introduction

The quality of illustrative examples receives increasing attention, and research on dictionary examples have done from the following perspectives: (1) extracting good example candidates from corpora (Kilgarriff et al., 2008; Kosem et al., 2011) and (2) investing a particular components of examples, such as grammatical structure (Ishii & Minn, 2015), vocabulary (Ishii, 2011a; Ishii, 2011b), and readability (Notohara, 2015).

GDEX teams have operationalized variables of three criteria of good examples Atkins and Rundell (2008) suggest to meet: Natural and Typical, informative, and intelligible to users, so that they can automatically extract good example candidate from corpora by using a tool loaded in a software, Sketch Engine (Kilgarriff et al., 2004). The operationalized variables include not only positive variables, which examples should have, but also negative variables. By giving the sentences with positive variables good evaluation and penalizing the sentences with negative variables, GDEX provides the user with example candidates, which can be good illustrative examples based on the rule with operationalized variables, from the sentence with the higher score, 1.

However, when considering intelligibility, it goes without saying that whether or not a particular illustrative example is intelligible to users depends on the proficiency levels of dictionary users. Also, what advanced learners can do with illustrative examples and what less proficient learners can do with them are expected to be different. Kawamto & Tono (2015) observed how Japanese EFL learners used a Japanese-English dictionary for their encoding purposes, and they found how learners used illustrative examples differed among groups at different proficiency levels. During their observation, advanced learners used the information obtained from examples with their existing grammatical knowledge as well as their existing knowledge of vocabulary to maintain cohesiveness of their writing. On the other hand, less proficient learners had difficulties using examples effectively. This finding indicates that the difficulty levels of illustrative examples should adjust to the proficiency levels of the users of dictionaries, and a right way of selecting and presenting examples should be contrived so that all the users can interpret illustrative examples tuned to their levels.

Automatically extracting example candidates from corpora, as GDEX does, is a fantastic line of research. However, although GDEX has improved more for precision enhancement, in order to enable GDEX to select the best example candidates automatically form corpora, those examples have to be good in the sense that they are appropriate in light of proficiency levels of target users. To do so, additional criteria will be necessary. This might be stricter rules of controlling operationalized variables they define; however, it is also possible any other measurement other than those variables could contribute to the classification of examples into groups of examples at the same difficulty levels. In order to explore this possibility, profiling information of
learners would be extremely useful.

The purpose of this study is to explore the criteria of good illustrative examples in learners’ dictionaries adjusted to proficiency levels of the target users by observing how Japanese EFL learners of English actually use illustrative examples in their writings.

1.1 Research questions

Research questions in this study are as follows:

RQ1 What factors determine the difficulty levels of illustrative examples in dictionaries?
RQ2 How should illustrative examples in dictionaries be adjusted to the proficiency levels of the target dictionary users in terms of the factors found in RQ1?
RQ3 How do the factors found in RQ1 interact with the way illustrative examples are presented in a dictionary according to the proficiency levels of the target users?

2. Methods

Learners’ actual use of illustrative examples should be observed under as natural environment as possible; however, such an observation does not allow the author to generate the finding, which helps her to conclude what the criteria of ‘good’ illustrative examples are. Experiment with controlled variables is an only way to learn the truth in terms of variables causing difficulty of using illustrative examples. Therefore, in this study, observation was first made in order to narrow down the candidates of variables (Study 1), and then experiment with such variables was done (Study 2).

2.1 Study 1

As mentioned above, the first part of study (Study 1) was an observational study. The purpose of the first part of study was to closely observe how different proficiency learners used illustrative examples in their writings in natural environments, especially placing focus on (1) when they felt necessity of referring to examples, (2) how they chose one example to use from many, (3) what kind of information and in which form, in a form of phrase or word, they extracted from examples, and (4) how they used the information which they extracted from examples.
2.1.1 Participants

Four learners were selected for each of the three proficiency levels: elementary, intermediate, and advanced, and the 12 Japanese EFL learners were observed how they used illustrative examples in their L1-L2 translation task, using think-aloud protocols.

2.1.2 Procedures

Each of the 12 participants took part in a series of tasks, starting with (A) a questionnaire, followed by (B) a training session, (C) a L1-L2 translation task.

(A) Questionnaire

Questionnaire consisted of two parts. The first part of questionnaire had seventeen questions, and they asked the participants’ general information, their day-to-day dictionary use, their dictionary instruction, and the frequencies of their dictionary use and the purposes of dictionary consultation. The second part of the questionnaire was administered to the 12 participants in order to decide the operationalized definition of English proficiency levels in Study 1 (see Section 1.2). For the questionnaire, the Japanese version of the list of CEFR-J CAN-DO descriptors were administered, and they were asked to intuitively assess themselves on a four-point scale. As a result of questionnaire, the operationalized variables of English proficiency levels was defined as follows:

Operationalized variables of English Proficiency levels (Study 1)
- Elementary learners $\leq A1.2$
- $A2.1 \leq$ Intermediate learners $< B2.1$
- $B2.1 \leq$ Advanced learners

(B) Training task

The main task of this study was a L1-L2 translation task, and think-aloud protocols were conducted in order to elicit information in terms of what 12 Japanese EFL learners thought and what kind of information they wanted to obtain. The protocol allows researchers to elicit their participants’ mental process; however, it is unnatural for participants to think aloud what they are thinking. Therefore, it is said that precise instruction, model, and practice are necessary (Dörney, 2007) so that participants are comfortable enough to think aloud during taking part in experiment. Each of the 12 participants in Study 1 also took an approximately 15-minute training task.
Table 1 List of Dictionaries Used in This Study

<table>
<thead>
<tr>
<th>Levels of Dictionaries</th>
<th>Titles of Dictionaries</th>
<th>Edition</th>
</tr>
</thead>
<tbody>
<tr>
<td>General purpose</td>
<td>Meikyou Japanese Dictionary</td>
<td>2\textsuperscript{nd} edition</td>
</tr>
<tr>
<td>Elementary</td>
<td>Junior Progressive English Dictionary</td>
<td>1\textsuperscript{st} edition</td>
</tr>
<tr>
<td>Upper-Intermediate</td>
<td>O-LEX Japanese-English Dictionary</td>
<td>2\textsuperscript{nd} edition</td>
</tr>
<tr>
<td>Intermediate</td>
<td>CORE-LEX English-Japanese Dictionary</td>
<td>2\textsuperscript{nd} edition</td>
</tr>
<tr>
<td>Upper-Intermediate</td>
<td>Longman Active Study Dictionary</td>
<td>5\textsuperscript{th} edition</td>
</tr>
<tr>
<td>Advanced</td>
<td>Longman Dictionary of Contemporary English</td>
<td>6\textsuperscript{th} edition</td>
</tr>
</tbody>
</table>

(C) L1-L2 translation

The main task of Study 1 was a L1-L2 translation task. They were asked to translate approximately 170 to 220 Japanese characters (the same passage but a shorter passage for less proficient learners) into English, using the same set of seven printed dictionaries: one Japanese monolingual dictionary, two Japanese-English dictionaries, two English-Japanese dictionaries, and two English monolingual dictionaries (see Table 1).

2.2 Study 2

In Study 1, a careful observation was made to examine how learners at different proficiency levels used illustrative examples in dictionaries in their L1-L2 translation tasks, and their typical behavior of using illustrative examples in dictionaries was described using a think-aloud protocol (see 3.1). Although the close observation revealed Japanese EFL learners’ reference skills, it was still difficult to generalize the results and conclude the exact factors which determine the difficulty levels of examples. In order to make a solid proposal for the criteria of ‘good’ illustrative examples, another experiment study was desirable. The purpose of Study 2 was to verify whether the following four variables: (1) formal properties of illustrative examples, (2) the number of illustrative examples, (3) distance between Japanese and English, and (4) availability of mediation in Japanese, were actually factors which determined the difficulty of using illustrative examples. These factors included the one that the author wanted to verify but she could not gain any results from Study 1, as well as possible factors which were found to contribute to the difficulty levels from the results of Study 1.

Four operationalized variables of illustrative examples:

1. Formal properties of illustrative examples \[\text{phrasal vs. sentential} \text{ examples}\]
   The different formal properties: a form of full-sentence examples (sentential examples)
and a form of not full-sentence examples (phrasal examples) will be compared in terms of the accuracy rate of how well learners can extract information from phrasal/sentential examples and use information in their writings.

(2). Number of illustrative examples \([\text{single vs. multiple illustrative examples}]\)
The accuracy rate of how well learners can extract information from examples and they can use information will be compared between the case where they are provided with only one example (single example) and where they are provided with more than two examples (multiple examples).

(3). Distance between Japanese and English \([\text{high vs. low transparency}]\)
The results of Study 1 showed that how appropriately learners could extract information from illustrative examples largely depended on how apparent the English grammatical structure of illustrative examples were corresponding to the Japanese grammatical structure of translation of examples (transparency). Therefore, the author tries to verify whether the extent of transparency (high vs. transparency) makes any difference in how well learners can extract information from examples.

(4). Availability of mediation in Japanese \([\text{mediation available vs. not available}]\)
Some dictionaries published in Japan (e.g., the series of O-LEX and CORE-LEX dictionaries) provide the dictionary users with Japanese mediation in order to help them interpret illustrative examples well. Whether or not Japanese mediations will actually help Japanese EFL learners to extract information from examples, even though examples do not show close relationships between Japanese and English (low transparency), will be examined in Study 2.

2.2.1 Participants

172 students, 65 third year junior high school students, 36 freshman students from two different universities, and 71 sophomore students from three different universities, participated in the second half of the study, Study 2. All of the participants were first assumed to be learners at the elementary level. Students who had passed the pre-level 2 English Language Proficiency Test of STEP were grouped into the lower intermediate levels, and those who had passed higher levels of tests than the level 2 English Language Proficiency test of STEP were grouped into the intermediate levels. One thing to be mentioned here is that the proficiency levels in Study 2 were not exactly the same as those operationally defined in Study 1.
Table 2 is a comparison table of the proficiency levels of the participants in Study 1 and Study 2, and it describes how the participants’ proficiency levels differed between Study 1 and Study 2.
### Table 2 Comparison Table of Proficiency Levels of the Participants in Study 1 and Study 2

<table>
<thead>
<tr>
<th>CEFR-J</th>
<th>PreA1</th>
<th>A1.1</th>
<th>A1.2</th>
<th>A2.1</th>
<th>A2.2</th>
<th>B1.1</th>
<th>B1.2</th>
<th>B2.1</th>
<th>B2.2</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td></td>
<td>&lt;----</td>
<td></td>
<td>&gt;</td>
<td></td>
<td>&gt;</td>
<td></td>
<td>&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Elementary levels</td>
<td>Intermediate levels</td>
<td>Advanced levels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 2</td>
<td>&lt;--------&gt;</td>
<td>Lower-in</td>
<td>&lt;---&gt;</td>
<td>Int.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### 2.2.2 Procedures

The participants in Study 2 were asked to translate ten Japanese sentences into English. There were three sets of tasks: Task A, Task B, and Task C. Participants who were assigned to any of those three tasks (see Table 3 for the details) were asked to translate the same Japanese sentences, but they were provided with different illustrative examples.

### Table 3 Number of Participants in Each of Three Tasks and Proficiency Levels

<table>
<thead>
<tr>
<th></th>
<th>Task A</th>
<th>Task B</th>
<th>Task C</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>32</td>
<td>31</td>
<td>30</td>
<td>93</td>
</tr>
<tr>
<td>Low-intermediate</td>
<td>19</td>
<td>23</td>
<td>21</td>
<td>63</td>
</tr>
<tr>
<td>Intermediate</td>
<td>3</td>
<td>9</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>54</td>
<td>63</td>
<td>55</td>
<td>172</td>
</tr>
</tbody>
</table>

### Table 4 How the operationalized variables were built in Tasks

<table>
<thead>
<tr>
<th>Question #</th>
<th>Task A</th>
<th>Task B</th>
<th>Task C</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Variable 1: Forms of illustrative examples]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q 1</td>
<td>phrase</td>
<td>phrase</td>
<td>sentence</td>
</tr>
<tr>
<td>Q 3</td>
<td>sentence</td>
<td>sentence</td>
<td>phrase</td>
</tr>
<tr>
<td>Q 4</td>
<td>phrase</td>
<td>sentence</td>
<td>phrase</td>
</tr>
<tr>
<td>Q 5</td>
<td>phrase</td>
<td>phrase</td>
<td>sentence</td>
</tr>
<tr>
<td>Q 8</td>
<td>sentence</td>
<td>phrase</td>
<td>phrase</td>
</tr>
<tr>
<td>Q 9</td>
<td>phrase</td>
<td>phrase</td>
<td>sentence</td>
</tr>
<tr>
<td>[Variable: Number of illustrative examples]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q 5</td>
<td>Single</td>
<td>Multiple</td>
<td>Single</td>
</tr>
<tr>
<td>Q 7</td>
<td>Multiple</td>
<td>Single</td>
<td>Multiple</td>
</tr>
<tr>
<td>[Variable: Distance between English and Japanese]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4 illustrates how the illustrative examples were controlled with the operationalized variables, and the tasks were carefully designed to investigate how different properties of examples defined in the operationalized variables affect the participants’ accuracy rates in translation.

3. Results

3.1 Results of Study 1

The 12 Japanese EFL Learners’ dictionary consultation, especially their references to illustrate examples, during their L1-L2 translation tasks was observed in terms of the following four consultation processes: (1) for what purpose do learners of different proficiencies refer to illustrative examples, (2) how do they choose one illustrative example from all of the provided examples in dictionaries, (3) what kind of information do they extract from the illustrative examples they choose to use, and (4) how do learners use or integrate the information in their contexts, or their translation tasks in this case. Here, the typical behavior of each of the three English proficiency levels in terms of the four consultation processes will be briefly described.

3.1.1 For what purpose do learners of different proficiencies refer to illustrative examples?

For what purpose Japanese EFL learners referred to illustrative examples differed among the group of three proficiency groups: elementary, intermediate, and advanced learners. The purpose of references to illustrative examples were different among the different proficiencies of learners in terms of how specific the purpose they had at the time that they decided to consult a dictionary. Elementary learners did not have any intention of referring to illustrative examples; rather they consulted dictionaries in order to search for the English equivalents of particular Japanese words. On the other hand, dictionary references for the purpose of referring to illustrative examples were observed in advanced and intermediate learners. However, the sense of purpose of referring to them was first shown in the condition that intermediate learners knew a word in their minds. Some of them thought aloud that they wanted to check the illustrative examples to see how the word was used; however, it was usually for the second reference of a particular word, but not for the first reference. They felt the necessity to refer to illustrative examples because they did not know how to use a
particular word, although they had the word to use on their minds after dictionary consultations. On the other hand, advanced learners, who had a larger size of vocabulary, sometimes had words in their minds without any dictionary consultations. However, they were not sure how to use the word, which motivated them to consult dictionaries to verify how the word should be used in their writing. Of course, it was also the case that they had no idea of how to translate Japanese words into English. Even in this case, their purpose of referring to illustrative examples was more specific than those that intermediate learners had. Advanced learners wanted not only to know how the word was used but also find better examples that fit well to their writings they had already written.

3.1.2 How do they choose one illustrative example from all of the provided examples in dictionaries?

How to choose one illustrative example to use in their translation was also different among the three proficiency levels of learners. Elementary learners depended on the Japanese translation of examples. Some of the elementary learners could choose one even from the large number of illustrative examples, although they took considerable time to finally find one. Some dictionaries provide the users with a large number of illustrative examples to give them a variety of choices. In the case that information is divided into several categories, elementary learners know why there are tons of examples. The distinctions also help elementary learners to narrow down the choice of the illustrative example to use. On the other hand, when many illustrative examples are provided without any distinctions of meaning, elementary learners, who depend on Japanese translations of illustrative examples, do not know why dictionaries present so many examples, although the Japanese translations are very similar to each other.

More proficient learners have a considerable amount of knowledge; therefore, they could interpret the provided examples, which helped them to choose the appropriate one. They knew how dictionary examples consisted of grammatical structures. Therefore, even though a large number of illustrative examples was provided, they could choose one from them. Advanced learners seemed to be more skillful, and advanced learners often consulted dictionaries for relatively specific purposes. They knew exactly what kind of information they wanted to obtain from illustrative examples. Therefore, they could choose one example in a shorter time. Also, because of their specific purposes of dictionary consultation, they sometimes decided not to use examples in the case that dictionaries did not provide the information that they wanted.
3.1.3 What kind of information do they extract from the illustrative examples they chose to use?

The results of the observation about the use of illustrative example showed that learners at different proficiency levels differed greatly in the ability of extracting appropriate information from the illustrative examples in a dictionary. More proficient learners seemed to have more varieties in the way to extract information from the illustrative examples. Elementary learners had a limited range of knowledge; therefore, they had no choice but to largely depend on the Japanese translation of the provided illustrative examples. Elementary learners searched for the corresponding words or phrases in the illustrative examples of Japanese words, which they were going to translate, in the translation of the examples. That is, elementary learners depended on the typographical information such as types of fonts and how the words/phrases were highlighted, at the same time they depended on the Japanese translation. Depending on such information is a practical skill for elementary learners; however, extracting information by depending on the Japanese translation and the font can cause elementary learners to fail to extract the appropriate information. On the other hand, more proficient learners had more considerable knowledge of language; therefore, they could analyze the illustrative examples in terms of how they consisted of grammatical components and interpret them well. These skills helped intermediate and advanced learners to extract the appropriate information from illustrative examples.

3.1.4 How do learners use or integrate the information in their contexts, or their translation tasks in this case?

How appropriately learners integrate information they extracted from illustrative examples into their translation was different among learners at the three proficiency levels. More proficient learners had a wider range of knowledge of both vocabulary and grammar; therefore, they could combine information they obtained from illustrative examples with their existing knowledge. Most of them did not seem to have difficulty even in the case that they needed to integrate several pieces of information from examples. However, they made several mistakes, of course, because they were still learners of English and they were in the process of acquiring the language.

On the other hand, elementary learners often failed to integrate information into their writing. Because of their limited knowledge of the language, how accurately they could integrate was closely related to how they extract information from examples. Even though proficient learners did not extract enough information to make meaningful sentences, they could supply information if necessary, using their existing knowledge. However, elementary learners could not do so. Also, the two factors: how transparent the relationships of illustrative examples in English and the Japanese translation were and how similar the provided illustrative examples were to Japanese
passage which they tried to translate, seemed to determine the difficulty of how well elementary learners could make major or minor changes into information they extracted, as well as how quickly they could choose one examples to use and how well they can extract information.

3.2 Results of Study 2

The results of Study 2 will be briefly reported in this section because of space limitation, discussing whether the four variables (1) formal properties of illustrative examples, (2) the number of illustrative examples, (3) distance between Japanese and English, and (4) availability of mediation in Japanese, are actually contributing factors which determine the difficulty levels of illustrative examples.

In Study 2, the answer to each of ten questions was scored as one if learners successfully used the information in a target-like manner or zero if they did otherwise. Both the number of correct and incorrect answers, which was calculated by subtracting the number of correct answers from the number of all answers, was counted for comparison. Here, it should be mentioned that the correct answers do not necessarily mean that the participants did write sentences without any mistakes. If learners correctly extracted the target information under corresponding headwords, the learners were considered to perform as expected.

In order to examine complex relationships among variables, layered chi-squared tests were conducted. A software, js-STAR (version 2.9.9j, version □), was used for all of the statistical analysis performed in this thesis.

Phrasal vs. Sentential examples

Table 5 provides the result of comparing phrasal examples with sentential examples in terms of how well learners at each of the three proficiency levels could extract information from each form of the illustrative examples. It provides information, including (A) the number of observed valid answers, (B) the number of correct answers, (C) the number of incorrect answers, and (D) the accuracy rate.

<table>
<thead>
<tr>
<th>Phrasal Examples</th>
<th>Elm</th>
<th>Lower-inter</th>
<th>Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td>the number of all data observed</td>
<td>190</td>
<td>128</td>
<td>32</td>
</tr>
<tr>
<td>the number of correct answers</td>
<td>152</td>
<td>114</td>
<td>30</td>
</tr>
<tr>
<td>the number of incorrect answers</td>
<td>38</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>the accuracy rate</td>
<td>80%</td>
<td>89%</td>
<td>94%</td>
</tr>
</tbody>
</table>
Table 5 shows that there was a tendency for Japanese EFL learners to extract appropriate information from provided dictionary examples more correctly when they were provided with phrasal examples than when they were provided with sentential examples. Elementary learners could extract information from phrasal examples at the accuracy rate of 80%, while they could extract information from sentential examples at the accuracy rate of 76%. The lower-intermediate learners and the intermediate learners did better with phrasal examples (89% versus 77% and 94% versus 92%, respectively), as well. In spite of the fact that all of the three proficiency level of learners could extract the appropriate information from phrasal examples at a higher accuracy rate than they could from sentential examples, the differences of the lower-intermediate learners’ accuracy rate between when they were provided with phrasal examples and when they were provided with sentential examples seem to be relatively bigger.

Table 6 Accuracy Rate of Integrating Information into Their Writing

<table>
<thead>
<tr>
<th></th>
<th>Elm.</th>
<th>Lower-int.</th>
<th>Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phrasal Examples</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the number of all data observed</td>
<td>190</td>
<td>128</td>
<td>32</td>
</tr>
<tr>
<td>the number of correct answers</td>
<td>69</td>
<td>72</td>
<td>22</td>
</tr>
<tr>
<td>the number of incorrect answers</td>
<td>121</td>
<td>56</td>
<td>10</td>
</tr>
<tr>
<td>the accuracy rate</td>
<td>36%</td>
<td>56%</td>
<td>69%</td>
</tr>
<tr>
<td><strong>Sentential examples</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the number of all data observed</td>
<td>238</td>
<td>153</td>
<td>39</td>
</tr>
<tr>
<td>the number of correct answers</td>
<td>69</td>
<td>61</td>
<td>26</td>
</tr>
<tr>
<td>the number of incorrect answers</td>
<td>169</td>
<td>92</td>
<td>13</td>
</tr>
<tr>
<td>the accuracy rate</td>
<td>29%</td>
<td>40%</td>
<td>67%</td>
</tr>
</tbody>
</table>

Table 6 describes how well each of the three proficiency levels of learners could use information they extracted from examples. Table 6 shows that all of three levels of learners, elementary, lower-intermediate, and intermediate learners, drew the right
conclusions at a lower rate of accuracy than one of extracting appropriate information from either form of illustrative examples. There was still a tendency that the more proficient the learners were, the higher the accuracy rate was at which learners could integrate information into their writings, regardless of the different forms of illustrative examples. When comparing each proficiency levels of learners’ accuracy rate of integrating information from phrasal examples with their accuracy rate of integrating information from sentential examples, the accuracy rate was higher when they were provided with phrasal examples, and the tendency was observed in all of the three groups of different proficiencies. The elementary learners’ and intermediate learners’ accuracy rate of integrating information from phrasal examples was slightly higher than those from sentential examples (36% versus 29% for elementary learners and 69% versus 67% for intermediate learners); on the other hand, the intermediate learners’ accuracy rate of integrating information from phrasal examples, which was 56%, seems to be much higher than the accuracy rate of when they were provided with sentential examples (40%).

However, the results of layered chi-squared test showed that there were not any significant differences in the accuracy rate of both how well the elementary and intermediate learners could extract information from illustrative examples and how well they could use information in their writings between when they were provided with phrasal examples and they were provided with sentential examples. On the other hand, the difference in terms of formal properties of examples made significant differences in how well the lower-intermediate learners could extract information from examples and use the information in their writings $\chi^2 (1, N = 280) = 6.221; p < .05$; $\chi^2 (1, N = 281) = 6.859; p < .01$; $\phi = 0.156$), and the residual analysis showed that phrasal examples were easier for the lower-intermediate learners to extract information from the types of examples and apply the information to their writings than sentential examples.

*Single vs. multiple examples*
and Table 8 describes how well each of the three proficiency levels of learners could extract information from either single or multiple examples and how well they could use information they extracted from examples, respectively.

Table 7 Accuracy of Extracting Information from Illustrative Examples
**Table 7** Accuracy of Extracting Information from Illustrative Examples

<table>
<thead>
<tr>
<th></th>
<th>Elm</th>
<th>Lower-in t.</th>
<th>Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single Examples</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the number of all data observed</td>
<td>90</td>
<td>50</td>
<td>16</td>
</tr>
<tr>
<td>the number of correct answers</td>
<td>51</td>
<td>30</td>
<td>16</td>
</tr>
<tr>
<td>the number of incorrect answers</td>
<td>39</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>the accuracy rate</td>
<td>57%</td>
<td>60%</td>
<td>100%</td>
</tr>
<tr>
<td><strong>Multiple examples</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the number of all data observed</td>
<td>87</td>
<td>62</td>
<td>16</td>
</tr>
<tr>
<td>the number of correct answers</td>
<td>59</td>
<td>54</td>
<td>13</td>
</tr>
<tr>
<td>the number of incorrect answers</td>
<td>28</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>the accuracy rate</td>
<td>68%</td>
<td>87%</td>
<td>81%</td>
</tr>
</tbody>
</table>
shows that the more proficient the learners were, the better they could extract appropriate information from examples, regardless of whether they were provided with a single example or they were provided with multiple examples. There was only one exception; the lower-intermediate learners extracted information from multiple examples at higher rate of accuracy than the intermediate learners (87% versus 81%). When comparing the accuracy rate of extracting information from single examples with the accuracy rate of extracting information from multiple examples, the elementary learners and the lower-intermediate learners did better when they were provided with multiple examples. However, the intermediate learners reached the right answer at a higher rate of accuracy when they were provided with single examples. It is possible that the more proficient learners can extract appropriate information, regardless of the number of illustrative examples.

Table 7 shows that the elementary learners could integrate information slightly better when they were provided with single examples than when they were provided with multiple examples (39% versus 36%). Also, the intermediate learners could integrate information they obtained from single examples at a much higher accuracy rate than when they did so with multiple examples (81% versus 56%). On the other hand, the lower-intermediate learners showed the opposite tendency. They could integrate information into their writings when they used multiple examples at a higher accuracy rate than when they used single examples (52% versus 73%). There was still the tendency that the more proficient the learners were, the higher accuracy rate the learners could integrate information into their writings at, regardless of the different number of illustrative examples they were forced to use.
When comparing the numbers of that each of the three proficiency levels of learners could extract correct information from a single example with they could do so from multiple examples, significant differences were not observed in the group of elementary learners, \( \chi^2(1, N = 177) = 1.888; \text{n.s.; } \varnothing = 0.103 \), and the result of Fisher’s exact test shows that there were not any significant differences, \( p_{\text{two-tailed}} = 0.2258; \text{n.s.; } \varnothing = 0.322 \) in the group of intermediate learners; on the other hand, the difference in terms of the number of correct answers which the lower-intermediate learner could reach between the two cases where they were provided with single/multiple examples was significant, \( \chi^2(1, N = 112) = 9.442; p < .01; \varnothing = 0.290 \). Also, when comparing how accurate the Japanese EFL learners integrated information they obtained from single examples into dictionaries with how well they integrate information obtained from multiple examples, the relations between the number of provided examples and the elementary and intermediate learners’ frequencies of reaching the correct/incorrect answers were not significant, \( \chi^2(1, N = 117) = 0.806; \text{n.s.; } \varnothing = 0.022 \); \( \chi^2(1, N = 32) = 1.309; \text{n.s.; } \varnothing = 0.202 \). On the other hand, the relationships with these two variables were significant in the lower-intermediate levels, \( \chi^2(1, N = 112) = 4.204; p < .05; \varnothing = 0.194 \). The residual analysis showed that the lower-intermediate learners could extract information from illustrative examples and use information in their writings significantly better when they were provided with multiple examples.

**Table 8** Accuracy Rate of Integrating Information into Their Writings

<table>
<thead>
<tr>
<th></th>
<th>Elm</th>
<th>Lower-int.</th>
<th>Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single Examples</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the number of all data observed</td>
<td>90</td>
<td>50</td>
<td>16</td>
</tr>
<tr>
<td>the number of correct answers</td>
<td>35</td>
<td>26</td>
<td>13</td>
</tr>
<tr>
<td>the number of incorrect answers</td>
<td>55</td>
<td>24</td>
<td>3</td>
</tr>
<tr>
<td>the accuracy rate</td>
<td>39%</td>
<td>52%</td>
<td>81%</td>
</tr>
<tr>
<td><strong>Multiple examples</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the number of all data observed</td>
<td>87</td>
<td>62</td>
<td>16</td>
</tr>
<tr>
<td>the number of correct answers</td>
<td>31</td>
<td>45</td>
<td>9</td>
</tr>
<tr>
<td>the number of incorrect answers</td>
<td>56</td>
<td>17</td>
<td>7</td>
</tr>
<tr>
<td>the accuracy rate</td>
<td>36%</td>
<td>73%</td>
<td>56%</td>
</tr>
</tbody>
</table>

When comparing the numbers of that each of the three proficiency levels of learners could extract correct information from a single example with they could do so from multiple examples, significant differences were not observed in the group of elementary learners, \( \chi^2(1, N = 177) = 1.888; \text{n.s.; } \varnothing = 0.103 \), and the result of Fisher’s exact test shows that there were not any significant differences, \( p_{\text{two-tailed}} = 0.2258; \text{n.s.; } \varnothing = 0.322 \) in the group of intermediate learners; on the other hand, the difference in terms of the number of correct answers which the lower-intermediate learner could reach between the two cases where they were provided with single/multiple examples was significant, \( \chi^2(1, N = 112) = 9.442; p < .01; \varnothing = 0.290 \). Also, when comparing how accurate the Japanese EFL learners integrated information they obtained from single examples into dictionaries with how well they integrate information obtained from multiple examples, the relations between the number of provided examples and the elementary and intermediate learners’ frequencies of reaching the correct/incorrect answers were not significant, \( \chi^2(1, N = 117) = 0.806; \text{n.s.; } \varnothing = 0.022 \); \( \chi^2(1, N = 32) = 1.309; \text{n.s.; } \varnothing = 0.202 \). On the other hand, the relationships with these two variables were significant in the lower-intermediate levels, \( \chi^2(1, N = 112) = 4.204; p < .05; \varnothing = 0.194 \). The residual analysis showed that the lower-intermediate learners could extract information from illustrative examples and use information in their writings significantly better when they were provided with multiple examples.

**High vs. low transparency & Mediation available vs. not available**

Table 9 provides that the number of all observed valid answers, the number of correct answers, the number of incorrect answers, and the accuracy rate when each of the three different proficiency levels of learners: elementary, lower-intermediate learners, and
intermediate learners, was provided with (1) examples which had relatively clear relationships between Japanese and English (Close), (2) examples which did not have clear relationships between Japanese and English (Far), and (3) examples which did not have clear relationships, but includes Japanese mediations for help (Far + mediation). Table 9 indicates that there was a tendency that the higher proficiency levels of the learners were, the more accurately they could extract appropriate information from illustrative examples, regardless of how close the illustrative examples in English and the Japanese translation were and whether or not Japanese mediations were provided. There was only one exception, where lower intermediate learners could appropriately extract information from the illustrative examples, which showed close relationships between examples in English and the Japanese translation (Close), at the lowest accuracy rate (45%), while elementary learners could extract appropriate information from illustrative examples at the accuracy rate of 65%.

Table 9 Accuracy of Extracting Information from Illustrative Examples

<table>
<thead>
<tr>
<th></th>
<th>Elm.</th>
<th>Lower-int.</th>
<th>Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Close</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the number of all data observed</td>
<td>54</td>
<td>40</td>
<td>7</td>
</tr>
<tr>
<td>the number of correct answers</td>
<td>35</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>the number of incorrect answers</td>
<td>19</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>the accuracy rate</td>
<td>65%</td>
<td>45%</td>
<td>100%</td>
</tr>
<tr>
<td>Far</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the number of all data observed</td>
<td>53</td>
<td>43</td>
<td>7</td>
</tr>
<tr>
<td>the number of correct answers</td>
<td>35</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>the number of incorrect answers</td>
<td>28</td>
<td>22</td>
<td>3</td>
</tr>
<tr>
<td>the accuracy rate</td>
<td>47%</td>
<td>49%</td>
<td>57%</td>
</tr>
<tr>
<td>Far + mediation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the number of all data observed</td>
<td>57</td>
<td>43</td>
<td>12</td>
</tr>
<tr>
<td>the number of correct answers</td>
<td>22</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>the number of incorrect answers</td>
<td>35</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>the accuracy rate</td>
<td>39%</td>
<td>53%</td>
<td>75%</td>
</tr>
</tbody>
</table>

The results of the layered chi-squared test show that there were not any significant differences in how well Japanese EFL learners could extract information from illustrative examples between when they were provided with dictionary examples which differs in the closeness of Japanese and English. The differences were not significant in all three groups of proficiency levels: the elementary learners, $\chi^2(1, N = 117) = 0.688; \text{ n.s.}; \varnothing = 0.077$, the lower-intermediate learners, $\chi^2(1, N = 112) = 0.051; \text{ n.s.}; \varnothing = 0.025$. The result of Fisher’s exact test shows that there were not significant differences in how well the intermediate learners could
extract information from illustrative examples, regardless of differences in the closeness of Japanese and English, \( p = 0.1923; \alpha = .05 \).

Comparing the number of correct answers when Japanese EFL learners were provided with the Japanese mediations with the number of correct answers when they were not provided with the mediations, the result shows that there were significant differences in the groups of elementary learners, \( \chi^2(1, N = 120) = 2.805; .05 < p < .10; \varphi = 0.153 \). On the other hand, the result does not show that whether or not Japanese mediations were provided did not make any significant differences in the groups of lower-intermediate learners, \( \chi^2(1, N = 83) = 0.306; n. s.; \varphi = 0.06 \). Fisher’s exact test showed that there were not significant differences in the groups of intermediate learners as well, \( p_{\text{two-tailed}} = 0.6169; \varphi = 0.185; \text{the odds ratio is 0.44} \). The residual analysis shows that the elementary learners could extract information examples at a significantly lower rate of accuracy when they were provided with Japanese mediation.

4. Discussion

In this section, the following three research questions of this paper will be discussed based on the results of Studies 1 and 2:

RQ1 What factors determine the difficulty levels of illustrative examples in dictionaries?

RQ2 How should illustrative examples in dictionaries be adjusted to the proficiency levels of the target dictionary users in terms of the factors found in RQ1?

RQ3 How do the factors found in RQ1 interact with the way illustrative examples are presented in a dictionary according to the proficiency levels of the target users?

RQ1 What factors determine the difficulty levels of illustrative examples in dictionaries?

Table 3 describes the factors which were found in Studies 1 and 2 to determine difficulty levels of illustrative examples for Japanese EFL learners. It provides (1) factors determining difficulty levels, (2) whether the factors are positive (P) to help learners to use examples effectively, and (3) whether the factors are negative (N) to cause learners to have difficulty in using the examples.
Table 10 Factors Which can Determine the Difficulty of Illustrative Examples

<table>
<thead>
<tr>
<th>Study 1</th>
<th>Study 2</th>
<th>Proficiency levels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Elm.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Elm.</td>
</tr>
<tr>
<td>Vocabulary</td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>Typographical highlight</td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>Classification of meanings</td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>Forms of illustrative examples (phrasal examples)</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>Forms of illustrative examples (sentence examples)</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>Number of illustrative examples (single example)</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>Number of illustrative examples (multiple examples)</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>Distance between illustrative examples and their translations</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>Availability of Japanese mediation</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>Japanese translation</td>
<td></td>
<td>P/N</td>
</tr>
</tbody>
</table>

RQ2 How should illustrative examples in dictionaries be adjusted to the proficiency levels of the target dictionary users in terms of the factors found in RQ1?

RQ3 How do the factors found in RQ1 interact with the way illustrative examples are presented in a dictionary according to the proficiency levels of the target users?

4.1 Vocabulary

The result of Study 1 indicated that even though the Japanese translations were provided, less proficient learners had no idea what unknown words meant. Proficient learners have a wider knowledge of not only vocabulary but also grammar; therefore, even though they encounter a few unknown words in illustrative examples, they can guess which English word corresponds to a particular word in Japanese translation. On the other hand, it is challenging for less proficient learners to even find the English equivalents from examples. It might be a surprising fact for some; however, it might be natural at the same time when considering the viewpoint that learners can understand texts when they know at least 95 percent of all words (Laufer, 1989). Most of the previous studies, which try to automatically extract illustrative examples from corpora, use the frequency of words as a factor (Kilgarriff et al., 2008; Didakowski et al., 2012). This study supports that the frequency of words is one of the most important factors determining the difficulty of illustrative examples; therefore, lexicographers can adjust difficulty levels of illustrative examples by replacing a word with another.

Lexicographers of bilingual dictionaries might completely depend on the presence of the Japanese translations of illustrative examples. However, the result of Study 1
shows that the Japanese translations can help them well when dictionary users encounter only one unknown word. In the case that they encounter more than one unknown words, it is puzzling for less proficient learners to find how each of the English words correspond to Japanese words in the translation.

4.2 Number of illustrative examples

The number of dictionary examples provided in dictionaries can be a factor which determines the difficulty levels of illustrative examples. A larger number of illustrative examples caused less proficient learners to have difficulty in choosing one example to use. Study 1 showed that less proficient learners had difficulty in choosing one illustrative example when a large number of illustrative examples were provided, although Study 2 showed that the number of illustrative examples did not make any significant differences in how well the elementary learners could use illustrative examples.

The availability of online dictionaries allows lexicographers to include as many examples as they want; however, the results show that advanced learners, who have enough knowledge of both vocabulary and grammar to interpret illustrative examples, appreciate a plenty of illustrative examples shown in dictionaries. On the other hand, a long list of illustrative examples shown in online dictionaries when one consults a word in online dictionaries confuses less proficient learners. Less proficient learners have difficulty in choosing one from the long list of illustrative examples, and they face the difficulty of using the information in their writing next even though they succeed in choosing the appropriate one. The different number of illustrative examples made significant difference in how well the lower-intermediate learners could extract information and apply the information to their writings. Regardless of the number of provided illustrative examples, the intermediate learners could extract information and apply the information relatively well, because of their wider knowledge of vocabulary and grammar. On the other hand, the elementary learners could neither extract information well nor apply the information well, regardless of the number of illustrative examples. The reasons behind this can be that information they obtain from single illustrative examples was not enough, as single examples were for the lower-intermediate learners. However, when they were provided with multiple examples, it is possible that they could not deal with the excess of information to make meaningful sentences. If learners have enough knowledge of the language, they can make a decision on whether they should use or they should not use the obtained information, and they can even think of how to combine information from an example with information from another example, while using their existing knowledge. Kawamoto and Tono (2015) found that how Japanese EFL learners use illustrative examples differ between learners at a particular level and learners at the other proficiency levels. The differences seemed to be caused by how well they can combine
information they obtain from examples with their existing knowledge, which was a similar finding to Study 2. How three different levels of learners dealt with the number of examples was different.

4.3 Formal properties of illustrative examples (phrasal versus sentential examples)

The differences in terms of forms of illustrative examples, phrasal examples or sentential examples, were closely related to how much information illustrative examples had. This is one reason behind how well each of the three different proficiency levels of learners could extract information from illustrative examples and integrate information they obtained from phrasal/sentential examples into their writings. The shorter length of phrasal examples was usually shorter than one of sentential examples. Shorter length of illustrative examples means that learners extract information from smaller choices. This fact seems to be one reason why all learners at three different proficiency levels could extract information when they were provided with phrasal examples better than when they were provided with sentential examples. However, the result of the chi-squared test shows that different forms of illustrative examples, phrasal versus sentential examples, did not make any significant differences in how well the elementary learners and the intermediate learners can extract information. The different forms made significant differences only on the lower-intermediate learners.

On the other hand, the shorter length of phrasal examples also means that learners have to supply a larger number of words in order to make meaningful sentences, using information they obtain from examples. Phrasal examples and sentential examples are the same in terms of the condition that learners make major or minor changes into information in examples, such as supplying extra words and replacing a word with another, in order to apply the information to make meaningful sentences. However, sentential examples have more information in terms of what kinds of words (e.g., POS and subject/object) proceed or follow a particular headword; on the other hand, phrasal examples do not provide such information. That is, when learners use phrasal examples, they have to think of what kinds of words they should supply.

4.4 Distance between Japanese and English

The results of Study 1 showed that distance between Japanese and English seemed to be a factor which determined the difficulty of appropriately using illustrative examples. However, the results of Study 2 showed that the different distance between illustrative examples and their Japanese translations did not make any difference in how Japanese EFL learners could extract information. When comparing how well they did among the three groups of different proficiency levels of learners, the intermediate learners can
extract information from examples, which have close relationships between English and Japanese, significantly better than the other two levels of learners. The results indicated that the distance could be a factor which determines the difficulty levels of examples for more proficient learners. The different distance did not have any negative effect on less proficient learners.

4.5 Usefulness of mediation in Japanese

The results implicate that the availability of Japanese mediations can help more proficient learners to extract information, but not less proficient learners. The results showed that far from not helping them, the availability of Japanese mediations can make less proficient learner get confused. The results suggest that sufficient knowledge of not only English but also Japanese is necessary to make full use of information available in illustrative examples. Although adult learners can be elementary learners of course, it is true that elementary learners are still young. At the same time that young learners are in the process of acquiring English, their knowledge of Japanese is not fully developed. That is, young learners at the elementary proficiency levels have limited knowledge of not only English but also Japanese. The types of tasks, the L1-L2 translation tasks administered in both Study 1 and Study 2, might require the participants to have knowledge of Japanese. Limited knowledge of not only English but also their native language, Japanese, was one cause which leads to more difficulty for less proficient learners to extract appropriate information from examples and apply the information.

4.6 Japanese translation

Bilingual dictionaries provide the Japanese translation as well as illustrative examples, so that the dictionary users can understand what the illustrative examples mean. However, the result of Study 1 indicated that even though the Japanese translations were provided, less proficient learners had no idea what unknown words meant. Proficient learners have a wider knowledge of not only vocabulary but also grammar; therefore, even though they encounter a few unknown words in illustrative examples, they can guess which English word corresponds to a particular word in Japanese translation. On the other hand, it is challenging for less proficient learners to even find the English equivalents from examples.

4.7 Typographical highlight

One should not end the debate on the criteria of “good” illustrative examples by reaching the conclusion that typographical highlighting helps to learners to extract
appropriate information from illustrative examples. However, the author should mention here that Study 1 showed that typographical highlighting actually helped elementary learners, who were found to largely depend on typographical information. Typographical highlights allow the lexicographer to provide with information in terms of the grammatical structure of the examples, so that dictionary users can interpret how to make major or minor changes into the example to write meaningful sentences, using the illustrative examples. However, too much typographical highlighting can cause dictionary users to get confused; therefore, how effectively typographical highlighting was done should be considered carefully.

5. Conclusion

It is common knowledge that the quality and quantity of illustrative examples in reference materials should increase as proficiency levels go higher, but there is no clear understanding on exactly what criteria should play significant roles in determining the levels of difficulty of dictionary examples. This study provides empirical evidence on this important issue of how to control factors of selecting examples which are good for a particular target user at a given proficiency level. However, we still do not know properly how to control these variables and the combination of variables to adjust difficulty levels of examples. Further studies will be needed to examine how these variables work together to determine the levels of difficulties and ease of examples, so that this study would add significant value to the field of lexicography and applied linguistics.

References

Dictionaries


Other literature


A Study on the Homonyms in the Advanced-level International Chinese

Wenting Wei
The Education University of Hong Kong
s1122190@s.eduhk.hk

Shan Wang*
The Education University of Hong Kong
eduhk2018@gmail.com

Abstract

As an import component of language communication, vocabulary plays a significant role in teaching Chinese as a foreign language. In the advanced level of Chinese language learning, the amount of vocabulary that language learners acquire is the key for them to improve their communicative competence. This paper examines the homonyms from The Graded Syllable, Characters and words for the Application of Teaching Chinese to the Speakers of Other Languages (the Research Group of the Graded Syllable, Characters and words for the Application of Teaching Chinese to the Speakers of Other Languages 2010) that underwent definition changes in the fifth, sixth and seventh edition of Contemporary Chinese Dictionary (CCD). First, this paper compares the senses of all advanced-level words in the fifth, sixth and seventh edition of CCD and then selects and analyzes 5 homonyms in detail. The change in definition is mainly due to the increase or deletion of contents. Secondly, this paper examines the 5 homonyms in four learners’ dictionaries and compares them with CCD. They are A Dictionary of Chinese Usage:8000 words (Center for Chinese Linguistics PKU 2000), A Learner’s Chinese Dictionary – illustrations of the usages (Liu & Chen 2005), A Dictionary of HSK (Shao 2000), and The Commercial Press Learner’s Dictionary of Contemporary Chinese (Lu & Lv 2006). It is found that different dictionaries deal with them in different ways. Third, this paper uses the CCL corpus to annotate the senses of the 5 homonyms. The sense frequency information shows that if the senses are sorted by frequency, the order will be quite different from the existing dictionaries, which shows the importance of sense frequency in learners’ dictionaries. In sum, this study displays certain guiding significance in examining how homonyms should be treated in international Chinese teaching, especially in dictionaries.

Keywords: international Chinese teaching, advanced level, homonym, learner's dictionary, sense frequency
1. Introduction

Dictionary plays an important role in learning language. “As a bridge between native and foreign languages, Dictionary is an important tool for cross-cultural communication” (Zhao 2004). Most beginners may choose some bilingual dictionaries which are easy to acquire for assisting learning, “A bilingual dictionary acts as a supportive tool in learners’ initial stage. One of its most important functions is to help learners start the transition of using monolingual dictionaries.” (Zhang 2002). It’s thus obvious that as the beginners reach intermediate or even advanced level, their choice of dictionaries should also be adjusted.

International students with advanced level can use fluently Chinese for daily communication. “Vocabulary is very important in Teaching Chinese as a Foreign Language (TCFL). Only when the students grasp vocabulary can they communicate with Chinese. Otherwise, their communication will fail without vocabulary.” (Feng 2005). The demand of vocabulary for learners at this stage grows quickly. A monolingual dictionary can help those advanced learners immerse in the atmosphere of Chinese, which contributes to enlarging learners’ vocabulary storage and improving their accuracy so as to improve the level of communication.

Chinese monolingual dictionaries have a long history. Early in the Han Dynasty, Xu Shen compiled the first Chinese dictionary in China, Origin of Chinese Characters, and “the first dictionary of Chinese-English in the world” was A Dictionary of Chinese Language (Hua Ying Dictionary) which was not published until 1815 (or Chinese and English dictionary). By contrast, Chinese monolingual dictionaries have a deeper accumulation of precipitation. In addition, “monosyllabic vocabulary is mainly derived from the authoritative national newspaper” (Lu 2003). The scale, authority and normative of the monosyllabic dictionary outperform the bilingual dictionary, especially in the professional vocabulary, slang, dialect words and homograph, variant-form words. And most of the bilingual dictionaries are compiled based on the monolingual dictionaries.

Interpretations in the Chinese monolingual dictionary are not only for Chinese mother tongue learners, but also for the advanced Mandarin learners. Monolingual dictionary is an important learning tool and vocabulary source for advanced learners. The analysis of the changes in the interpretations of the monolingual dictionary can help the advanced learners to understand the vocabulary meanings and evolution of Chinese vocabulary more clearly and intuitively. Chinese monolingual dictionaries are more suitable for advanced learners both in theory and in practice. “with a strong normalization, theoretical, contemporaneity, CCD is recognized as the most authoritative reference tool in modern Chinese dictionary” (Zhang 2007). This study suggests that CCD can be used as a reference book for advanced learners.

While the advanced learners input large numbers of vocabularies, they also need to understand different types of vocabulary, including homographs, variant-form
words, homophones, polyphonic words and words with variant pronunciations. As an important vocabulary concept, homograph’s definition still receives many disputes in academia. So it requires learners to synthesize all different points of different schools, which virtually increases the difficulty of learning. Secondly, due to the characteristics in homographs, advanced learners rely heavily on the interpretations of vocabulary to distinguish them. It requires learners to understand and distinguish the similar interpretations of homographs, which is definitely a difficulty and challenge for advanced learners.

This paper selects nearly 80 polysemy words from the advanced vocabularies from The Graded Syllable, Characters and words for the Application of Teaching Chinese to the Speakers of Other Languages, (the Research Group of the Graded Syllable, Characters and words for the Application of Teaching Chinese to the Speakers of Other Languages 2010), the selected vocabularies which are also homographs in Contemporary Chinese Dictionary. This study found that the definitions of 11 homographs in fifth, sixth, seventh edition have been changed. This paper compares the changes, analyzes the reasons of the changes, and explores the paradigm that is suitable for advanced learners by comparing other dictionaries.

This study chooses the fifth, sixth and seventh version of Contemporary Chinese Dictionary as a blueprint, and all the interpretations, sentences derive from CCD. Therefore, it provides a scientific guarantee and solid theoretical basis for the source of the article corpus. This paper summarizes the changing characteristics of the interpretations of the synonyms in the vocabulary dictionaries with three different versions, and uses the data analysis method to summarize the changing laws so as to analyze the reasons for the changes. According to combination of the retrieval sentences in CCL corpus and the interpretations in CCD, this paper makes scientific analysis of the arrangement of senses.

It not only provides a theoretical reference for the revision and compilation of CCD and other related dictionaries compilation, but also helps advanced learners to take advantage of CCD to learn Chinese better.

2. Research Status

2.1 The Authority of Contemporary Chinese Dictionary

CCD is compiled by Institute of Linguistics CASS (Chinese Academy of Social Sciences), the dictionary with highest authority and normative language in China. CCD is not only an important source of corpus for Chinese mother tongue learners, but also a reliable reference tool for learners who learn Chinese as a second language. Since the first edition in 1978, CCD had been revised and replenished several times. The latest seventh edition was published in September 2016. The continuously
amendment of CCD plays a great role in promoting the spread of mandarin, Chinese standardization and cultural and educational development”.

“Mr. Lu Jianming thinks the ideal dictionary should have the following characteristics: clear targets; appropriate words; specification in the written forms of characters, words and pronunciation; accurate and condensed interpretations; proper illustration, plain language; feature in science, accuracy, simplification, and practicability” (Huang 2009), CCD “has reached a new height in the theory and compilation level as well as in the quality of editing. It’s a model of published dictionary”. CCD is an indispensable basis for Chinese learners and Chinese cultural workers. Many scholars and experts have been working on studying CCD since it was published in 1978. Numerous related papers were published later. For example, Li Feng’s examined the new words in Contemporary Chinese Dictionary (6th edition), and Wan, Ru & Cao, Wei’s Analysis the Revision in the Quantitative of Applications in Contemporary Chinese Dictionary (5th edition), which examines the feature of applications in Contemporary Chinese Dictionary (5th edition).

In ‘Contemporary Chinese Dictionary ‘Symposium Proceedings’, Lv & Hu (1996) mentioned that CCD has not only been widely used in China, but also has become an essential book for foreigners to learn Chinese. It is not only the reference book for native learners, but also an important learning tool for second-language learners owing to its authority and normalization.

Comparing the interpretations between CCD and other four dictionaries, Zhang (2008) counted the number of verbs in the same interpretation. He found that there were many similar interpretations among these dictionaries. The interpretations in these dictionaries inherited those from CCD to a great extent. CCD is a model of Chinese dictionaries’ compilation.

<table>
<thead>
<tr>
<th>Learner’s dictionary</th>
<th>The number of verbs which presents the same interpretations in the five learner’s dictionaries and Contemporary Chinese Dictionary</th>
<th>Proportion in the 74 verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contemporary Chinese Dictionary(1995)</td>
<td>17</td>
<td>23%</td>
</tr>
<tr>
<td>A Dictionary of Modern Chinese Usage(1995)</td>
<td>34</td>
<td>46%</td>
</tr>
<tr>
<td>A Learner’s Chinese Dictionary: Illustrations of the Usages(2005)</td>
<td>33</td>
<td>45%</td>
</tr>
</tbody>
</table>
Therefore, in terms of its practicality, authority, CCD is the ideal dictionary for advanced learners.

The fifth edition of CCD in 2005 is the one which was revised the most since the first publication in 1999. The fifth edition of CCD started to annotate word class, which makes it much easier to understand and reach “the peak” on the history of CCD. The number of research articles related to the seventh edition of CCD is not much because it wasn’t published for a long time, and most readers are not familiar with it.

Hence this study compares three different editions of CCD in different editions to research the changes of interpretations, which is for bringing some reference value to the compilation of Chinese learner’s dictionary. As a transition edition, the sixth version of CCD provides contextual information for those dictionaries, so this study also made a related introduction for it. In addition, it will help us to understand the stability of interpretations in CCD through the contrast of the changed interpretations in the continuous three versions CCD.

2.2 Function of Monolingual Dictionaries on Advanced Learners

Hu (1997) mentioned that we should require advanced students to use a dictionary, especially original Chinese Dictionary during advanced learning stage. For advanced learners, it is very important and necessary to choose a Chinese monolingual learner’s dictionary. The effect of Chinese monolingual learner’s dictionaries on advanced learners is reflected in the following three aspects.

Firstly, to get rid of the influence of mother tongue language, and to increase opportunities in touching and feeling Chinese (Zhang 2004). All monolingual dictionaries take advantage of the target language for query and interpretation, which not only prevents students from over depending on their first language, but also exercises the ability of students to use the target language, so that they have more opportunities and time to use the target language, and immerse in it.

Secondly, drawing inferences about other cases by analogy with one instance. Students can master harder words with the help of the words which have been learned before (Zhang 2004). Many compilers of monolingual dictionaries will try to pay attention to use some simple interpretations, and after understanding, deep thinking and predicting, advanced learners would master the vocabulary.

Thirdly, “the quantity and quality of monolingual dictionaries were assured” (Zhang 2002). The compilation scale and material supply of monolingual dictionaries is much higher than that of bilingual dictionaries. Only the specialty and the tolerance for vocabulary in monolingual dictionaries can meet the needs of advanced learners, especially in terms of the vocabulary collected, and the interpretations.
2.3 Characteristics of Advanced Chinese Learners in Learning Vocabulary

The main task of advanced Chinese learners is to deepen the understanding of meaning, and to clearly grasp the differences between the two languages. Meanwhile the learners in this stage should not only pay attention to the law of Chinese characters, but also understand the legitimacy of Chinese characters. There are three features of vocabulary learning for Chinese learners in this stage. One of them is that it is difficult to use those vocabularies which have learned, the other two are that the words fail to convey the meaning and they can’t understand the word with social and cultural factors (Zhang 1996). These are mainly due to the increased advanced vocabularies and the profound cultural tradition behind the words. According to requirements of HSK  

Chinese Language Proficiency Scales For Speakers of Other Languages”, HSK level 4 (elementary and intermediate level ) need to master 1200 words , while in level 6 (advanced level) 5000 vocabularies is required to master . It follows that the requirements of the breadth and depth of the vocabulary are very high”

2.4 Teaching Key Points and Difficult Points in the Teaching of Homographs

In Chinese speech, Wang (1995) said, “Homonym means that the pronunciation and the written form of the words are the same, but with different meanings. For example, jīn (巾), Towels, A scarf or shawl”. Then there are lots of papers enriching the concepts of homograph, such as, Cui (1957) The speech of acceptance in Contemporary Chinese; Wang & Wu (1983) Summary of contemporary Chinese vocabulary; Zhou (1995) The Dictionary of Homonym; Li (2005) A Theoretical Research about Homonymous Words; Cao (2010) Modern Chinese Vocabulary Study.

The accurate concept of Homonym wasn’t worked even if there were repeated deliberation. Although there are not too many homonyms, most of them belonged to the advanced vocabularies. Homonym is a specialized sphere of knowledge which the advanced learners can’t ignore and avoid.

“Homonym is widespread in Chinese, especially it’s a difficult point in Chinese teaching”, (Ren 2016). As an indispensable part, Homonym plays an important role in the Mandarin system and Mandarin teaching. Due to some historical reasons, international students are always confused by the homonym which has the same written form but with different meanings. Students not only need to understand the “pronunciation, shape” of homonyms, but also to differentiate and utilize the meanings of homonyms. Those unsolved problems in homonyms for native speakers are definitely much harder for international students. So how to lead the students to understand the concept of homonyms, and how to use it properly, is a difficult point in teaching Chinese as a second language.
3. Selection of the Advanced Chinese Homonym

The number and the difficulty of advanced Chinese vocabulary is greatly increased then other levels. Learning advanced vocabulary is not only the key point, but the difficult point.

The Graded Chinese Syllables, Characters and Words for the Application of Teaching Chinese to the Speakers of Other Languages is ‘the first national standard in Teaching Chinese to the Speakers of Other Languages’ (Li 2011). It didn’t come out until 17 years later after the publication of the Outline of the Graded Vocabulary for HSK in 1992, The Graded Chinese Syllables, Characters and Words for the Application of Teaching Chinese to the Speakers of Other Languages is a symbolic and contemporarily important project which takes 2 years to be researched. Thus, the advanced vocabularies in this study all come from The Graded Chinese Syllables, Characters and Words for the Application of Teaching Chinese to the Speakers of Other Languages, Research group (2010).

This paper examines all advanced vocabulary in CCD, in which five polysemous words were chosen from the homonyms to be dissected and expounded.

After comparing the interpretations of these 5 homonyms in the fifth edition (in 2005), the sixth edition (in 2012), and seventh edition (in 2017) of CCD, we are looking forward to obtain a better understanding of homonym and its interpretations.

4. The Transformations of the Interpretations in the Three Versions of “Contemporary Chinese Dictionary”

“The main task of the dictionary is generally borne by interpretation, the interpretation of a dictionary almost determines the quality of the dictionary”, (Han 1993). Thus it can be seen that how important the interpretation of a dictionary is to the learners. As the time goes by, language, a kind of culture carrier, made rapid development, and the interpretations of various dictionaries renews continuously too. Fu Huaiqing (2004) thinks that: “The main contents of the dictionary is to explain vocabularies”.

Comparing the interpretations of different dictionaries, figuring out the same and the different this article provides, very useful information for researching semantics.

Five interpretations in the three continuously versions of CCD have changed. As table 2 illustrates.
Table 2 Five changed interpretations in the three Continuously versions of *Contemporary Chinese Dictionary*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>④ gǒng (拱)</td>
<td>arch: arched door</td>
<td>arch: arched;</td>
<td>arch: arched;</td>
</tr>
<tr>
<td></td>
<td>arched dam</td>
<td>arched part of building;</td>
<td>arched part of building;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>arched gate</td>
<td>arched gate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bridge arch</td>
<td>bridge arch</td>
</tr>
<tr>
<td></td>
<td></td>
<td>arched dam</td>
<td>arched dam</td>
</tr>
<tr>
<td>④ mò (墨)</td>
<td>&lt;fig&gt; learning or the</td>
<td>&lt;Jiezhi&gt; learning or the</td>
<td>&lt;Jiezhi&gt; learning or the</td>
</tr>
<tr>
<td></td>
<td>ability to read:</td>
<td>ability to read:</td>
<td>ability to read:</td>
</tr>
<tr>
<td></td>
<td>uneducated;</td>
<td>uneducated;</td>
<td>uneducated;</td>
</tr>
<tr>
<td></td>
<td>unlearned;</td>
<td>unlearned;</td>
<td>unlearned;</td>
</tr>
<tr>
<td></td>
<td>unlettered;</td>
<td>unlettered;</td>
<td>unlettered;</td>
</tr>
<tr>
<td></td>
<td>completely illiterate</td>
<td>completely illiterate</td>
<td>completely illiterate</td>
</tr>
<tr>
<td>② wǎ (瓦)</td>
<td>earthen: earth basin</td>
<td>clay: clay basin</td>
<td>clay: clay basin</td>
</tr>
<tr>
<td></td>
<td>earthen ware</td>
<td>clay: clay jar</td>
<td>clay: clay jar</td>
</tr>
<tr>
<td></td>
<td></td>
<td>clay: clayware</td>
<td>clay: clayware</td>
</tr>
<tr>
<td>② yīxiàn (一线)</td>
<td>forefront: go to the</td>
<td>forefront: go to the</td>
<td>forefront: go to the</td>
</tr>
<tr>
<td></td>
<td>workshops and extend</td>
<td>workshops and extend</td>
<td>workshops and extend</td>
</tr>
<tr>
<td></td>
<td>solicitude to the workers at</td>
<td>solicitude to the workers at</td>
<td>solicitude to the workers at</td>
</tr>
<tr>
<td></td>
<td>the forefront of production.</td>
<td>the forefront of production.</td>
<td>the forefront of production.</td>
</tr>
<tr>
<td>⑥ zhèn (镇)</td>
<td>comparatively large</td>
<td>comparatively large</td>
<td>comparatively large</td>
</tr>
<tr>
<td></td>
<td>trading center</td>
<td>market or market town</td>
<td>trading center or town</td>
</tr>
</tbody>
</table>

(1) gǒng (拱)

Both of the sixth and the seventh versions add in an extra interpretation “arched part of building”. The latest version explains more fully than the fifth one. According to the renewed interpretation, the sixth, seventh versions add one more related example sentence “bridge arch” on the basis of the fifth version.

According to Contemporary Chinese Corpus, the frequency of usage in “bridge arch” ranked third, which is higher than “arched door”.

This study suggests that the added example sentence can be changed to a higher frequency “bridge arch”, which will meet the needs of readers better.

(2) mò (墨)
The interpretations of this character in sixth, seventh version are the same. Compared with the fifth version, the only difference is the rhetorical devices, metaphor and <Jiezhi>. “Metaphor is a kind of rhetorical devices which is to describe an object, taking advantage of of another one which is similar in nature but different in essence” (Yuan 1982). “‘Jiezhi’ means to choose a new constative which is related to the meaning of Original sentence in a particular situation” (Zhang 2009). After comparing these two rhetorical devices, this paper finds that “metaphor” which is only emphasis the similarity beside the essence, however, “Jiezhi” may reflects the inner relationship between two objects. It is clear that the compilers changed the interpretation after careful consideration.

(3) wǎ (瓦)

There are two meanings of “earth”: ①dirt, ②clay (Mo, 2001). The interpretations in the fifth, sixth, seventh version of Contemporary Chinese Dictionary says that “clay” is a kind of sticky earth which with less sand. “clay” is just one kind of “earth”, so it’s not strict to employ these two words with the same meaning universally.

According to Contemporary Chinese Dictionary, the character niá n (黏) equals to nián (黏). Furthermore, based on Chinese Poetics Dictionary, Fu, Xu, & Wang (1999) expound that the character nián (黏) and nián (黏)are the same words. So these two words are interchangeable.

According to the copra of Center for Chinese Linguistics PKU (CCL), the use frequency of “niántǔ 粘土” is much higher than “niántǔ 黏土”; thus, this paper suggests that the interpretation in the sixth, seventh version can be changed to “clay (niántǔ 粘土)”.

(4) yīxiàn (一线)

Although the interpretation in the fifth version is complicated, the explanation is much more clear and comprehensive. On the contrary, the interpretation in the other two versions can’t explain “yīxiàn(一线)”effectively with the word “forefront”.

(5) zhèn (镇)

According to Contemporary Chinese Dictionary, “trading center” means “market town or town”, we can see that “trading center” is a much more widen concept than “town”. However, the interpretation of “market or market town “and “trading center or town” in the sixth, seventh version are extremely cumbersome. To avoid confusion, this paper suggests to keep the interpretation in the fifth version or change it to “market town or town”
5. The comparison of Homonyms between Contemporary Chinese Dictionary and other Dictionaries and the research in teaching Homonyms

Dictionary of Homonyms • Foreword (Zhou 1995) defines Homographs in both broad and narrow aspects: In broad sense, homograph refers to a series of words with the same written form, no matter the pronunciation of these two words are same or not. However, in narrow sense, homograph means two words with same written form but in different pronunciation.”. Although this book already fully considered the two aspects in broad and narrow senses, but it still didn’t been recognized by academia. The concept of homograph has been discussed for a few decades by many scholars, but they still didn’t reach unanimous agreement.

CCD is a Chinese Dictionary, mainly targeted for native speakers. However, some other dictionaries, such as A Dictionary of Chinese Usage:8000 words, A Learner’s Chinese Dictionary–illustrations of the usages, A Dictionary of HSK, The Commercial Press Learner’s Dictionary of Contemporary Chinese. All of these dictionaries are mostly used by second language learners. Different target groups mainly decide the orientation and keystone of the dictionaries, especially in terms of some special types of vocabulary, for example, Homonyms may be presented differently in various dictionaries.

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>gǒng (拱)</th>
<th>mò (墨)</th>
<th>wǎ (瓦)</th>
<th>yīxiàn (一线)</th>
<th>zhèn (镇)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contemporary Chinese Dictionary</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>A Dictionary of Chinese Usage:8000 words</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>A Learner’s Chinese Dictionary – illustrations of the usages</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NULL</td>
<td>NO</td>
</tr>
<tr>
<td>A Dictionary of HSK</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NULL</td>
<td>NO</td>
</tr>
<tr>
<td>The Commercial Press Learner’s Dictionary of Contemporary Chinese</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NULL</td>
<td>NULL</td>
</tr>
</tbody>
</table>

This paper retrieves five homographs separately in four dictionaries, including A Dictionary of Chinese Usage:8000 words, A Learner’s Chinese Dictionary – illustrations of the usages, A Dictionary of HSK, and The Commercial Press Learner’s Dictionary of Contemporary Chinese.( See the appendix 1)
According to the comparison, there are considerably different interpretations of these five homographs in the five dictionaries. On the whole, two kinds of trend are presented. (See Table 4). (1) A word is regarded as “Homograph” in CCD but as “Polysemy” in other four dictionaries. For instance, the other four dictionaries regard “gǒng (拱)” as “Polysemy”. (2) Some of the homographs are used less frequently in *Contemporary Chinese Dictionary*, and the other four dictionaries may delete these usages. For example, there isn’t any interpretations of “mò (墨 ²)” in the four dictionaries; there isn’t any interpretations of “wǎ (瓦 ²)” in the three dictionaries; Three dictionaries present the interpretation of “zhèn (镇)” but without any explanation of “zhèn (镇 ²)”

Table 4 Processing in different dictionaries

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>gǒng(拱)</td>
<td>Homograph</td>
<td>Polysemy</td>
<td>Polysemy</td>
<td>Polysemy</td>
<td>Polysemy</td>
</tr>
<tr>
<td>mò(墨)</td>
<td>Homograph</td>
<td>NO mò (墨 ²)</td>
<td>NO mò (墨 ²)</td>
<td>NO mò (墨 ²)</td>
<td>NO mò (墨 ²)</td>
</tr>
<tr>
<td>wǎ(瓦)</td>
<td>Homograph</td>
<td>NO wǎ (瓦 ²)</td>
<td>NO wǎ (瓦 ²)</td>
<td>Polysemy</td>
<td>NO wǎ (瓦 ²)</td>
</tr>
<tr>
<td>yīxiàn(一线)</td>
<td>Homograph</td>
<td>NULL</td>
<td>NULL</td>
<td>NULL</td>
<td>NULL</td>
</tr>
<tr>
<td>zhèn(镇)</td>
<td>Homograph</td>
<td>NO zhèn (镇 ²)</td>
<td>NO zhèn (镇 ²)</td>
<td>NO zhèn (镇 ²)</td>
<td>NULL</td>
</tr>
</tbody>
</table>

It is not difficult to find that compilers take the users into consideration to a greater extend. For some dictionaries which show cater for native speakers, the compilers try to refine the distinctions between homonyms and give detailed explanations to meet the deep-seated exploration of the native speakers. For the dictionary which focuses on second language users, the compilers simplify the classification to the greatest extent, perhaps because it is not urgent to take into account the need for a second language user to distinguish between homonyms and polysemy, the excessive input of these details may deepen the user’s confusion and reduce their learning interest and efficiency.

The problem of the homonym is not only embodied in various dictionaries, but also in teaching Chinese as a foreign language. Even if the homonym occupies a very small part of the modern lexical system, it still has its own important value. In the process of teaching Chinese as a foreign language, the phonetic identity of the homonym brings some pressure and distress to both of the teachers and students. On one hand, it is difficult for the teachers to fully explain the meaning of various
associations. On the other hand, it may cause difficulty in comprehension for the students without cultural background and language environment. Meanwhile, due to the small number of homonyms and the relatively low utilization rate, it usually leads to the ignorance of homonyms and the systematic study in teaching Chinese as a foreign language. Therefore, the particularity of homonym itself and the ignorance of teaching make homonym become a heavy difficulty point in teaching Chinese as a foreign language.

6. Sense Ordering for Homonym

6.1 Methods of Sorting Interpretations

The process of giving interpretations to vocabulary is complex and the arrangement of their meanings has always been a very important but difficult problem. The function of the dictionary is to help learners get the most knowledge in the most convenient way. The arrangement of the interpretations not only has the significant influence on dictionary users, but also on exploring the arrangement of interpretations. Interpretations of the polysemous word can be arranged in the following three ways: the frequency order of meanings, sequential logic relationship, and the sequential semantic development (Wang & Wu 2015).

6.2 Sense Arrangement Tagging

Since the 1980s, as a spanking way to explore the language, the corpus has been widely recognized and fully flourished. The corpus in Center for Chinese Linguistics PKU is a significant basis to Corpora Research, which “has been approved as the humanities and social science research base of National Universities by the Ministry of Education”, so this paper selects some related corpora from CCL for analyses.

This paper randomly chooses a part of homonymous corpora from CCL to conduct annotation. In order to ensure the scientificity of the corpora, some invalid ones, such as the repeated paragraphs, some place names, name of articles, and interpretations in dictionary, were removed from this paper. Simultaneously, to ensure a sufficient amount of data of corpus, the ratio of the number of final selected corpus to the number of its interpretation is not less than 50:1. For example, there are 7 interpretations of “gǒng (拱)”. Hence, this paper has to choose over 350 valid corpora.

6.3 Frequency of Sense Statistics

(1) gǒng (拱)
As a senior homograph, the two words “gǒng (拱)” have 7 senses, while there are 6490 search results in CCL Corpus. So this paper chooses 550 corpora from CCL and conduct annotation and sequencing towards their senses.

Among the 550 corpora, there are only 372 valid sentences. According to the result of annotation, the number and the frequency about sense of “gǒng (拱)” will be shown in Table 5.

**Table 5** Frequency of Sense in “gǒng(拱)”

<table>
<thead>
<tr>
<th>Word</th>
<th>Arrangement of interpretations in the seventh “Contemporary Chinese Dictionary”</th>
<th>CCL Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>gǒng拱¹</td>
<td>① cup one hand in the other before the chest (in salutation)</td>
<td>118 31.7%</td>
</tr>
<tr>
<td></td>
<td>② surround</td>
<td>53 14.3%</td>
</tr>
<tr>
<td></td>
<td>③ V hunch; arch</td>
<td>25  6.7%</td>
</tr>
<tr>
<td></td>
<td>④ arched, arched part of buildings</td>
<td>141 37.9%</td>
</tr>
<tr>
<td></td>
<td>⑤ N (Gǒng) a surname</td>
<td>12  3.2%</td>
</tr>
<tr>
<td>gǒng拱²</td>
<td>① V push with one’s body</td>
<td>21  5.7%</td>
</tr>
<tr>
<td></td>
<td>② V sprout up through the earth</td>
<td>2   0.5%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>372 100.0%</td>
</tr>
</tbody>
</table>

As illustrated in the figure 5, the number showed a decreasing tendency in both the interpretation ①、②、③、⑤ of “gǒng拱¹”, and the interpretation ①、② of “gǒng拱²”. The arrangement of sense of “gǒng拱” in *Contemporary Chinese Dictionary* keeps consistent with its frequency in CCL Corpus.

However, the interpretation ④, which has been changed in the sixth and seventh version of *Contemporary Chinese Dictionary*, shows a high frequency with 141 times in CCL Corpus which ranks the first. According to the principle of concord, this paper suggests to change this interpretation ④ to interpretation ①.

(2) mò (墨)

Senior homographs “mò (墨)” show 11 interpretations in *Contemporary Chinese Dictionary*, there are 38326 search results in CCL Corpus. This paper chooses 1139 corpora at random, and tag the 550 valid corpora. The results are shown in Table 6.
Table 6 Frequency of Sense in “mò (墨)”

<table>
<thead>
<tr>
<th>Word</th>
<th>CCL Corpus</th>
<th>Number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>mò 墨 ①</td>
<td>China (or Chinese) ink; ink stick; black stick made of coal soot or burned pine soot for writing, painting, etc. occasionally of other materials for different colours.</td>
<td>83</td>
<td>15.09%</td>
</tr>
<tr>
<td>mò 墨 ②</td>
<td>(in a broad sense) pigment; ink</td>
<td>106</td>
<td>19.27%</td>
</tr>
<tr>
<td>mò 墨 ③</td>
<td>handwriting or painting</td>
<td>52</td>
<td>9.45%</td>
</tr>
<tr>
<td>mò 墨 ④</td>
<td>&lt; indication&gt; learning or the ability to read: uneducated</td>
<td>7</td>
<td>1.27%</td>
</tr>
<tr>
<td>mò 墨 ⑤</td>
<td>carpenter’s line marker; rules and regulations</td>
<td>43</td>
<td>7.82%</td>
</tr>
<tr>
<td>mò 墨 ⑥</td>
<td>black: pitch-dark</td>
<td>53</td>
<td>9.64%</td>
</tr>
<tr>
<td>mò 墨 ⑦</td>
<td>&lt; fml&gt; corruption; graft; embezzlement</td>
<td>1</td>
<td>0.18%</td>
</tr>
<tr>
<td>mò 墨 ⑧</td>
<td>tattooing the face and painting it black (a punishment in ancient China); also qíng 黥</td>
<td>2</td>
<td>0.36%</td>
</tr>
<tr>
<td>mò 墨 ⑨</td>
<td>(Mò) Mohist school</td>
<td>36</td>
<td>6.55%</td>
</tr>
<tr>
<td>mò 墨 ⑩</td>
<td>N (Mò) a surname</td>
<td>75</td>
<td>13.64%</td>
</tr>
<tr>
<td>mò 墨 ⑪</td>
<td>Mexico</td>
<td>92</td>
<td>16.73%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>550</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

According to the annotations, we could find that the arrangement of sense in frequency order isn’t consistent of the existed arrangement.

Firstly, the arrangement of sense in frequency order is: ② > ① > ⑥ > ③ > ⑤ > ④ > ⑧ > ⑦. It’s obviously that the arrangement is discrepant from each other. However, due to the nuance between interpretations ①, ② in some corpora, it’s hard to identify all of them very clearly. Interpretations ⑦, ⑧ are not used frequently in modern Chinese, and they are in the minority. Sampling annotations can’t represent the difference comprehensively; thus we can ignore the gap of frequency between these two senses. The frequency of items ③⑤⑥ is much higher than the item ④. According to the principle of sense arrangement, it’s better to switch over them to meet the needs of readers.

Secondly, the frequency of interpretations ⑨⑩ is relatively high. However, according to the sense annotations, it is found that the collocation with these two senses remains extremely simple. Almost all of the corpora show the results: Mohists, Mozi, and Modi. In view of their simple structures, this paper suggests to keep these two senses in the original position even though they get high frequency.

Thirdly, despite the high frequency of “mò 墨 ⑪”, it should be placed to the end since it’s a proper noun.

(3) “wǎ 瓦”
In *Contemporary Chinese Dictionary*, there are 4 senses in the senior homonym “wǎ 瓦”. And we find 74746 search results from CCL Corpus, among which the 688 corpora were selected and we conduct annotation toward the retained 203 valid corpora. The statistical results are shown in Table 7 below:

<table>
<thead>
<tr>
<th>Word</th>
<th>Arrangement of interpretations in the seventh “Contemporary Chinese Dictionary”</th>
<th>CCL Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Number</td>
</tr>
<tr>
<td>wǎ 瓦¹</td>
<td>① N tile; vaulted, flat or semi-cylindrical slab of hard material such as baked clay or cement, laid in rows to cover roofs</td>
<td>124</td>
</tr>
<tr>
<td></td>
<td>② clay</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>③ N Wǎ's surname</td>
<td>15</td>
</tr>
<tr>
<td>wǎ 瓦²</td>
<td>Abbr. for 瓦特 wǎ tè</td>
<td>21</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>203</strong></td>
</tr>
</tbody>
</table>

As the sense annotations showed, the frequency of sense annotations coincides with the substantially arrangement.

(4)“yīxiàn(一线)”

Homographs " yīxiàn(一线)" has 3 interpretations, and 11760 search results in CCL Corpus. This paper randomly selected 180 valid corpora to tag from the whole 223 corpora. The statistical results are shown in Table 8 below:

<table>
<thead>
<tr>
<th>Word</th>
<th>Arrangement of interpretations in the seventh <em>Contemporary Chinese Dictionary</em></th>
<th>CCL Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Number</td>
</tr>
<tr>
<td>yīxiàn 一线¹</td>
<td>① war front</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>② forefront</td>
<td>70</td>
</tr>
<tr>
<td>yīxiàn 一线²</td>
<td>A ray of; a gleam of</td>
<td>77</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>180</strong></td>
</tr>
</tbody>
</table>

The results are shown in Table 8, the gap of the use frequency between those two interpretations in the text is up to 20%. In the light of arrangement of meanings, this
paper suggests to switch over these two senses to help the users to look up common sense more conveniently.

(5) “zhèn(镇)”

There are 10 interpretations in the senior homographs “zhèn(镇)”, and the search results in CCL Corpus is 116519. This paper chooses 1469 corpora at random, and annotates the valid 652 ones. Statistics are shown in Table 9 below.

<table>
<thead>
<tr>
<th>Word</th>
<th>Arrangement of interpretations in the seventh “Contemporary Chinese Dictionary”</th>
<th>CCL Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>zhèn 镇</td>
<td>① V press down; keep down; case 21 3.22%</td>
<td>Number</td>
</tr>
<tr>
<td></td>
<td>② calm 68 10.43%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>③ keep pace by force; garrison 162 24.85%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>④ garrison post 26 3.99%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>⑤ N town; administrative division generally under the jurisdiction of a country 240 36.81%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>⑥ N comparatively large trading center or town 129 19.79%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>⑦ N cool sth. with cold water or ice 5 0.77%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>⑧ (Zhèn) N a surname 0 0.00%</td>
<td></td>
</tr>
<tr>
<td>zhèn 镇</td>
<td>① adv (oft. used in the early vernacular)often; frequently 0 0.00%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>② (oft. used in the early vernacular)whole period 1 0.15%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>652 100.00%</td>
<td></td>
</tr>
</tbody>
</table>

It seems that the tagging frequency of “zhèn(镇)” didn’t follow the ordering principle, but in fact it is also reasonable.

We can divide the interpretations of “zhèn(镇)” into four semantic categories: First kind: the interpretations from ①-④ can be classified as “press down; garrison”, accounting for 42.49%; The second kind: the interpretations ⑤⑥ should be “village; town”, accounting for 56.6%; The third: the interpretation ⑦ “cool sth” and the interpretation ⑧ “a surname”.

In the view of semantic categories, this paper suggests to change Category 2 to the front. It’s obviously that the interpretation ⑤ marks a higher frequency than others, so it’s better to switch it to be the interpretation ①. The sense arrangement in every semantic category could follow the principle of sorting by frequency, for example, as for the interpretation ④ in the Category 1, there is a big difference between the annotation sequence and the interpretations. In accordance with an
ordering principle in valuable meanings, the paper suggests item ④ should be rearranged based on its frequency, which will be more reasonable. Although we didn’t find any corpora related to the interpretation “a surname” in category 4, we still suggest to keep it, because we have to take into account the limited corpora which may lead to some error.

This paper suggests to adjust the following order according to the arrangement in semantic categories.

**Table 10** Arrangement of Sense According to Linguistic Group

<table>
<thead>
<tr>
<th>New Arrangement</th>
<th>Arrangement of interpretations in the seventh “Contemporary Chinese Dictionary”</th>
<th>Number</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>⑤ <strong>N</strong> town; administrative division generally under the jurisdiction of a country</td>
<td>240</td>
<td>36.81%</td>
</tr>
<tr>
<td>2</td>
<td>⑥ <strong>N</strong> comparatively large trading center or town</td>
<td>129</td>
<td>19.79%</td>
</tr>
<tr>
<td>3</td>
<td>③ keep pace by force; garrison</td>
<td>162</td>
<td>24.85%</td>
</tr>
<tr>
<td>4</td>
<td>② calm</td>
<td>68</td>
<td>10.43%</td>
</tr>
<tr>
<td>5</td>
<td>④ garrison post</td>
<td>26</td>
<td>3.99%</td>
</tr>
<tr>
<td>6</td>
<td>① <strong>V</strong> press down; keep down; case</td>
<td>21</td>
<td>3.22%</td>
</tr>
<tr>
<td>7</td>
<td>⑦ <strong>V</strong> cool sth. with cold water or ice</td>
<td>5</td>
<td>0.77%</td>
</tr>
<tr>
<td>8</td>
<td>⑧ (Zhèn) <strong>N</strong> a surname</td>
<td>0</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

The arrangement of senses in the four dictionaries keep consistent with that in *Contemporary Chinese Dictionary*. With the help of sense annotation, this paper finds that there are some differences between the sense arrangement among the five reference dictionaries and the results of the frequency statistics in Corpus. On the one hand, in accordance with the sorting sense in frequency, it’s convenient for second language learners to know the use of frequency, and to look up the common sense directly which will enhance the effect of learning. Some English dictionaries, such as *Random House Dictionary*, mainly adopts this kind way in sorting arrangement. (Hu&Wang 2012). On the other hand, if there is a strong semantic link between different senses, we can also arrange sense according to semantic category, and sort the interpretations according to its frequency within each semantic category.
7. Conclusion

Homographs are the two words with the same written form but with different meanings, which have their specific difficulties in language teaching. This paper chooses some advanced vocabularies from *The Graded Syllable, Characters and words for the Application of Teaching Chinese to the Speakers of Other Languages* (the Research Group of the Graded Syllable, Characters and words for the Application of Teaching Chinese to the Speakers of Other Languages 2010), and pays close attention to the 5 vocabularies which are homographs in *Contemporary Chinese Dictionary*. First of all, this paper compares the transformation of the interpretations of the 5 words in the fifth, sixth, seventh version of *Contemporary Chinese Dictionary*. Secondly, this paper makes further explorations of the 5 words in *A Dictionary of Chinese Usage: 8000 words* (Center for Chinese Linguistics PKU 2000), *A Learner’s Chinese Dictionary — illustrations of the usages* (Liu & Chen 2005); *A Dictionary of HSK* (Shao 2000); *The Commercial Press Learner’s Dictionary of Contemporary Chinese* (Lu & Lv 2006). After comparing with the interpretations in *Contemporary Chinese Dictionary*, it’s clear to see that the interpretations show great difference in different learner’s dictionaries. Thirdly, this paper takes advantage of CCL Corpus to annotate the sense and count the frequency. With the combination of the arrangement of sense in five dictionaries, this article figure outs the importance of Semantic Information Frequency in a learner’s dictionary, and provides two methods in sense arrangement, such as, Semantic Information Frequency and semantic category. This study also has great significant influence on the in-depth discussions in how to teach homographs in teaching Chinese as a second language.

Acknowledgements

This project is sponsored by Knowledge Transfer Funds of The Education University of Hong Kong. The project number is KT_2016-2017- 0012.

Reference


The Research Group of “The Graded Syllable, Characters and words for the Application of Teaching Chinese to the Speakers of Other Languages” M 2010.
The graded Syllable, characters and words for the application of teaching Chinese to the speakers of other languages. Beijing: Beijing Language University Press.


Xu, Shen. 100-121. Interpretation of Chinese characters.


### 拱:

<table>
<thead>
<tr>
<th>释义</th>
<th>词典</th>
</tr>
</thead>
<tbody>
<tr>
<td>①两手相合，臂的前部上举：～手</td>
<td>《现代汉语词典》2016年第7版</td>
</tr>
<tr>
<td>②环绕：～卫；～众星～月～四山环～的大湖。</td>
<td></td>
</tr>
<tr>
<td>③肢体弯曲成弧形：～肩缩背；～嘴～腰。</td>
<td></td>
</tr>
<tr>
<td>④成弧形的建筑物，也指建筑物成弧形的部分：～式门楼～桥～连～坝。</td>
<td></td>
</tr>
<tr>
<td>⑤(Gǒng)姓。</td>
<td></td>
</tr>
<tr>
<td>①用身体撞动别的东西或拨开土地等物体：用身子～开了大门；猪用嘴～地；蚯蚓从地下～出许多土来；一个小儿从人群里～出去了。</td>
<td>《HSK中国汉语水平考试词汇8000词词典》</td>
</tr>
<tr>
<td>②植物生长，从土里向外钻或顶：～出来了。</td>
<td></td>
</tr>
<tr>
<td>①两手相合，臂的前部上举：～手～手礼；～手向大家～了～手。</td>
<td>《学汉语用例词典》</td>
</tr>
<tr>
<td>②环绕：～卫；～众星～月～天津、保定、张家口是北京的～卫城市；记者们像群星～月一样，簇拥着他走进大厅。</td>
<td></td>
</tr>
<tr>
<td>③肢体弯曲成弧形：～背～嘴～腰；～肩缩背；他～着嘴，满脸的不高兴；小花猫～了～腰，从桌子上跳了下来。</td>
<td></td>
</tr>
<tr>
<td>①用身体撞动别的东西或拨开土地等物体：～门；门被～开了；他双手抱着一摞书，用肩膀～开了大门；小孩儿从人群里～了出来；蚯蚓从地下～出许多土来。</td>
<td>《汉语水平考试词典》</td>
</tr>
<tr>
<td>②植物生长，从土里向外钻或顶：～出来，～不出来；～出一寸高了；麦苗～出土来了。</td>
<td></td>
</tr>
</tbody>
</table>

【近义词】④⑤顶
【构词】拱别/拱火/拱立/拱门/拱桥/拱手/拱卫/拱肩缩背

【义１】用身体把别的东西撞开或推开(push up or forward with one's body; (of pigs, etc.) dig earth with the snout; (of earthworms, etc.) wriggle through the earth): 小孩儿用屁股把门～开了；老母猪用嘴巴～地。
【义２】两手相合；环绕(cup one hand in the other before the chest; surround; arch):～抱～门～桥～这个冠军我们是～手相～的～大家围着她就像众星～月似的。

【动】①用身体顶开或推开(别的东西):他用力拱开了门；这个孩子使劲从人群中拱了出去。②动物用嘴顶开或推开(别的东西):小猪在地上拱出了
一个坑里小狗用嘴拱开门跑去了一。
【义 1】铺屋顶用的建筑材料，一般用泥土烧成的，也有用水泥等材料制成的，形状有拱形的、平的或半个圆筒形的等。
【义 2】用黏土烧成的：～盆 | ～罐 | ～器。
【义 3】(Wǎ)姓。
瓦 2 Wǎ
瓦特的简称。1 秒钟做 1 焦的功，功率就是 1 瓦。

### 词汇
- 瓦房
- 瓦匠
- 瓦砾
- 瓦垄
- 瓦舍
- 瓦解冰消
- 瓦盆
- 瓦罐
- 瓦解

### 用例
- 房顶用的建筑材料，一般用泥土烧成的，也有用水泥等材料制成的，形状有拱形的、平的或半个圆筒形的等。
- 砖瓦匠
- 瓦片
- 瓦房
- 瓦解
- 烧瓦
- 盖瓦
- 瓦砾
- 瓦解
- 瓦解

### 一钱
- 架线
- 瓦片
- 瓦解
- 瓦砾
- 瓦解
- 瓦解

### 一线
1. 战争的最前线。
2. 第一线：深入车间慰问～工人。
3. 数量词，形容极其细微：～阳光 | ～光明 | ～希望 | ～生机。
### 镇:

<table>
<thead>
<tr>
<th>词典</th>
<th>释义</th>
</tr>
</thead>
</table>
| 《现代汉语词典》2016年第7版 | ①动压；抑制：～纸|～痛|他一说话，就把大家给～住了。  
②安定：～静|～定。  
③用武力维持安定：～守|坐～。  
④镇守的地方：军事重～。  
⑤各行政区划单位，一般由县一级领导。  
⑥各较大的市集或集镇  
⑦动把食物、饮料等同冰块放在一块儿或放在冷水里，冰箱里使凉：冰～汽水|把西瓜放在冷水里～一～。  
⑧(Zhèn)名姓。 |
| 镇2 | ①副  
时常(多见于早期白话，下同)：十年～相随。  
②表示整个的一段时间：～日(整天)。 |
| 《HSK中国汉语水平考试词汇大纲汉语8000词词典》 | ①较大的市集：集～|我们去～上买些东西|这里有个很大的集～，各种吃的都有。  
②行政区划单位，一般由县一级领导：～长|～子|城～|我们～的～长很年轻|这个～子不大，但很干净|我们县共有五个～。  
【提示】镇，姓。【构词】镇尺/镇服/镇静剂/镇市/镇守/镇纸/镇子  
①压，抑制：～纸|～痛|你拿这块石头把地图～住，别让风刮跑了|这种药可以～痛，但不能从根本上治你的病|把冷毛巾放在头上|②用武力维持安定：～守|服|他被头儿的威严～住了|这支军队～守在祖国的边疆|有你在这儿～着，敌人不敢来。③把食物、饮料等同冰块放在一起或放在冷水里使凉：冰～汽水|冰～啤酒|把西瓜放在凉水里～一。 |
| 《学汉语用例词典》 | (名)行政区划单位，一般由县一级领导。 (town, an administrative division generally under the jurisdiction of a county) 常做主语、宾语、定语。[量]个  
例句  
附近几个镇近几年变化都很大|我们镇不太大。|这里原来只是一个无名小镇，现在成了远近闻名的旅游景点。|近几年乡镇企业发展很快。|镇服装厂每年都能赚很多外汇。  
(动)①压，抑制。(press down; keep down; ease)常做谓语。
### 例句
他一说话，就把大伙儿给镇住了。 | 这个孩子不听话，你给我镇他。
安定:用武力维持安定。 (calm; guard; garrison) 常用于构词语：
镇定镇静镇守镇例句：总司令亲自坐镇指挥。
把食物、饮料等同冰块或冷水放在一起，或放入冰箱使变凉。
(cool with cold water or ice) 常做谓语、定语。
例句：把西瓜放在冰箱里镇一镇。 | 我喜欢喝冰镇啤酒。

<table>
<thead>
<tr>
<th>《汉语水平考试词典》</th>
<th>镇丙（名）</th>
</tr>
</thead>
</table>
| 【义 1】行政管理上的单位，在村与县之间，也泛指农村人口集中
的居民点 (town; garrison post): 村～|市～|乡～|各～～长都到县里开会去了 |
周庄是个古朴的江南小～。

<table>
<thead>
<tr>
<th>《商务馆学汉语词典》</th>
<th>镇丁（动）</th>
</tr>
</thead>
</table>
| 【义 2】压住；抑制 (ease; press down; guard): 这种药有～痛的作用。
【义 3】～定: 使安静 (calm; tranquil): ～定！妈妈显得格外～静。
【义 4】把食物、饮料等同冰块、
冷水放在一起，或在冰箱里使温度降低 (chill; cool with cold water or ice): 冰～汽水 |
把西瓜冰～一下。 | 无 |
A Chinese-English Dictionary for Production: With “有效” as an Example

Anmin Wang
School of Foreign Languages, Lanzhou Jiaotong University
am_w@mail.lzjtu.cn

Abstract

With the headword “有效” in Chinese-English (henceforth C-E) dictionaries targeting Chinese translators or English learners as an example, the article looks into the potential ways to improve its treatment to better meet their productive needs. Five leading medium or large C-E dictionaries claiming to serve this very purpose seem to have adopted more or less the same way of dealing with it: listing out several “equivalents” after the headword, then following them with illustrative examples and their corresponding English translations. The translations of these examples mainly focus on illustrating the usage of the equivalents provided. In the meantime, the free translations for the headword in context, from which an equivalent can not be extracted, are generally not enough. Hence it can be argued that these equivalents, examples and translations offered have not fully aided dictionary users in their output. Without differentiating the fine differences between the equivalents in terms of aspects such as collocation, semantics, stylistics, and seeking for effective ways to highlight such differences, dictionary users may confuse their usages or not be fully aware how to put these equivalents into correct use. Various labels can thus be added for the equivalents to better meet users’ productive needs. The examples and their translations can be further improved by increasing the number of examples, balancing phrase and sentence ones, improving the illustrating quality, and including more freely translated headwords in context. Chinese-English parallel corpora can be used for this end as well. Productive C-E dictionaries can be more user-friendly in terms of providing necessary aid for productive purpose.

Keywords: “有效”, Chinese-English dictionary, production, Chinese-English parallel corpus, free translation

Introduction

The entry structure of many current productive C-E dictionaries is almost the same, with English equivalents following a Chinese headword. However, those equivalents
are simply listed out, without further explanations of their differences or similarities in terms of semantics, syntax, pragmatics, and so on. The illustrative examples of a headword and their translations, including its free translations in different contexts, appear after the equivalents. However, most translations of examples just serve to illustrate the usage of different equivalents. The total free translations of a headword are relatively few.

The C-E dictionaries for native Chinese serve the productive purpose (e.g. Wang 1998: 40; Li 2009: 65), namely C-E translation, English writing, etc. Therefore, the meaning of the equivalents needs to comply with that of a Chinese headword and they should also be standard English usages (Pan 2003: 89). Through equivalents and the translated illustrative examples, C-E dictionaries also need to “show the grammatical, [semantic, pragmatic, etc] features of English expressions and save the users the trouble of looking up their usages further in monolingual English dictionaries or English-Chinese ones” (Zhao 2006: 97-98). This implies that translating meaning in C-E dictionaries needs to be further improved. Chinese-English parallel corpora can help attain this goal, especially in terms of enhancing the quality of the freely translated illustrative examples, which will benefit dictionary users’ production in English.

With the treatment of “yǒuxiào 有效” in productive C-E dictionaries as an example, the paper aims to look into the features of supplying equivalents, examples and their translations in such dictionaries. The paper is comprised of four sections. The first one focuses on the definition of “yǒuxiào 有效” and how it is treated in five representative large or medium productive C-E dictionaries. The common features of those dictionaries in treating it are summarized in the second section. In the third section, some principles for providing equivalents and the translation of examples in such dictionaries are put forth, followed by the conclusion as the last section.

1. The Definition of “yǒuxiào 有效” and Its Treatment in Five C-E Dictionaries

In Contemporary Chinese dictionary (5th edition) (2012: 1581), “yǒuxiào 有效” is marked as a verb. It is defined as “能实现预期目的；有效果(can realize the desired goals; have effects—my translation)”, followed by examples: ～措施|这个方法果然～ (effective measures|The method turns out to be effective—my translation). For native Chinese, the definition and the examples suffice to enable us to understand the meaning and the usage of this word, even if whether the part of speech of “yǒuxiào 有
"有效" in the first example is a verb or adjective may be open to doubt.

The latest version of five representative large or medium productive C-E dictionaries has been selected. They are *New Age Chinese-English Dictionary* (the 2nd edition), *New Century Chinese-English Dictionary* (the 2nd edition), *A Chinese-English Dictionary* (the 3rd edition), *ABC Chinese-English Comprehensive Dictionary* (2nd edition), and *The Chinese-English Dictionary* (the 3rd edition) (henceforth abbreviated as *New Age*, *New Century*, *ABC*, *A Chinese-English*, and *The Chinese-English*, with their editions omitted). If not specified otherwise, the editions of the five dictionaries are the same as have been mentioned above. How “yǒuxiào 有效” has been treated in those five dictionaries will be analyzed below.

As for the part of speech of “yǒuxiào 有效”, among the five dictionaries above, it is marked or interpreted as an adjective in four, and a verb in just one. In *New Century* (Du 2016: 2123), *A Chinese-English* (Yao 2010: 1712), *ABC* (DeFrancis 2003: 1190), it is marked as an adjective. The compilers of *The Chinese-English* did not provide its part of speech. However, based on that of its translated equivalents, the compilers have interpreted it as an adjective as well. “yǒuxiào 有效” is labeled as a verb in *New Age*, just the same as that provided in *Contemporary Chinese Dictionary*. Thus it is understandable that in *New Century*, *A Chinese-English*, *The Chinese-English*, and *ABC*, the equivalents are all adjectives. By contrast, the compilers have offered six equivalent expressions for it in *New Age*, all of which are “verb+object” or “link verb+predicative”.

The translated equivalents in five C-E dictionaries have a lot in common. Those in *New Century*, *A Chinese-English* and *The Chinese-English* are the same: “efficacious, effective, valid”. Apart from the three words, “available” has been supplemented in *ABC*. The English equivalents in *New Age* for “yǒuxiào 有效” are as follows: “produce/have an effect; be efficacious/effective/authentic/valid”. Except “authentic”, the adjectives working as predicates are the same as the three equivalents above. The equivalents of “yǒuxiào 有效” in the five dictionaries are summarized in Table 1.1.

| Table 1.1 Translation equivalents provided by compilers in five C-E Dictionaries |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Total equivalents              | 4               | 3                | 3                | 3               | 6               |
| Equivalents                    | efficacious, effective, valid | efficacious, effective, valid | efficacious, effective, valid | efficacious, effective, valid | produce/have an effect; be efficacious |
Nevertheless, for the benefits of dictionary users, the compilers of productive C-E dictionaries “cannot simply list isolated equivalents which lack empirical basis, [since] whether they are native or correct is still open to doubt” (Liu & Li 2012: 21). Suppose those equivalents are different in semantics, syntax, pragmatics, etc., they should be labeled correspondingly. This could help relieve users of further consulting English-Chinese dictionaries (Hui 2004: 1) or monolingual English dictionaries for the usage of those equivalents.

Apart from the equivalents in productive C-E dictionaries, the illustrative examples and their translations are also conducive to dictionary users’ production. If the equivalents have differences in the aspects mentioned above, compilers can also show the differences through the translations of examples. In other words, C-E dictionary compilers need to provide suitable examples and their translations to help users output. The translation of examples should not be confined by equivalents. It can show “different expressions in the target language in different collocations, contexts and situations” (Zhao and Jiang 2014: 12) or the translations of a headword in different contexts to dictionary users for reference. The examples and their translations in five C-E dictionaries are presented in Table 1.2. The details of these illustrative examples in five C-E dictionaries and their English translations have been attached at the end of the paper as Appendix 1.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total examples</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>Total translations illustrating the usage of headwords</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>Free translations of headwords</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1.2 Classification of illustrative examples and their translations in five C-E dictionaries

In analyzing the illustrative examples of “yǒuxiào 有效” and their translations, if the translations just illustrate the usage of an equivalent, the equivalent is underlined. This is to differentiate such translations from its free translation in contexts. (Refer to

1.Here examples only refer to those in which “yǒuxiào 有效” is not used as a term.
Appendix 1) With the latter as the criterion, no such translations have been offered in ABC and The Chinese-English. Only one has been provided in A Chinese-English and New Age, accounting for 1/6 of its total examples. Six such translations have been provided in New Century, constituting 1/3 of the translation of total examples for “yǒuxiào 有效”. It ranks No. 1 among the five dictionaries in providing free translations of “yǒuxiào 有效”, representing its compiling concept of “Putting users first, Focusing on practicality” (Hui 2004: 1).

2. Similarities in Five C-E Dictionaries in Treating “yǒuxiào 有效”

The five dictionaries above share some features in treating “yǒuxiào 有效” in terms of providing equivalents, and giving examples and their translations.

2.1 Only Listing Its Equivalents, with Their Semantic, Stylistic, Register and Disciplinary Features Unspecified.

To begin with, the semantic scope of the English equivalents of “yǒuxiào 有效” is not the same, as is shown in Figure 1.1.

![Diagram](attachment:image.png)

**Figure 1.1** The semantic relationship between the equivalents of “yǒuxiào 有效”

As Figure 1.1 indicates, in conveying the meaning of “yǒuxiào 有效”, the semantic scope of five equivalents, namely, “effective”, “authentic”, “available”, “efficacious” and “valid”, partly converges. The last oval containing elliptical marks stand for its other potential equivalents, whose semantic scope may also overlap with that of other ones to different extent. However, without clearly defining the semantic scope of those translated equivalents, users may take them as absolute synonyms, thus misuse them.
Meanwhile, the stylistic, disciplinary, and other features of those translated equivalents listed are not provided by the compilers of the five C-E dictionaries above. This may lead to their misuse of the equivalents as well. For instance, “be authentic” is included in New Age as one of the equivalent English translation of “yǒuxiào有效”. Nevertheless, the adjective “authentic” has a disciplinary feature, which has not been indicated clearly in English-Chinese dictionaries. In The English-Chinese Dictionary (2nd edition) (Lu 2010: 115), the English word has been labeled to be used in the legal field only. In New Age English-Chinese Dictionary (Hu 2016: 167), aside from offering the same disciplinary label, it is also pointed out before the Chinese equivalents of “yǒuxiào有效” that it collocates with “deed” to convey the Chinese meaning. Besides, when “efficacious” is used to express the meaning of “yǒuxiào有效”, it is a formal usage (Oxford Advanced Learner’s English-Chinese Dictionary 1997: 463). However, it ought to be noted that its Chinese counterpart “yǒuxiào有效” does not have such a stylistic feature. Therefore, such a feature needs to be highlighted for the good of users.

2.2 Examples and Their Translations

2.2.1 The Number of Examples is Generally Inadequate

According to the Table 1.2 above, the number of examples in the five C-E dictionaries is generally not enough. Except the fact that the compilers of New Century has presented 18 illustrative examples, A Chinese-English and New Age have each provided 6 such examples. Only 2 examples have been offered in The Chinese-English. By contrast, no such examples have been provided in ABC. Therefore, it can be argued that overall the number of examples can be further increased to be conducive to dictionary users’ production in English.

2.2.2 Using Phrase or sentence Examples Can be Carefully Weighed

Whether to use phrase or sentence examples can be weighed carefully. In terms of the division between phrase and sentence examples, it can be found that 14 phrase examples and 4 sentence ones have been offered in New Century. The Chinese-English has provided one phrase and one sentence example, while New Age has 2 phrase ones and 4 sentence ones. A Chinese-English has offered 3 phrase and sentence examples each. Therefore, except New Century, A Chinese-English and The Chinese-English have offered equal number of phrase and sentence examples. New Age has provided more phrase examples than sentence ones. Only if a sentence example is indispensable to illustrating the usage of English equivalents or the free translation of the headword in various contexts, otherwise the number of sentence examples can be reduced. Correspondingly, the number of phrase examples can be further increased, to make the
entry more informative to better meet dictionary users’ productive needs.

### 2.2.3 Illustrating Quality Can be Further Improved

It can be argued that the quality of illustrating can be further improved as well. For instance, the translations of some examples are for illustrating the usage of the same English equivalent. Therefore, the examples can be integrated into one. For example, *A Chinese-English* has offered “有效合同” and “有效证件” as two examples. However, the translations of the two Chinese phrases both serve to illustrate the usage of “valid”. Therefore, they can be combined into one. Similarly, *New Century* has provided “有效合同/协议” and “有效护照/证件/身份证” as phrase examples. The two examples can also be integrated into one, since they all show the usage of “valid” as well.

Some translation equivalents need to collocate with certain other words, which have not been highlighted through the translated illustrative examples. For instance, in *New Century*, “依然有效 remain in force/effect” is offered as one example for “yǒu xiào 有效”. Owing to the lack of the information on the subject in the corresponding English translation, dictionary users may put any subject before the two English expressions. However, the subjects that go with them are actually limited. According to *New Century English-Chinese Dictionary* (Hu 2016: 834), “in effect” mainly collocate with “law” and “rule,” while “in force” to with law, rule and system. “available” is listed as one of the equivalents of “yǒu xiào 有效” in ABC. Nevertheless, the subjects go before it ought to be bus/train tickets and play tickets (Lu 2010: 118). Meanwhile, the typical collocation of each equivalent, like “adjective+preposition”, has not been highlighted as well, to call dictionary users’ attention to its usage.

### 2.2.4 Freely Translated “yǒu xiào 有效” in Contexts Can be Further Increased

Among the translation of examples, the free translation of “yǒu xiào 有效” in different contexts can be increased. Most of the translations serve to illustrate the usage of equivalents. The proportion of freely translated headword in various contexts is generally low. As the Table 1.2 shows, even in *New Century*, such kind of examples only account for one-third of its total. Meanwhile, even if some free translations of “yǒu xiào 有效” in contexts are accurate and native, their stylistic features are ignored in those C-E dictionaries. For instance, “hold good” has been included as one free translation for “yǒu xiào 有效” in contexts. However, it should be an informal usage
(Lu 2010: 804). In the same vein, “do the trick” as a free translation of “yǒuxiào 有效” is also informal (Lu 2010: 2169; Hu 2016: 2768).

Therefore, to offer translation equivalents, examples and their translations for “yǒuxiào 有效” in a productive C-E dictionary, the following three principles can be followed. This is to overcome the shared features in the five C-E dictionaries above and be more helpful to dictionary users’ production.

3. Three Principles for Providing Equivalents, Examples and Their Translations in Productive C-E Dictionaries

3.1 Various Features of Equivalents Should be Highlighted

The various features of equivalents in a productive C-E dictionary, such as register, stylistics, semantics, can be emphasized. Such a dictionary aims at helping dictionary users output in English. Many productive C-E dictionaries have marked the part of speech of headwords, which helps to render the meaning of a headword accurately. The part of speech of the translated equivalents of the headwords is usually the same as that of a headword. Since explanatory equivalents are a big obstacle for productive tasks, like translation and writing, such dictionaries ought to provide equivalents directly “insertable”. Nevertheless, if the features of “insertable” equivalents in semantics, style, register, discipline, and so on, are not highlighted or clearly specified, users may misuse them.

The semantic scope of different equivalents of “yǒuxiào 有效” varies. Therefore, we can make the fine difference between them. To distinguish “effective”, “efficacious”, “valid” Shi (2006: 1130), “available” and “authentic”, the following table can be provided. The explanation of their usage is given in Chinese.

<table>
<thead>
<tr>
<th>Equivalent</th>
<th>Chinese Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>available</td>
<td>主要指车票、戏票等有效 [This word mainly refer to bus/train, or play ticket is still valid—my translation]</td>
</tr>
<tr>
<td>authentic</td>
<td>主要指契约有效 [It is mainly used to indicate a deed is still legally valid—my translation]</td>
</tr>
<tr>
<td>effective</td>
<td>泛指达到一切预期效果 [It refers to achieving the desired effects generally—my translation]</td>
</tr>
<tr>
<td>efficacious</td>
<td>主要指药物、治疗等有效果 [It mainly refers to the effects of medicine, treatment, etc.—my translation]</td>
</tr>
</tbody>
</table>
valid  主要指文件、证件、合同、票据等在规定期限或范围内有效

| valid | 主要指文件、证件、合同、票据等在规定期限或范围内有效 | It mainly imply documents, certificates, contracts, tickets, and so on, are effective or usable within a fixed period or scope.—my translation |

For the equivalents with distinctive features in terms of discipline, style, register, and so on, we can adopt two ways to cope with them. One way is list such English words as equivalents. Meanwhile, labels can be added in front of the corresponding equivalents to highlight these features. For example, before “efficacious” the stylistic label of “＜正式＞ (formal—my translation)” can be added. Meanwhile, “authentic” can also be taken as one of the equivalents, with “【律】 (Law—my translation)” added before it. The rules concerning the usages of translation equivalents, if expressed incorrectly or summarized wrongly, will make dictionary users commit errors in their production. We can also highlight these features of discipline, stylistics, register, and so on, by providing explanations in parenthesis in Chinese behind the translation of an example sentence of a headword, which will be illustrated in 3.2.2.

3.2 Examples and Their Translations

3.2.1 The Number of Examples Can be Increased

On the whole, the number of illustrative examples for a headword in a productive C-E dictionary, like “yǒuxiào 有效”, can be increased. The examples ought to have various commonly used Chinese phrases, and sentences showing the usage of a headword, and the length of such sentences should vary (Pan 2003: 91). Even if it is hard to set a premium number for such examples of a headword, it is preferable that such a number can be larger, so as to be helpful to dictionary users’ production.

3.2.2 The Use of Phrase and Sentence Examples Can be Balanced

The phrase and sentence examples in a productive C-E dictionary ought to be balanced. To save space and provide more examples in helping dictionary users output, its compilers can use sentence examples in case that phrase ones may cause problems in users’ understanding (Mei 1999: 43) of an example, prevent their production in English, and so on. Therefore, “依然有效” in New Century ought to be replaced with a sentence example: 这些规则/法律会依然有效。“These rules/laws will remain in force/effect.” “rules” and “laws” appear in bold, to highlight the fact that they are typical collocates of “in force/effect”. We can add another example: 这一系统会依然有效。The system
will remain in force, which serve the same purpose. C-E dictionary compilers can enable its users to derive grammatical, collocational, and other information from examples and their translations. Therefore, if “authentic” is not listed as one of the equivalents, we can add a bracketed explanation after the translation of the example to emphasize its disciplinary feature: 该契约依然有效。The deed is still authentic.（1

该词为法律用语 [This is a legal usage.—my translation]. The example appears as a full sentence to highlight the typical collocation in its translation. When consulting a C-E dictionary, people want to know how an equivalent is used in context and this includes collocation, idioms (Hui 2016: 3), register, stylistic features, and so on. In other words, for C-E dictionary users, the more important thing is when users see the English translation of an example, the translation can enable him to know how to translate a Chinese sentence or write out a correct and native expression. The dictionary compilers have the responsibility to help users to achieve the goal. The balancing of phrase and sentence examples also serves to attain this goal.

3.2.3 The Illustrating quality Can be Improved

The illustrating quality in productive C-E dictionaries can be further enhanced, so as to provide the largest amount of information possible in the limited dictionary space. For example, “有效合同/协议” and “有效护照/证件/身份证” in New Century can be combined into one: “有效合同/协议/护照/证件/身份证” = “valid contract/agreement/passport/documents/identity card”. The translations all illustrate the usage of “valid”. Meanwhile, we can also offer the following phrase examples for “yǒuxiào 有效”: 采取有效方法/手段/步骤/措施 take effective methods/means/steps/measures. All the translation equivalents of Chinese nouns can collocate with “effective”. In this way, the possible collocations of “yǒuxiào 有效” in Chinese, together with their translations, is presented in the least amount of space.

C-E dictionary compilers can help dictionary users to learn the specific differences between different English equivalents (Lu and Tian 2013: 48). For example, if we do not list “available” as one of the equivalents of “yǒuxiào 有效”, the limitation between the subject and verb collocation needs to be highlighted. For this purpose, we can offer the following ticket: 这些车票/戏票在发售当日有效。These bus/train/play tickets are available on the day of issuance. The expressions in bold indicate the possible collocates of “available” in this case. At the same time, aside from being used as attributive in Chinese, “yǒuxiào 有效” can also appear as a verb in Chinese, which
can be rendered into corresponding “be+adjective” structure. Therefore, in offering examples, the following example can be provided: “这种证件在三年内有效。” It can be translated as “This type of documents is valid for three years.” The “for” in bold implies the habitual preposition following “valid”, which is intended for catching dictionary users’ attention.

3.2.4 Freely Translated Headword in Various Contexts Ought to be Increased

The number of free translation of a Chinese headword in different contexts in productive C-E dictionaries ought to be increased. Many usages in Chinese-English translation are special expressions in specific translating contexts, thus hard to be abstracted into specific senses and impossible to provide an equivalent (Shi and Huang 2016: 2-3). Even if it is hard to draw an equivalent out of this type of translation, the translation of a headword is vivid and native. Therefore, such expressions ought to be selected to be the examples of freely translated headwords in contexts. Translations like them can help dictionary users go beyond seeking for exact equivalent expressions (Pan 2003: 90). Even if “hold/hold good/be good for/do the trick” have been offered as the examples of the free translations of “yǒuxiào 有效” in context, other expressions, such as “work/do the work/job, expire” can be further supplemented. For example, “这些药在 2018 年 2 月六日前有效。” can be translated into “These medicine will expire by 6th, Feb, 2018.” However, no equivalents can be abstracted from this translation.

3.3 Using Large Chinese-English Parallel Corpora

Large Chinese-English parallel corpora can be utilized to improve the quality of a productive C-E dictionary. Since Chinese-English corpora have an overall impact on compiling such a dictionary, they are discussed separately here. The five dictionaries above do not mention making use of such corpora for compilation. However, such corpora can help C-E dictionary compilers “select accurate equivalents” (Li 2011: 69-70), which is tremendous work for them. The equivalents drawn from good translations (Qian 2001: 80) are directly insertable in translation, thus very helpful in meeting productive needs.

Aside from helping compilers’ selecting equivalents, such corpora can also help them find suitable examples and their translations. For English and Chinese, most vocabulary do not match on the one-to-one basis (Li 2006: 42). Therefore, it is hard to build up the correspondence between English and Chinese on the lexical level, but such matching on the sentence level is comparatively easy to attain (Li 2009: 63). The correspondence on the sentence level between English and Chinese can thus provide the free translations of a headword for compilers.

The translation of examples in productive C-E dictionaries ought to seek for the
functional equivalence, paying little attention to the equivalence in terms of part of speech, syntax and superficial meaning (Huang 2014: 16). These freely translated examples can help dictionary users understand the subtle differences of the usage of various translations of a headword in different contexts, the flexibility of language usage and the subtlety in [English] diction (Pan 2003: 91). They can “expand the vision of translators and language learners, which is far more valuable for translating practice than the limited list of translated equivalents in general C-E dictionaries” (Liu and Zhang 2012: 20).

**Conclusion**

The paper looks into the treatment of “yǒuxiào 有效” in five large and medium productive C-E dictionaries first and summarized their shared features in terms of the provision of equivalents, examples and their translations. Such dictionaries aiming to help with Chinese-English translation need to pay attention to highlighting the different features of equivalents. As for the examples and their translation, those dictionaries should increase the number of examples, balance the number of phrase and sentence ones, and carefully weigh what type of examples ought to be used. The illustrating quality can be further improved and the free translation of a headword in various contexts should be increased as well. To compile such dictionaries, compilers should make good use of large Chinese-English parallel corpora, particularly to provide more free translations of a headword in contexts. The enhancement of the translating quality of productive C-E dictionaries can still be approached from diverse perspectives in the future, particularly exploring the role of large Chinese-English parallel corpora in helping achieve this goal.

**References**


Zhao, G. 2006. Some Thoughts on Compiling Chinese-English Dictionaries.
Lexicographical Studies, 1: 94–100.

**Appendix 1** The classification of illustrative examples for “yǒuxiào 有效” and their translations in five C-E dictionaries

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total illustrative examples</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>Illustrating the usage of equivalents</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>Details</td>
<td>1. 有效合同 valid contract 2. 有效证件 valid documents 3. 采取有效措施 take effective measures 4. 这药治疗哮喘病有效。This is an efficacious (or effective) drug for asthma. 5. 这张车票三日内有效。This train ticket is good (or valid) for three days.</td>
<td>1. 有效预防天花 be efficacious in preventing smallpox 2. 做有效工作 do effective work 3. 形成有效的约束机制 establish effective and restraint mechanisms 4. 一年/三月内有效 be valid for one year/three months 5. 有效版权effective copyright 6. 有效补救办法 effective remedy 7. 有效措施/管理 effective measures/management 8. 有效合同/协议 Valid contract/agreement 9. 有效合同/证件/身份证 valid passport/documents/identity card 10. 有效治疗方法 effective treatment 11. 有效日期 effective date/date of validity 12. 新药片非常有效。The new tablets are tremendously effective.</td>
<td>1. 有效方法 an effective method 2. 有效时间 effective/net time 3. 此证明一周内有效。This certificate is valid/good within the first week of its issuance. 4. 两种文本同样有效。Both texts are equally authentic. 5. 这种药治疗肝炎非常有效。It is a very effective drug for hepatitis.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Freely translated headwords in context</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Details</td>
<td>1. 这项指示依然有效。This</td>
<td>1. 实现有制与市场经济的有效结合 effect a good interface between public ownership and</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Directive still holds good.</td>
<td>Market economy remains in force/in effect. 2. 依然有效 remain in force/in effect. 3. 在任何时间和地点都有效 hold good at all times and all places. 4. 本合同不再有效。The contract will no longer stand. 5. 本协议/法律仍然有效。The contract/law still holds. 6. 这种药应该很有效。This medicine should do the trick.</td>
<td>This regulation still holds good.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Contrastive Study of Pragmatic Information between Learner’s Dictionaries

Yueting Rong
Guangdong University of Finance and Economics
rongyueting@163.com

Abstract

Pragmatic information (PI) which falls into the category of meaning and conveys speaker’s meaning is one of the most essential items to be included in learner’s dictionaries (LDs). However, researches on PI in Chinese dictionaries for foreign learners (CDFLLs) are fragmentary with the result that CDFLLs are still barren of PI, which greatly reduces their utility and quality. So far no consensus seems to have been reached regarding some issues. For instance, no set of rules have been formulated for a systematic treatment of PI; The definition and classification of PI are still under dispute; What kinds of PI can be incorporated into CDFLLs and how to present PI etc.. In view of this, this study aims at exploring contrastive study of PI between English learner’s dictionaries and CDFLLs with the purpose of improving CDFLLs.

Keywords: PI, Learner’s dictionary, CDFLL, EDFLL, investigation.

1. Survey overview

1.1 Words with PI investigated

The concept of pragmatic information (PI) was proposed by Leech & Thomas (1987) for the first time in their attempt to combine PI with the compilation of learner’s dictionaries (LDs). They also pointed the importance of PI to foreign language learners and the absence of systematic inclusion of PI in LDs: “The study of pragmatics is very important in terms of the needs of foreign students' communication. Surprisingly, so far no one has seriously tried to enter the pragmatic information into the dictionary for foreigners to learn.

The researchers' opinions are not consistent on what is PI. Leech & Thomas (1987) argues that PI is “not the same as the literal meaning, unconventional, conventionally meaning of the speaker.” Hartmann & James (2000: 111) argues that PI is “information about the speaker's emotions, attitudes and socio-cultural rules, as well as the selections of the speaker's vocabulary, politeness, formal level, and so on,”. Atkins & Rundell (2008: 422-426) argues that “PI is extra-linguistic information that conveys
not linguistic meanings, but speakers' attitudes in different contexts and falls within the category of socio-cultural competence.”

Although these researchers have different descriptions of PI, they all share three common factors of what is PI: “situational”, “affective and attitudinal” and “speaker”. Hartmann & James's “formal”, Atkins & Rundell's “different scenarios” are related to “situational” factor. Hartmann & James's “speaker's emotion, attitude expression”, Atkins & Rundell's “speaker's attitude” and so on and “emotion and attitude” factor. In the above description, the choice of PI is related to “speaker”, for example, Leech & Thomas explicitly refers to PI as “speaker's meaning”. In addition, Hartmann & James and Atkins & Rundell also argue that PI is culturally relevant and belongs to the category of socio-cultural competence.

Based on the above viewpoints, the present study suggests that PI involves three elements: speaker attitude, context, and culture. As long as one of the three elements mentioned above, it can be regarded as PI, and as long as this information is conventional, it can be included in the dictionary.

PI is mainly attached to the typical PI vocabularies, such as hedges, which convey speaker’s attitude. The ambiguity of the objective world determines the fuzziness of human thinking. The PI contained in hedges is an important content for foreign language learners to grasp. Common hedges in English are: probably, a bit, somewhat, kind of and so on. Common hedges in Chinese are: 好像, 可能, 大约, 有点, and so on. “可能” is a typical hedge in Chinese, which is used to express the views euphemistically, and to keep the negative side of the person, such as “我可能来不了了”

1.2 Dictionaries investigated

In this study, five English Dictionaries for Foreign learners (EDFLLs) are: Oxford Advanced Learner's Dictionary, 8th (2010) (Oxford); Collins COBUILD Advanced Learner's Dictionary, 6th, 2008(Collins); Macmillan English Dictionary for Advanced Learners 2nd, 2007 (Macmillan”); Longman Dictionary of Contemporary English, 5th, 2009 (Longman); and Cambridge Advanced Learner's Dictionary, 2008 “(Cambridge). These five EDFLLs chosen are regarded as top five learner’s Dictionaries in English language.


In this study, the five CDFLLs are chosen to be investigated because they meet at least three requirements: they must be published by well-known publishing houses;
they should be at least medium-sized versions; and they must be well-known and well-tested works.

1.3 The words investigated

12 hedges which are investigated are: kind of, probably I am afraid, pretty, somewhat, approximately, somewhat, about, possibly, a bit, perhaps; The 12 hedges are basic words coming from Oxford 3000. 12 hedges in Chinese are: 十分、可能、好像、据说、大约、差不多、有点1、左右、几乎、稍微、通常 and 多半. The Chinese hedges are also basic words in Chinese language.

Due to the differences between English and Chinese, the Chinese words investigated can not be one-to-one correspondence with the English words investigated.

2. An Empirical Study of English Learner ’s Dictionaries

2.1 The survey of PI inclusion in EDFLLs

The PI inclusion in EDFLLs is showed in Table 23.1

<table>
<thead>
<tr>
<th>DICs Words</th>
<th>Oxford</th>
<th>Longman</th>
<th>Collins</th>
<th>Cambridge</th>
<th>Macmillan</th>
</tr>
</thead>
<tbody>
<tr>
<td>kind of</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>Probably</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>I am afraid</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>Pretty</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Quite</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>somewhat</td>
<td>C</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>approximately</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>to a/some degree</td>
<td>A</td>
<td>C</td>
<td>D</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>About</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>Possibly</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>a bit</td>
<td>C</td>
<td>C</td>
<td>D</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Perhaps</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
</tbody>
</table>
Notes: A means “absence of entry”; B means “absence of PI”; C means “insufficient of PI” and D means “sufficient of PI”

(1) Absence of entries

Absence of entries means that the word is not included into the dictionary as an entry, and the term does not appear as sub-entry.

The entries “to a / some degree and approximately” belong to the core of the English vocabulary, but they are not included as entries and it is not suitable to do this.

(2) The absence of PI

The absence of PI means that the surveyed dictionaries list the investigated vocabulary as an entry, but only provides the conceptual meaning of the term, and does not provide pragmatic information.

Collins only treats to a / some degree as sub-entry, and does not provide definition but an example, which is not enough to show the PI of hedges:

to a/some degree partly What you say is true to some degree. Collins

(3) Insufficiency of PI

The insufficiency of PI mainly exists in two cases: the PI is not provided in the definition, but the example involved PI; both the definition and example provide PI, but the PI is not enough.

In terms of hedges, the other four English learner’s dictionaries except Collins have the same problem for the processing of a bit: only synonyms of the entry word is provided in the definition. Such as a little, slightly and to a small degree, PI is not provided, but the examples provide a lot of PI. The following examples in an italic part are related to the PI in Table 2.2. (Notes: Examples not related to PI omitted).

Table 2.2 Survey of PI of a bit in EDFLLs

<table>
<thead>
<tr>
<th>DICs</th>
<th>definition</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxford</td>
<td>rather, a little</td>
<td>I can lend you fifty pounds, if you want. That should help a bit.</td>
</tr>
<tr>
<td>Longman</td>
<td>slightly or to a small degree</td>
<td>That’s a bit odd. ‘Are you sorry to be leaving?’ ‘Yes, I am a bit.’ Arent’you being a little bit unfair? I think you’re a bit too young to be watching this.</td>
</tr>
<tr>
<td>Cambridge</td>
<td>slightly</td>
<td>That was a bit silly, wasn’t it? I’m a bit nervous.</td>
</tr>
</tbody>
</table>
(4) Sufficiency of PI

Sufficiency of PI means that the dictionary provides PI in definition, examples and register, and the amount of PI is enough. The five English learner’s dictionaries investigated provide sufficient PI.

*I am afraid* is a typical hedge, euphemistic expression of bad news or rejection and it is a politeness strategy. Five English learner’s dictionaries deal with *I am afraid* as follows: (1) the definition involves its politeness strategy; (2) the examples are rich and there are 5-6 examples; (3) “Longman” provides the domain register “spoken”; “Collins” provides the domain label [SPOKEN, POLITENESS].

*Perhaps* and *probably* are both typical hedges and they are also politeness strategy for tactfully express their views, to avoid offending others. “Collins” deal with perhaps and probably in a suitable way: (1) definition provides the speaker's attitude; (2) examples is vivid. For example, *He probably thinks you're both crazy!* (3) Creat pragmatic label [VAGUENESS], so that PI is conveyed clearly.

**perhaps**

You use perhaps in opinions and remarks to make them appear less definite or more polite. [VAGUENESS]

*Perhaps the most important lesson to be learned is that you simply cannot please everyone.*

*His very last paintings are perhaps the most puzzling.*

**probably**

1. If you say that something is probably the case, you think that it is likely to be the case, although you are not sure. [VAGUENESS]

*The White House probably won't make this plan public until July.*

*Van Gogh is probably the best-known painter in the world.*

2. You can use probably when you want to make your opinion sound less forceful or definite, so that you do not offend people. [VAGUENESS]

*He probably thinks you're both crazy!*  

Collins

2.2 The survey of PI inclusion in CDFLLs

| Hedges | DICs | 可 | 好 | 据 | 大 | 差 | 不 | 有 | 点 | 左 | 几 | 稍 | 通 | 多 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

Table 2.3 A survey of PI inclusion in CDFLLs
(1) Absence of entries

As the survey vocabulary is the basic vocabulary in Chinese, so the absence of entries is a big pity.

(2) Absence of PI

Hedges”好像” means “not very sure of ideas or views,” as a politeness strategy, tactfully express their views and plays a moderating role in utterance. For example, “她好像来不了了,” “事情好像没那么容易解决 “and “好像” is used to express their views or statements of negative information, to ease the tone.”Business Hall”, “Contemporary Chinese” and “Chinese Proficiency” only provide conceptual meanings of “as if” and do not provide PI as hedges.

好像

很像：这几天天气真热，好像夏天了。青年人好像早晨八九点钟的太阳。

好像

足球场的草地好像铺在地上的绿色毯子→足球场的草地跟铺在地上的绿色毯子有相同和相似的地方。例：这姑娘长得真漂亮，好像一朵鲜花一样。天空的白云好像一朵朵飘动的棉花。太阳升起来的时候，好像一个大火球。
好像

非常像：这块石头好像一只老虎。

CPTD

Obviously, in the entry of “好像”, the above three dictionaries only gives the “like” as a metaphorical usage, and the PI of “好像” that is not very sure of the idea or view”, “euphemistically make suggestions”. PI of “好像” is missing.

(3) Insufficiency of PI

“A bit” as a hedge, euphemistically convey negative information or emotions, to alleviate the negative face. The five DICLs provide only a somewhat “conceptual” meaning in the definition, but in the examples they provide “somewhat” PI as hedges.

有点

稍微：……我有点不明白你说什么。……

LCDCPH

有点

今天有点热。→ 今天不是很热，但比较热。例 ……你不给她打电话，有点不合适吧？……

LDCC

有点

表示程度不高：稍微。多用于不如意，不愉快的事情：……你这么说似乎有点不讲道理。咱们这么做有点不太妥当吧。……

CCW
有点

表示略微；稍微：……这句话说得有点叫人摸不着头脑。……

有点

表示程度不高；略微或稍微（程度比有一些更高）(多用于不如意的情况)：……

我稍微有点累，想休息一会。

Obviously, the five CDFLLs only provide the concept of “somewhat”, that is, “slightly, relatively, slightly” and so on, without mentioning its PI, that is “for the negative expression of euphemism or emotion. But the examples are containing PI, such as the example in” business hall” “I do not understand what you say”; example in “Contemporary Chinese” “You do not give her a call, a little inappropriate”; Words “;” you say that seems a bit unreasonable; we do a bit less appropriate “;” “example” in 8000 words This sentence is a bit confusing. “Obviously, these examples of “a little” do not convey the conceptual meaning “slightly”, but PI which is a communication strategy, to retain each other's negative face.

“恐怕” also contains two pragmatic messages: speculation and worry. But “contemporary Chinese” only conveys the “guess” meaning, such as:

恐怕

天这么阴，恐怕要下雨了。→大概要下雨。例：他感冒了，恐怕不会参加晚会。

(4) Sufficiency of PI

It’s a pity that words investigated are not given sufficient PI in CDFLLs.

3. Analysis of PI inclusion between learner’s dictionaries

3.1 Analysis of PI inclusion in EDFLLs

(1) to distinguish between pragmatic information and conceptual meaning
Pragmatic information and conceptual meaning are the basic meanings of vocabulary. In order to master a vocabulary, both are indispensable. Rundell (2007) notes in Macmillan's second edition, “Dictionary users need a range of information, not just meaning, but information about collocation, register, and speaker's attitude.” It can be seen that “Macmillan” treats pragmatic information as important as conceptual meaning. For example, “Macmillan” provides four senses in the entry “pig”, of which the first sense is the conceptual meaning, the other three are with the pig-related pragmatic information, namely: “nasty person; difficult or annoying; “Insulting to the police”, as follows:

(2) To refine PI

EDFLLs treat conceptual meaning and PI equally and respectively and this treatment is closely related to Corpus “A large, diversified corpus is the basis of dictionary compilation,” Rundell, editor of McMillan, describes in the foreword of the dictionary. The five EDFLLs were compiled on a corpus basis. Collins is a pioneer in the combination of corpus and lexicography, and all the words and sentences in the dictionary come from its own large corpus, COUBILD English Corpus. “Collins”, other EDFLLs have adopted a corpus-based dictionary compilation model. Longman established the Longman Corpus Network, Cambridge established the Cambridge Learner's Corpus, Oxford established the Oxford Corpus Collection, Macmillan, It has the World English Corpus (World English Corpus). These corpora are constantly updated. Corpus as the basis, coupled with advanced corpus analysis and extraction technology and advanced compilation concept, make the learners of English learning lexicon comprehensively describe the lexical information as possible.

Corpus-based PI processing, senses are treated in detail, the amount of PI is sufficient. For example, “McMillan” analyzes the various usage of “about” as hedges in corpus and present five senses of “about” as hedges:

about

1 You use about to introduce who or what something relates to or concerns….
2 When you mention the things that an activity or institution is about, you are saying what it involves or what its aims are…
3 You use about after some adjectives to indicate the person or thing that a feeling or state of mind relates to. …
4 If you do something about a problem, you take action in order to solve it…. 
5 When you say that there is a particular quality about someone or something, you mean that they have this quality….

Macmillan

3.2 Analysis of PI in CDFLLs

The typical vocabularies which carry PI are hedges and discourse markers, which are characterized by the conceptual meaning, the expression of the signifier and the
signified relationship, the pragmatic information, and the communicative mood of the speaker and attitude. Unfortunately, the EDFLLs only focuses on the conceptual meaning of the words and ignores the PI. Therefore, based on investigation, it is found that the main problem of the inclusion of linguistic PI is that the PI is not mentioned in the definition, but the examples are related to the pragmatic information, so the conceptual and pragmatic information are confused. And Even both the definition and examples convey PI, PI is still insufficient.

(1) Follow the linguistic dictionaries format

Linguistic dictionary is mainly used for “query meaning”, and learner’s is mainly used for “learn to use words.” (Chen Chuyuan 2005). Modern Chinese Dictionary is a generally accepted linguistic dictionary. Its compilation purpose is as stated in its foreword: “It is compilation in accordance with the provisions of the State Council to promote simplified Mandarin”. Obviously, “Modern Chinese Dictionary” and the EDFLLs have different purposes of editing and dictionary user groups, thus must be different in the dictionary information inclusion. From the purpose of compilation, the Modern Chinese dictionary is mainly to provide the concept of vocabulary and other related language knowledge, while the an EDFLL should not only provide conceptual meanings, but also provide a series of lexical-related grammatical information, pragmatic information and stylistic information Lexical information to help L2 learners construct vocabulary cognitive context. However, an embarrassing fact was found in the investigation of PI in the CDFLLs for foreigners in. The main current situation of the CDFLLs are still to copy Modern Chinese Dictionary as a blueprint, completely or partially, copy definition and examples in Modern Chinese Dictionary. For example, the definition of “donkey” in “8000” is exactly the same with the definition of donkey in Modern Chinese Dictionary.

The above examples fully show that the CDFLLs are not based on the corpus to analyze and extract the conceptual meaning or pragmatic information, but by imitating or copying the traditional Chinese linguistic dictionaries compilation model as its main compilation paradigm, so the pragmatic information in the CDFLLs neglect is the inevitable result.

(2) Affected by the ontology of language, only focus on conceptual meaning and ignore pragmatic information

It is found that the CDFLLs are influenced by the structuralist methodology of the ontology of the language, statically studies the language, separates the linguistic knowledge from the knowledge of the world, ignores the many factors outside the language (such as social factors, cultural factors, language Users, etc.) on the language of the language as a self-contained, closed meaning of the symbol system. Therefore, the CDFLLs only provide the conceptual meaning of the entry words and ignores the pragmatic information related. For example:
It can be seen from the above cited examples that the vocabulary is divorced from its context, CDFLLs are concerned only with the conceptual meaning of the words, and ignores other factors such as the pragmatic information related to the vocabulary. Thus, it can not reflect the “real usages of words”. The meaning conveyed by the dictionary is isolated, fragmentary, resulting in the dictionary users only know their “meaning”, but I do not know its “usage.”

(3) PI is confused with conceptual meaning

In the investigation it is found that, even if the PI and conceptual meaning is included in the CDFLLs, they are mixed up. That is, PI is not treated as a sense and is not provided in the definition, but the usage of PI is provided in the examples. The confusion of PI and conceptual meaning reflects that CDFLLs consider conceptual meaning to be the whole of dictionary meaning, and does not recognize the existence of the concept of PI. However, the pervasiveness of the pragmatic information makes it possible for the CDFLLs to bring PI into the examples. such as:

有点

稍微：......我有点不明白你说什么。......

LCDCPH)

In the above cited examples, we can see that the definitions only relates to conceptual meaning, but some examples are concerned with PI. This is enough to explain the mix between conceptual and PI.

4. Conclusion: the Strategies of PI inclusion in CDFLLs

The absence of PI in CDFLLs is the result of the backwardness of the compilation idea, and it is also a result of underdevelopment of compilation techniques and compilation methods. Due to the lagging state of corpus and corpus analysis software development, a considerable number of Chinese dictionary compilers have to rely on the traditional citation files in the process of dictionary compilation. The lack of linguistic representation in citation files is also reflected in the limited availability of citation sources as well as personal preferences and habits during the excerpt. Both paper and electronic citation files are problematic in that the true use of the language cannot be viewed in a panoramic manner. Dictionaries cannot keep up with the pace of information technology, it is difficult to achieve modernization. In the modernization of dictionaries, the corpus of dictionaries should be set up first. If CDFLL compilation is not corpus based, those dictionaries are not dictionaries which reflect the true language but the dictionaries which reflect “expert language sense”.

647
In view of the shortcomings in the construction of PI in the CDFLLs in Section 6.2.1 and the benefits of corpus in the collection of information in the dictionary compilation, the strategies of PI inclusion in CDFLLs are as follows:

(1) **Large-scale corpus is supposed to build as an important basis for dictionary compilation, rather than the traditional language dictionary as a blueprint for editing**

At the specific operational level of dictionary compilation, corpora technology brings many beneficial changes, such as choosing words, frequency annotation, providing examples and description of language prompts. On the basis of the corpus, the collection of the information in the learner’s dictionary is more authentic, objective and scientific. (WEI Xiang-qing, 2009; Rundell, 2009b; Wang Fufang et al., 2004) The irresistible potential energy of corpus in lexicographic compilation has led lexicographers to take a new step in terms of reliability, breadth and depth. And the introduction of corpus has made dictionary editors look at dictionary compilation from a new perspective, which has led to the change of lexicographical paradigm.

(2) **To establish the word frequency**

The establishment of word frequency that is to determine the frequency of use of information, which relates to the selection of words. For example, “McMillan”, based on corpus analysis, builds 7500 as core words, the decoding and encoding of these words information is what the second language learners must master. “Macmillan” provides these words in detail to provide decoding and encoding information, and to identify them by asterisk, the use of high and low frequency of a star to three stars to represent the way, such as:

hello ★★★ basically ★★ cruelty ★

“Longman”, through the word frequency statistics, builds 3000 words which are used commonly in the English spoken and writing, such as:

red S1 W1; hello S1; persuade S3 W2; crazy S21

S1 stands for 1000 common words in spoken language, S2 stands for 2000 commonly used words in spoken language; S3 stands for 3000 words commonly used in spoken language; W1 stands for 1000 words most commonly used in writing; W2 stands for 2000 words most commonly used in writing; W3 represents the most commonly used 3000 words in writing.

On the basis of fully studying the corpus, CDFLLs are supposed to build entry words based on corpus analysis.

(3) **Accurate PI and detailed PI**
Accurate PI is a very accurate description of the meaning of words under the support of large amounts of corpus data. Such as bump off in the dictionary is usually interpreted as “to kill somebody”. But after a great deal of research on corpus data, the Longman Language Activator (2003) interprets bump off as:

bump off: an informal word meaning to kill someone or arrange for them to be killed, especially because they know about things you have done wrong, or are dangerous to you.

Refinement of PI refers to the analysis of a large number of language data in the corpus, the detailed division of PI. For instance the English discourse marker hello, contains abundant pragmatic information, Longman divided its meaning into six senses. If there is no corpus support, it is difficult to make such detailed division.

**hello**

1 used as a greeting when you see or meet someone…
2 used when answering the telephone or starting a telephone conversation…
3 used when calling to get someone’s attention …
4 used when you think someone is not acting sensibly or has said something stupid …. 
5 British English used to show that you are surprised or confused by something…
6 say hello to have a quick conversation with someone …

(4) **Highlight the PI collocation information**

Vocabulary can be used only when used in conjunction with other vocabularies. In the corpus of support, the dictionary will be collocated with the information included in the dictionary. “Macmillan” supply the dictionary users with 506 collocation information. For example, under the term advice, there are adjectives and verbs that are often used with advice:

**Adjectives frequently used with advice**
expert, financial , legal , medical , practical , professional

**Verbs frequently used with advice as the object**
accept , ask for , disregard , follow , get , give , heed , ignore , need , obtain , offer , provide , receive , reject , seek , take , want

(5) **PI is supposed to be authentic and reliable**

Examples in CDFLLs are an integral part of an entry. The important role of examples is to pass, and strengthen the word information, to provide information on the structure of the syntax, to provide collocation information, transmission of cultural information. In the traditional dictionary compiling paradigm, due to the limitation of the compiling technique, when it is difficult to find the examples related to the word-words, some examples are often adapted or self-compiled. Such an example, while conveying the correct grammar, syntax information, but not true, authentic language, the drawback
lies in the inability to convey the authenticity of the language. In the corpus-based lexicographic paradigm, the scope of examples is more extensive, with authenticity, reliability, and sometimes vivid. For example, “Collins” probably provides examples of hedges:

**probably**

1. ....
2. VAGUESS You can use probably when you want to make your opinion sound less forceful or definite, so that you do not offend people.

*He probably thinks you're both crazy!*

Collins

(6) To provide usage notes

The corpus, based on the analysis of a large number of language use facts, extracted the best evidence of the use of the language. By analyzing the learners’ corpus, the dictionary editors can understand the difficulties that the learners may feel in their study, and confuse the differences of the words in the dictionary, so as to help the learners solve the problems. For example “Longman” uses the way “Grammar” (grammar tips) to give extra explanation to some words:

*information…..Grammar Information is an uncountable noun and has no plural form. Use a singular verb after it: The information was not passed on to the hospital.*

CDFLLs can be used to summarize the usage information of the PI related to the entry words by analyzing the corpus and present it to the dictionary users in the form of “usage notes” and so on.

References

A. Dictionaries


B. Other Literature


Weak-Form Representation in Online English Learner's Dictionaries

Wenlong Sun
Bilingual Dictionary Research Centre, Nanjing University
sherwinlong@163.com

Abstract

The use of weak forms is an essential feature of English speech and can exert an influence on EFL learners’ oral production and listening comprehension. The most common weak-form words basically fall into the category of function words or grammatical words, the acquisition of which usually happens at an early stage of the EFL learners’ learning process. In terms of the weak-form representation in learner’s dictionaries, the dilemma faced by the lexicographers is that few EFL learners seem to consult a dictionary for the pronunciation of function words, though some of them (including a fair number of advanced EFL learners) have difficulties in acquiring the weak forms or fail to foster the awareness of weak-form pronunciations. Meanwhile, indicating pronunciation in dictionaries has long been considered to be a secondary business while the representation of weak forms in learner’s dictionaries seems to be the ‘chicken ribs’, i.e. each learner’s dictionary would choose to include the weak forms in it, yet the representation of weak forms is too oversimplified, of which the learnability is weaken to a large extent. Owing to the rapid development of electronic lexicography, the representation of phonetic information in learners’ dictionaries seems to be greeted with a heaven-sent opportunity. Online learner's dictionaries represent the trend of future pedagogical lexicography. The present study, in order to redress the anti-phonetic bias of current lexicography, first makes an inventory of fifty-one most commonly used weak-form words, they are: seven determiners (a, an, any, his, some, the, your), eight pronouns (he, her, him, she, them, us, we, you, me), six conjunctions (and, as, but, or, than, that), nine prepositions (at, by, for, from, into, of, onto, per, to), nineteen auxiliary verbs (am, are, be, been, can, could, do, does, had, has, have, is, must, shall, should, was, were, will, would), and one adverb (there). Then this paper examines the representation of the fifty-one weak-form words in the online versions of the ‘Big Five’: Oxford Advanced Learner’s Dictionary, Longman Dictionary of Contemporary English Online, Cambridge Learner’s Dictionary, Macmillan Dictionary, and COBUILD Advanced English Dictionary. A comparison is made among the ‘Big Five’ in the three following aspects: the breadth of weak-form representation, the depth of weak-form representation, and the consistency of weak-form representation. Finally, some suggestions relating to the improvement of weak-form representation in the online English learner’s dictionaries are made.

Keywords: phonetic representation, weak form, online learner’s dictionary
1. Introduction

The use of weak forms is an essential feature of English speech. Almost all the words which have both a strong and weak form belong to a category that may be called function words, filling in between content words, and making sentences ‘work’, grammatically (Kelly 2000; Roach 2008). Compared with the strong forms, the weak forms of these structure words are prone to “show reductions of the length of sounds, centralisation of vowels towards /ə, ɪ, ʊ/ and the elision of vowels and consonants” (Gimson 1980:261; Cruttenden 2014:273), which is a hard nut for nonnative English learners to crack. In EFL teaching practices, it is common to find that a fair large number of EFL learners, especially those whose mother tongues are significantly different from English, tend to pronounce the strong forms of those structure words as they are pronounced in isolation. Although it is possible to use only strong forms in speaking, and usually they can still be understood by other speakers of English, the negative impact is twofold. First, an “all-strong-form” pronunciation is detrimental to one’s oral production since most native speakers of English find it unnatural and foreign-sounding (Deterding 2006; Lee et al. 2006; Roach 2008). Second, the failure of using weak forms correctly is a barrier to listening comprehension (Hua 1991; Dauer and Browne 1992; Xu 2009), and even gives rise to listening anxiety (Wei 2006; Bekleyen 2009). Therefore, “learners who aim at a native English accent (British or American) must learn the weak forms of function words and regard them as the regular pronunciations” (Cruttenden 2014:276).

Accordingly, there is an urgent need for high quality, effective materials, especially computer-based materials with audio demonstrations, for EFL learners’ acquisition of weak forms. The learner’s dictionaries are a beneficial complement to EFL course books and English pronunciation teaching, and the weak forms “should be recorded first in English dictionaries and taught first in the textbooks, especially those printed for Chinese Learners” (Zhang 2002:150). Although, the learners’ dictionaries, which are specially designed for EFL learners, have a tradition of valuing phonetic transcription, it seems ironically that lexicographers do not appear to have much concern for pronunciation (Piotrowski 1987; Lewis 2014), and indicating pronunciation is often under-estimated by the critics of dictionaries as being a “derivative business” (Magay 1979: 99) or “secondary importance” (Magay 1981:86). Even in today’s world, the state of the art in the lexicographic treatment of pronunciation “appears to be sadly anachronistic” (Sobkowiak 2008: 68).

Owing to the rapid development of electronic phonolexicography during the final decade or so of the last century, the representation of phonetic information in learners’ dictionaries seems to be greeted with a heaven-sent opportunity to improve the EFL learners’ pronunciation. On the one hand, the electronic dictionaries have broken through the space limitation existing in the print dictionaries; on the other hand, the rapid advancement of phonetic techniques has brought the electronic dictionaries a role

---

1 For example, languages such as Mandarin, French, Hindi, Italian, Spanish, Turkish, etc., are generally considered to be syllable-timed, which differs from the stress-timed language of English.

2 For instance, the Oxford Advanced Learner’s Dictionary engaged Windsor Lewis, Gimson, Ramsaran and Ashby respectively as its phonetics editors in order to cater to the phonolexicographic needs of EFL users.
transition from ‘silent teachers’ to ‘talking teachers’. However, “despite its enormous pedagogical potential, and unlike other sub-disciplines of learners’ lexicography, theoretical and applied phonolexicography had hardly been developing in the last decades” (Sobkowiak 2007: 150). The weak-form pronunciations in learner’s dictionaries have become chicken ribs (things of little value or interest), of which the reasons are complicated. For one thing, the function words that have weak-form pronunciations are the most commonly used words and are learned by EFL learners at an early stage (usually in elementary school or in the early period of secondary school). Therefore, few advanced learners would consult a dictionary for the pronunciation of a commonly used word, because they thought there is nothing new to their knowledge of the pronunciation of these function words. For another, even for the learners who are definitely in need of the usages of weak forms, they could hardly benefit from the current learner’s dictionaries since the weak-form representation is too oversimplified, only with the list of written transcriptions in most cases.

In order to redress “the anti-phonetic bias of current lexicography” (Sobkowiak 2007: 149), as well as raise the EFL learners’ awareness of weak forms, this paper makes an attempt to survey the weak-form representation in the online English learner’s dictionaries (hereafter called OELD) from three perspectives: (1) the breadth of weak-form representation; (2) the depth of weak-form representation; (3) the consistency of weak-form representation. Five OELD (hereafter called ‘the Big five’) are chosen in this study, they are: Oxford Advanced Learner’s Dictionary (hereafter called OALD), Longman Dictionary of Contemporary English Online (hereafter called LDOCE), Cambridge Learner’s Dictionary (hereafter called CLD), Macmillan Dictionary (hereafter called MD), and COBUILD Advanced English Dictionary (hereafter called COBUILD)\(^3\). Then, a comparison is made among the Big five in terms of the weak-form representation from the three mentioned aspects above. And finally, some tentative suggestions are made in order to optimize the weak-form representation of the OELD in the future.

2. Theoretical considerations of weak-form representation in OELD

2.1 The breadth of weak-form representation

The breadth of weak-form representation refers to the quality of consisting of weak-form information in OELD, which mainly involves the following three aspects: the scope of weak-form words, the scope of weak-form transcriptions and accompanying audio recordings, and the scope of weak-form variants.

First, in terms of the first aspect, the lexicographers should select out the weak-form words that need to be represented in OELD. Many phoneticians have made a list of the most commonly used weak-form word. For instance, Jones (1962) gave a

---

\(^3\) The portal sites of OALD, CLD, and COBUILD contain different dictionary types in order to meet various needs of the users. This present study merely chooses the component of learner’s dictionary from the three online dictionaries.
list of all the English words which have weak forms differing notably from the strong forms\(^4\), but he also pointed out that some weak forms of the words such as my, many, nor, so, such, etc., were seldom used by people; Gimson (1980), on the basis of Jones (1962), enumerated a list of the most common of weak-form words\(^5\), of which the weak forms were the first choice; O’Connor (1980) pointed out that the EFL learners must learn to use the weak forms of thirty-five English words if they want their English to sound English\(^6\); Roach (2008) shared a similar view with O’Connor (1980) and held that the EFL learners need to learn about roughly forty weak forms to help them to understand what they hear\(^7\); Zhang (2002, 2007), who is a student of Gimson, recommended forty-two most common weak forms for the Chinese English learners\(^8\). Therefore, it can be seen that the number of the most common weak-form words is roughly forty, which is the top priority for EFL learners to acquire. For the lexicographers, they have alternatives to make their own wordlist of the weak-form words when it comes to the weak-form representation in learner’s dictionaries, but this should be based on the most common weak-form word list.

Second, for the internet lexicography, it is a recognized practice to supply the written transcriptions with accompanying audio recordings for the sake of users. As to the pronunciations of weak-form words in the OELD, the audio recordings mainly include three parts: the recordings of weak forms (including weak-form variants), the recordings of strong forms, and the American pronunciations of both weak forms and strong forms\(^9\). For each written transcription of the weak-form words, the audio recording should be given either before or after the transcription.

Third, if one weak-form word has more than one pronunciation, the compliers also need to make a decision on how many weak-form variants should be included in the OELD? It is notable that there exist significant distinctions between native speakers and foreign learners of English. Taking and for instance, Cruttenden (2014) introduced six weak-form pronunciation variants, i.e. æn, ən, ə, n, and ð, for the native speakers in his phonetic work. However, for the EFL learners, ‘it’s unnecessary to list all of them but choose the commonly used forms which are easier for them to acquire, because most of the EFL learners have little knowledge of phonetics or phonology, and

\(^4\) The list includes sixty-one words, they are a, am, an, and, are, as, at, be, been, but, by, can (auxiliary), could, do (auxiliary), does (auxiliary), for, from, had (auxiliary), has (auxiliary), have (auxiliary), he, her, him, his, is, ma’am, many, me, must, my, nor, not, of, or, per, Saint, shall, she, should, Sir, so, some, such, than, that (conj., relative pron.), the, them, there, till, time(s), to, upon, us, was, we, were, who, will, would, you, your, respectively.

\(^5\) Gimson’s list includes forty-eight words, they are a, am, an, and, are, as, at, be, been, but, can (aux.), could, do (aux.), does (aux.), for, from, had (aux.), has (aux.), have (aux.), he, her, him, his, is, me, must, not, of, Saint, shall, she, should, Sir, some (unspecified quantity), than, that (conj. and rel. pron.), the, them, there (indef. adv.), to, us, was, we, were, who, will, would, you, respectively.

\(^6\) The thirty-five weak-form words are and, as, at, be, been, but, can (aux.), could, do (aux.), does (aux.), for, from, had (aux.), has (aux.), have (aux.), he, her, him, his, is, me, must, not, of, Saint, shall, she, should, Sir, some (unspecified quantity), than, that (conj. and rel. pron.), the, them, there (indef. adv.), to, us, was, we, were, who, will, would, you, respectively.

\(^7\) The forty weak-form words are the, a, an, and, but, that, than, his, her, your, she, he, we, you, him, her, them, us, at, for, from, of, to, as, some, there, can, could, have, has, had, shall, should, must, do, does, am, are, be, is, was, has, have, had, can, shall, will, would, you, your, respectively.

\(^8\) The forty-two weak-form words are further divided into six groups by Zhang, they are five conjunctions: and, as, but, that, than; eleven pronouns: he, him, his, she, her, me, them, we, us, you, that; sixteen aux. verbs: do, does, am, are, be, been, is, was, has, have, had, can, could, shall, will, must; four determiners: a, an, the, some; five prepositions: at, for, from, of, to; and one adverb: there.

\(^9\) In this present study, only the British pronunciations of weak-form words are examined, while the American pronunciations of weak-form words are not involved because the differences between the British and American pronunciations are not the focus of this study.
lexicographers should not expect all of them to be phoneticians, though some upper advanced learners may be. This issue demands utmost considerations from the dictionary-makers.

2.2 The depth of weak-form representation

The depth of weak-form representation can be defined as the great amount of knowledge that is related to the usage of weak forms. It is meaningless to merely list the written transcriptions and accompanying audio recording of the weak forms, because for the EFL learners, especially those who have little knowledge of weak forms, what they want most is to learn how to pronounce the weak forms properly in authentic context. Accordingly, the compliers should provide the usage notes of weak forms for the learners with specific examples so as to deepen the learners’ understanding of the weak forms. However, it is far from enough to include weak-form usage notes in learner’s dictionaries, the strong forms should also be placed equal importance as the weak forms since the two are interconnected with each other. Roach (2008) claimed that the strong forms had been neglected for a long time by phoneticians because they spared no efforts to introduce weak forms in their works while the significance of strong forms were not realized by both teachers and students. Roach’s comment is instructive because he realized the opposite-unity relation between weak forms and strong forms. The strong forms usually carry special meanings (e.g. for contrast) in a certain context and are relatively easier for the learners to acquire with explicit rules to follow. Therefore, the inclusion of strong-form usage notes should be an important component of weak-form representation in learner’s dictionaries.

Moreover, it is probably wise to choose one or two common weak forms and strong forms to construct an awareness-building exercise, despite some voices to the contrary: “It cannot realistically be seen as part of the dictionary's function to teach the sound system” (Brazil 1987: 161). EFL dictionaries are gradually transformed into one-stop learning resources, with grammatical, pragmatic and phonetic advice on board, a variety of self-study facilities, multiple suites of exercises and hyperlinks to even more support on the world-wide web (Sobkowiak 2007). In the era of internet lexicography, the didactic function of learner’s dictionary has not been brought into full play under the traditional concept that dictionary is a consulting tool rather than a learning tool. Actually, as early as the 1970s, Hausmann had made a distinction between a “learning dictionary” and a “consultation dictionary”; according to this typology, “either the learning dictionary is to be studied from beginning to end or its individual articles are to be studied in this way, whereas the consultation dictionary is to be used for consultations related to text reception and text production” (Tarp 2004: 248). However, due to the limitation of technology, Hausmann’s ‘distinction’ has long been considered to be a utopian scheme. Currently, thanks to the information and multimedia technology, dictionary-space limitation is no longer a stumbling block for the OELD, which has brought new vigor and vitality into carrying out the conception of ‘dictionary as a learning tool’. In an era of Internet lexicography increasingly
dictated by techniques, lexicographers need to develop and demonstrate learnability — the core conception of ‘learning dictionary’ — in order to stay relevant and succeed. Against this background, adding pertinent exercises of both weak forms and strong forms is a beneficial exploration when it comes to the depth of weak-form representation in OELD.

2.3 The consistency of weak-form representation

One persistent concern regarding dictionary-making is the issue of consistency, which is a major concern of phonetic lexicography when dictionary makers and critics ponder sound representation from the lexicographic perspective (Sobkowiak 2002). When it comes to the weak-form representation in OELD, lexicographers need to focus on the following three phonological consistency issues.

The first issue refers to the sequential consistency between weak forms and strong forms, i.e. which should come first when the two pronunciation forms are represented in the dictionary. Similarly, the second issue relates to the sequential consistency of different weak-form variants, i.e. how to rank weak-form variants when one weak-form word has more than one weak form? The third issue involves the consistency between the weak-form written transcriptions and their accompanying audio recordings, i.e., are the audio pronunciations match well with their transcriptions?

Overall, regarding the consistency of weak-form representation, the following issues need to be examined: (1) In terms of the precedence relationship between weak forms and strong forms, are there any inconsistencies among different dictionaries or within a particular dictionary? If any inconsistency is found, is there any motivations, either lexicographically or linguistically, accounting for that? Or is it just a careless mistake? (2) In terms of the precedence relationship among different weak-form variants, are there any inconsistencies among different dictionaries or within a particular dictionary? If any inconsistency is found, is there any motivations, either lexicographically or linguistically, accounting for that? Or is it just a careless mistake? (3) In terms of the consistency between the transcriptions and audio recordings within a particular dictionary, are there any mismatches?

3. The status quo of weak-form representation in OELD

This paper makes an inventory of the most common weak-form words on the basis of Zhang (2002, 2007) and Roach (2008). The list includes fifty-one function words, they are: seven determiners (a, an, any, his, some, the, your), eight pronouns (he, her, him, she, them, us, we, you, me), six conjunctions (and, as, but, or, than, that), nine prepositions (at, by, for, from, into, of, onto, per, to), nineteen auxiliary verbs (am, are, be, been, can, could, do, does, had, has, have, is, must, shall, should, was, were, will, would), and one adverb (there). On the basis of this weak-form wordlist, the status quo of weak-form representation in the OELD is examined in the following three aspects.
3.1 The breadth of weak-form representation in OELD

3.1.1 The coverage rate of weak-form written transcriptions

Table 1 The coverage rate of weak-form written transcriptions in Big five

<table>
<thead>
<tr>
<th></th>
<th>I / II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD</td>
<td>38/51</td>
<td>75%</td>
</tr>
<tr>
<td>COBUILD</td>
<td>40/51</td>
<td>78%</td>
</tr>
<tr>
<td>OALD</td>
<td>44/51</td>
<td>86%</td>
</tr>
<tr>
<td>LDOCE</td>
<td>46/51</td>
<td>90%</td>
</tr>
<tr>
<td>CLD</td>
<td>47/51</td>
<td>92%</td>
</tr>
</tbody>
</table>

(Note: I = the number of weak-form words with written transcription; II = the total number of weak-form words in this study; III = the coverage rate of weak-form written transcriptions)

From Table 1, it can be clearly seen that the coverage rate of weak-form words with written transcription can be roughly divided into three levels, among which CLD and LDOCE belong to the first level with a higher coverage rate of 92% and 90% respectively, OALD the second level with a medium coverage rate of 86%, and COBUILD and MD the third level with a lower coverage rate of 78% and 75% respectively. In CLD and LDOCE, there are four weak-form words (me, onto, was, were) and five weak-form words (any, by, do, will, would) that are provided with no weak-from written transcriptions; in OALD, seven weak-form words (any, or, by, does, is, will, there) lack weak-from written transcriptions; in COBUILD and MD, the number of weak-form words that is short of weak-from written transcription rises to eleven (any, his, your, him, you, by, into, onto, per, is, will) and thirteen (any, we, or, by, onto, been, do, is, should, was, were, will, would) respectively. In other words, if one learner wants to consult the Big five about the weak-form transcriptions of the fifty-one weak-form words, he is relatively ensured with a higher availability in CLD and LDOCE, and a lower availability in COBUILD and MD, and an in-between availability in OALD.

The crux of the matter behind the simplistic statistics is that what the norms are for lexicographers to supply weak-form words with written transcriptions? Apparently, no agreement has been reached in the same learner’s dictionary, let alone the agreement among the Big five. For example, in CLD auxiliary verbs such as am, is, are, be are provided with their corresponding weak forms while was and were are merely supplied with strong forms; prepositions to and into are supplied with both strong forms and weak forms while onto is only supplied with its strong form. Furthermore, among the Big five, there exists no total conformity since there is always at least one weak-form transcription that is included in one learner’s dictionary but excluded from
another. For instance, the weak-form transcriptions of *me, onto, was*, and *were* excluded from CLD, but are included in both OALD and LODCE; the weak-form transcriptions of *by* and *any* are excluded from MD, COBUILD, OALD and LODCE, but are included in CLD.

### 3.1.2 The coverage rate of weak-form audio recordings

In terms of the OELD, it is customary to supply the written transcriptions of headwords with accompanying audio recordings. The number of weak-form written transcriptions and weak-form audio recordings are counted respectively by artificial statistics. Table 2 shows the coverage rate of weak-form audio recordings in the Big five.

<table>
<thead>
<tr>
<th></th>
<th>I / II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>OALD</td>
<td>60/60</td>
<td>100%</td>
</tr>
<tr>
<td>MD</td>
<td>48/48</td>
<td>100%</td>
</tr>
<tr>
<td>COBUILD</td>
<td>41/42</td>
<td>≈100%</td>
</tr>
<tr>
<td>LDOCE</td>
<td>0/65</td>
<td>0</td>
</tr>
<tr>
<td>CLD</td>
<td>0/71</td>
<td>0</td>
</tr>
</tbody>
</table>

(Note: I = the number of weak-form audio recordings; II = the number of weak-form written transcriptions; III = the coverage rate of weak-form audio recordings)

From Table 2, it can be found that OALD, MD, and COBUILD basically stick to the principle of ‘one sound for one transcription’ when coping with pronunciations of weak-form words. Specifically, sixty weak-form written transcriptions in OALD and forty-eight weak-form written transcriptions in MD are all supplied with their accompanying audio recordings; out of the forty-two weak-form written transcriptions in COBUILD, forty-one possess the audio pronunciations\(^{10}\). These audio recordings are usually invoked by clicking the loudspeaker icons which are placed either before the written transcriptions (as in MD) or after them (as in OALD and COBUILD). Nevertheless, the sixty-five weak-form written transcriptions in LDOCE and the seventy-one in CLD are deprived of their accompanying weak-form audio sounds, with their strong-form audio recordings merely provided. Specifically, in LDOCE each weak-form word is supplied with two audio recordings, among which the red loudspeaker icon stands for the British pronunciation while the followed blue one represents the American pronunciation. The scenario, in CLD, is oddly similar to that in LDOCE, with the exception that the American pronunciation is placed before the British pronunciation. But one noteworthy fact the two shares is that both the British pronunciation and the American pronunciation are totally strong forms.

\(^{10}\) In COBUILD, the weak-form audio recording of *would* is missed, which may result from the carelessness of the compliers.
### 3.1.3 The average number of weak-form variants

In the OELD, the weak-form word is usually provided with more than one weak-form written transcriptions. For instance, in OALD, the weak-form word is furnished with three weak-form written transcriptions, /ənd/, /ət/, also /n/, so the number of weak-form written transcriptions can be regarded as an indicator of the weak-form variants to some extent. The statistics in Table 3 reveals more details of the weak-form variants included in the Big five separately.

**Table 3** The average number of weak-form variants in Big five

<table>
<thead>
<tr>
<th></th>
<th>I / II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>COBUILD</td>
<td>42/40</td>
<td>1.05</td>
</tr>
<tr>
<td>MD</td>
<td>48/38</td>
<td>1.26</td>
</tr>
<tr>
<td>OALD</td>
<td>60/44</td>
<td>1.36</td>
</tr>
<tr>
<td>LDOCE</td>
<td>65/46</td>
<td>1.41</td>
</tr>
<tr>
<td>CLD</td>
<td>71/47</td>
<td>1.51</td>
</tr>
</tbody>
</table>

(Note: I = the number of weak-form written transcriptions; II = the number of weak-form words with weak-form written transcriptions; III = the average number of weak-form written transcriptions per weak-form word)

From Table 3, it is notable that each of the Big five has recorded somewhat common weak-form variants by itself, and the average number of weak-form variants per weak-form word varies slightly. Among the Big five, CLD is the top one to include weak-form variants, which equips per weak-form word with 1.51 weak-form variants on the average. One of its striking features is that it offers users abundant weak-form variants of the auxiliary verbs. Specifically, it provides thirty-one weak-form variants for seventeen auxiliary verbs\(^\text{11}\), the average number of which is evidently greater than that of the other four dictionaries. By contrast, at the other extremity, COBUILD seems to be relatively conservative since it recommends merely one weak-form pronunciation to the EFL learners except for the two weak-form words *the* and *to*, both of which are transcribed with two weak-form pronunciations respectively. The remainder of the Big five, LDOCE, OALD, and MD, fall in between the ‘two extremes’ with an average number of weak-form variants 1.41, 1.36, and 1.26 respectively for each weak-form word.

\(^{11}\) The seventeen auxiliary verbs are *am, are, be, been, can, could, do, does, had, has, have, must, shall, should, was, were, would.*
3.2 The depth of weak-form representation in OELD

We have conducted a full investigation into the availability of the weak-form pronunciation notes in the Big five, however the findings are rather frustrating since most of the weak-form words are lacking in any weak-form usage. More specifically, CLD and MD fail to provide any usage note of weak-form pronunciation, while OALD, LDOCE and COBUILD merely supplies two (to, into), four (the, into, onto, to) and six (the, to, had, has, must, there) weak-form words with a simple weak-form usage respectively. Taking the two weak forms of to for example, on what occasion should the learners choose /tə/, and on what any other occasion should they choose /tu/? Table 4 shows the weak-form representation of to in Big five.

Table 4 The weak-form representation of to in Big five

<table>
<thead>
<tr>
<th></th>
<th>to</th>
</tr>
</thead>
<tbody>
<tr>
<td>OALD</td>
<td>/brə/ /tə/ ; name /tə/ ; be before vowels /tu/ ; name before vowels /tju/ ;</td>
</tr>
<tr>
<td>LDOCE</td>
<td>/tu/ ; before vowels tu ; strong tu /</td>
</tr>
<tr>
<td>COBUILD</td>
<td>PRONUNCIATION NOTE: Usually pronounced (ts /) before a consonant and (tu /) before a vowel, but pronounced (tu: /) when you are emphasizing it.</td>
</tr>
<tr>
<td>CLD</td>
<td>US /tə/ UK /tu/ strong /tə/ /tə/ weak /tə/</td>
</tr>
<tr>
<td>MD</td>
<td>ADVERB, PREPOSITION strong /tu/ ; weak /tu/ /tu/ ; /tə/ ; /tə/</td>
</tr>
</tbody>
</table>

From Table 4, it can be found that the weak-form representation of to can be roughly divided into three levels according to the detailedness of weak-form pronunciation notes. The CLD and MD fall into the first level which provides no pronunciation tip for the two weak forms of to. The EFL learners may be left in the dark when they need to make a choice between the two weak-form pronunciations. The OALD and LDOCE belong to the second level which offers pronunciation tip for one of the weak forms merely by adding that /tu/ is usually used before a vowel, while the usage of the other weak form is implicit which implies that the learners should be able to deduce that /tə/ usually appears before consonants. The third level is represented by the COBUILD, which not only provides pronunciation notes for the two weak forms of to, but also refers to one of its strong form usage (used for emphasis). On the whole, the depth of weak-form representation in OELD is oversimplified, which is merely the copy of its hard-paper counterpart, and there is still a long way to go to deepen the weak-form representation in OELD by taking full advantage of the internet lexicography.

3.3 The consistency of weak-form representation in OELD

3.3.1 The precedence relationship between weak form and strong form

In the practice of dictionary-making, the precedence relationship between weak form and strong form refers to which pronunciation should come first when it comes to the
arrangement of the two sorts of pronunciations. As for the phoneticians, they seem to have reached an agreement that the weak forms are more normal than strong forms in connected speech while the strong forms are considered to be rare, used occasionally to indicate some special meaning (Gimson 1980; Zhang 2002, 2007; Roach 2008; Cruttenden 2014). Gimson (1981:259) put it more bluntly that “it is clearly misleading to give the (strong) citation form for such words since in ordinary discourse the weak form is used in over 95% of occurrences.” As to the lexicographers, however, their opinions diverge from each other concerning the order of weak forms and strong forms in dictionaries. Table 5 demonstrates how the Big five grapple with the issue.

Table 5 The precedence relationship between weak form and strong form in Big five

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Weak Form</th>
<th>Strong Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>OALD</td>
<td>BrE /bat/</td>
<td>NAmE /bat/</td>
</tr>
<tr>
<td>LDOCE</td>
<td>but /bat/</td>
<td>strong /bat/</td>
</tr>
<tr>
<td>COBUILD</td>
<td>but for learners of English (but /bat/)</td>
<td></td>
</tr>
<tr>
<td>CLD</td>
<td>conjunction /UK STRONG /bat/ WEAK /bat/ US STRONG /bat/ WEAK /bat/</td>
<td></td>
</tr>
<tr>
<td>MD</td>
<td>ADVERB, CONJUNCTION, PREPOSITION strong /bet/</td>
<td></td>
</tr>
</tbody>
</table>

As is shown in Table 5, the arrangement of weak forms in Big five can be divided into two groups: the first group gives the weak form first with OALD, LDOCE, and COBUILD as its memberships; the second group, on the contrary, put the strong form in the first place with CLD and MD falling into it. For the former group, the weak-form pronunciation is regarded as unmarked, which is the normal and usual form that does not need the label of ‘WEAK’ or ‘weak form’, while the strong form is considered to be marked, which is the rare or occasional form with the label ‘strong form’ or ‘STRONG’ to show its special function (e.g. for emphasis). As to the latter group, considering the ontological pattern of lexicography, the dictionary is often considered a reference tool in the form of wordlist or wordbook, in which all entries “are looked at individually and separately and are not treated as an integrated whole” (Yong and Peng 2007: 4). This traditional approach has dominated dictionary making for thousands of years. Under the influence of this view, it makes sense that lexicographers give more priority to the citation forms in isolation (i.e. strong forms) when it comes to recording pronunciations in dictionaries. And this is also the practice carried into by CLD and MD. Therefore, the divergence between the two groups is somewhat objectively due to different philosophies relating to linguistics and lexicography.

However, when it comes to one specific dictionary, such divergence should be avoided in order to keep the stylistic rules of weak-form representation consistent. Yet after a closer examination of the precedence relationship between weak form and strong form in each of the Big five, one can still find a fair number of inconsistencies.
Owing to space limitation, this part merely takes OALD for example, in which the inconsistencies concerning the arrangement of weak forms and strong forms are listed as follows (see Table 6)

Table 6 inconsistencies of the arrangement of weak-form transcriptions in OALD

<table>
<thead>
<tr>
<th>Sample</th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
<th>Case 4</th>
<th>Case 5</th>
<th>Case 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>but conjunction</td>
<td>been</td>
<td>his determiner</td>
<td>your determiner</td>
<td>would modal verb</td>
<td>that conjunction</td>
<td>onto proposition</td>
</tr>
<tr>
<td>BrE /bʌt:/</td>
<td>BrE /bɛn/</td>
<td>BrE /hɪz/</td>
<td>BrE /jɔr/</td>
<td>BrE strong form /wud/</td>
<td>BrE /ðæt/</td>
<td>BrE /ɒn/</td>
</tr>
<tr>
<td>NAmE /bʌt/</td>
<td>NAmE /bɛn/</td>
<td>NAmE /hɪz/</td>
<td>NAmE /jɔr/</td>
<td>NAmE strong form /wud/</td>
<td>NAmE /ðæt/</td>
<td>NAmE /ɒn/</td>
</tr>
</tbody>
</table>

As can be seen from Table 6, in terms of the arrangement of weak-form transcriptions, all the six cases are inconsistent with the conventional way of ordering the weak form and strong form (as the ‘Sample’ but) in OALD to some extent. Specifically, the inconsistencies are listed as follows: (1) the first four Cases (from Case 1 to Case 4) give the strong form first which departs from the commonly accepted norm in OALD that the strong form is of secondary place; (2) the first three Cases (from Case 1 to Case 3), together with Case 5, the ‘strong form’ labels that should be placing before the strong-form transcription are all missing; (3) as to Case 3 and Case 4, the “weak form” labels are redundant; (4) the punctuation between the weak form and the strong form is always a semicolon while that in Case 1 is a comma; (5) the strong form pronunciation in Case 6 is left out though the strong form /ˈpʌnɪtə/ is only rarely heard. This kind of inconsistencies can also be found in CLD, MD, and LDOCE. If there are any linguistic or lexicographic motivations for the above hard-to-explain variations, it would be certainly too subtle for learners to discern.

3.3.2 The sequence of different weak-form variants

As mentioned above in section 3.1.3, each of the Big five has provided some kind of weak-form variants for some weak-form words, but the order of these variants varies from one dictionary to another. Taking the four weak-form words her, and, to, the for
instance, the arrangement of their weak-form variants in the Big five are listed respectively in Table 7.

### Table 7 The sequence of different weak-form variants in Big five

<table>
<thead>
<tr>
<th></th>
<th>her</th>
<th>and</th>
<th>to</th>
<th>the</th>
</tr>
</thead>
<tbody>
<tr>
<td>OALD</td>
<td>BrE /hə(r)/; /b/ (r)/; /ə(r)/</td>
<td>BrE /ənd/; /ən/; also /n/</td>
<td>BrE /tʃ/; before vowels /tu/</td>
<td>BrE /dʒ/ /dʒ/</td>
</tr>
<tr>
<td>CLD</td>
<td>WEAK /hə,ə/</td>
<td>WEAK /ənd,ən/</td>
<td>WEAK /tʃ, ə/</td>
<td>WEAK /dʒ, ə/</td>
</tr>
<tr>
<td>LDOCE</td>
<td>/ə, hə/</td>
<td>/ənd,ən/</td>
<td>/tʃ; before vowels /tu/</td>
<td>/ð ə ; before vowels /dʒ/</td>
</tr>
<tr>
<td>MD</td>
<td>weak /ə(r)/ /hə(r)/</td>
<td>weak /ən/ /ənd/</td>
<td>weak /tʃ, ə/</td>
<td>weak /dʒ, ə/</td>
</tr>
<tr>
<td>COBUILD</td>
<td>/hə/</td>
<td>/ənd/</td>
<td>/t ə / before a consonant and /tu/ before a vowel</td>
<td>/ðə/ before a consonant and /dʒ/ before a vowel</td>
</tr>
</tbody>
</table>

It can be found from Table 6 that there are mainly two factors relating to the arrangement of weak-form variants in the Big five. The first one is the length of written transcription, i.e. the weak-form variants can be arranged by their written transcription length, either in a descending order or in an ascending order. For example, both OALD and CLD sort the weak-form variants of *her* and *and* according to their written transcription length in a descending order, while MD sort them in an ascending order. However, it seems that LDOCE adopts a mixed method, with *her* in an ascending order while *and* in a descending order. The second factor is relevant to the distribution of weak-form variants represented by *to* and *the*. Among the Big five, nine out of ten put the weak-form variant appearing before a consonant on the first place while put the one appearing before a vowel on the second place, with CLD as the only exception. From an overall perspective, the occurrence of the weak-form variants is in a complementary distribution and highly context-dependent, and it is hard to figure out which one is more frequently used. And that is probably why the arrangement of weak-form variants in the Big five differs from each other when taking into consideration the above two factors.

### 3.3.3 The matching rate between written transcriptions and audio recordings

“As quite often happens in all e-dictionaries having both modes of phonetic representation, some entries exhibit a mismatch between transcription and recording” (Sobkowiak 2008:66-67). From section 3.1.2, it is known that among the Big five the most conspicuous discrepancy between the transcriptions and audio sounds could hardly be wider than that in LDOCE and CLD, because all the weak-form transcriptions in the two dictionaries are devoid of any audio pronunciation. Then, we will move on to at this point going into further details of the alignment among the
remainder of the Big five, i.e., the MD, OALD, and COBUILD, in which both the transcriptions and their accompanying audio recordings of the weak forms are reasonably complete. By clicking the loudspeaker icons beside the transcriptions one by one, we listened to all British recordings of the weak forms within the three dictionaries and checked whether they match well with their transcription counterparts or not. The checking results are listed in Table 8.

Table 8 The matching rate between weak-form written transcriptions and their accompanying recordings

<table>
<thead>
<tr>
<th></th>
<th>I / II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD</td>
<td>32/48</td>
<td>67%</td>
</tr>
<tr>
<td>OALD</td>
<td>58/60</td>
<td>97%</td>
</tr>
<tr>
<td>COBUILD</td>
<td>42/42</td>
<td>100%</td>
</tr>
</tbody>
</table>

(Note: I = the number of weak-form audio recordings at variance with the written transcriptions; II = the total number of weak-form audio recordings; III = the matching rate between written transcriptions and their accompanying recordings)

Table 8 shows that the matching rate in MD is merely 67% and as many as one-third of the weak-form transcriptions (sixteen) are inconsistent with their accompanying audio recordings. The major mismatched problems include: (1) The weak-form transcriptions of his, some, him, me, for, and, as, but, for, as well as the first weak-form transcriptions of he, her, have, are supplied with strong-form audio pronunciations. (2) The two weak-form audio recordings of into, you are reversed. If the users display the loudspeaker icon of /'ʃɪntə/ and /jʊə/ (the first weak-form pronunciation of into and you respectively), they will hear the pronunciations of /'ʃɪntə/ and /jʊə/ (the second weak-form pronunciation of into and you respectively), and vice versa. Moreover, the chances of mismatching between the weak-form transcriptions and their accompanying audio recordings have been reduced a lot in OALD, with only two inconsistencies being found: the British weak-form audio recordings of me and you are mispronounced as their parallel strong forms. Lastly, no discrepancy has been found between the forty-two weak-form transcriptions and their accompanying recordings in COBUILD.

4. Suggestions for a better weak-form representation in OELD

4.1 Suggestions for the breadth of weak-form representation

First and foremost, as to the number of weak forms in the OELD, what the lexicographers should always bear in mind is the targeted users of their dictionaries. That is to say, the dictionary-makers should make a distinction between the
dictionaries aiming at native speakers and those for the EFL learners. Taking Jone’s (1962) weak-form wordlist aiming at native speakers for example, the weak forms of such words on the list as *ma’am, many, saint, Sir, till, time, not*, etc., are not included in the OELD. It makes sense for the compliers to do so since it is too subtle for EFL learners to notice the differences between weak forms and strong forms of those words. By contrast, the wordlists of most common weak-form words worked out by O’Connor (1980), Roach (2008) and Zhang (2002, 2007) are more practical and suitable for the compliers to follow in the practice of pedagogical lexicography because they picked out the most basic weak-form words for the EFL learners. However, it is notable that the lexicographers should make some kind of supplement to the core weak-form wordlist out of the principle of systematization. For example, if *to* is provided with the weak forms, it would make no sense to exclude *into* and *onto* from the weak-form wordlist.

Second, in terms of the number of weak-form variants that is going to be included in the OELD, the lexicographers are supposed to keep a balance between the two extremes that ‘the more the better’ and ‘the less the better’. Taking *and* for example, both Jones (1997) and Cruttenden (2014) introduced six weak-form pronunciations for the native speakers, but there are some differences between them. Jones (1997) listed *ænd, æn, nd, n, m, n* in his English Pronouncing Dictionary, while Cruttenden (2014) introduced *æn, ən, ʊ, n, ənd, ðən* in his phonetic work without specific usage notes. Even the native speakers cannot tell exactly the differences among the six weak forms of *and*, as well as the differences between Jone’s scheme and that of Cruttenden’s. Under such circumstance, it is not wise to stick to the principle of ‘the more the better’. On the contrary, the principle ‘the less the better’ may be suitable to the beginners, when only one weak form is recommended. However, for the advanced learners, a balance is needed between the two extremes. For the above instance, it would be wise to include *ænd, æn, n, or ənd, ən* in the learner’s dictionaries since these chosen weak variants basically cover the most common weak-form usages of *and*, whether it is in slow, careful speech or in more rapid speech.

Third, as for the scope of audio recordings, although theoretically speaking, there are three possible ways to combine the written transcription and audio recording in the OELD, i.e. written transcription only, audio recording only, written transcription plus audio recording, it is predictable that supplying each weak-form written transcription with its audio counterpart would be a good choice for the EFL learners, the practice of which has been carried out by OALD, MD and COBUILD. By contrast, both LDOCE and CLD merely provide audio recordings for the strong-form transcriptions without any conspicuous motivations. Therefore, considering either from the perspective of comprehensiveness or learnability, it is wise for the OELD to include both the weak-form transcriptions and their accompanying recordings.

4.2 Suggestions for the depth of weak-form representation

To improve the depth of weak-form representation is the most important part in OELD because it is closely pertinent to the learnability of the learner’s dictionary. In the
present study, the learnability of the weak forms mainly refers to the capability of a learner’s dictionary product to enable its (potential) users to learn how to make full use of the weak forms effectively, which basically involves three didactic aspects: the usage of the weak-form pronunciation, the usage of the strong-form pronunciation, and pertinent exercises of both the weak form and strong form pronunciations. Taking that for instance, a tentative scheme is devised as follows, see Figure 1.

**Figure 1** A tentative scheme to improve the depth of weak-form representation

To begin with, as for the EFL learners, all the weak-form pronunciations should be provided with corresponding usages rather than just make a list of the weak forms without any instruction. This can be easily achieved in OELD with the aid of internet technology. Taking the weak-form pronunciation of *that* for example, the usage of weak form /ðæt/ is represented as follows, see Figure 2.

**Figure 2** The usage of weak-form pronunciation /ðæt/

In Figure 2, two usages of the weak-form /ðæt/ are introduced to the EFL learners, and each of them is accompanied with several most common examples in order to strengthen their understanding of this phenomenon. Compared with the paper dictionary which used to be called the ‘silent teacher’, the OELD should be turned into a ‘talking teacher’, in which “not only the phonetic representation of the headword
would be carefully thought over, but the entire entry would likely receive phonolexicographic attention" (Sobkowiak 2007: 149). Therefore, all the examples are supplied with audio recordings so that the learners can check the weak-form pronunciations in specific context.

Furthermore, the strong forms deserve at least as equal attention as, if not more than the weak forms, because the strong forms usually carry special meanings in ordinary colloquial speech and the strong-form rules are more regular. In Figure 3, three typical strong-form usages of *that* are listed with examples and their accompanying recordings.\(^{12}\)

![Figure 3 The usage of strong-form pronunciation /ðæt/](image)

In Figure 3, three strong-form usages of *that* are introduced with specific illustrations and recordings. It can be seen that the part of speech is a crucial factor connecting to the pronunciation of *that*. Needless to say, the pronunciation notes of strong forms are necessary and beneficial complements to those of weak forms. It is important for the EFL learners to remember that there are certain contexts where only the strong forms are acceptable and others where the weak forms are the normal pronunciation. Finally, considering the (potential) negative effects on the EFL learning process brought by the weak-form pronunciations, the design of pertinent exercises could revolve around two aspects: listening comprehension and oral production. Figure 4 and Figure 5 show these two aspects respectively.

---

\(^{12}\) In the examples, *that* is capitalized in order to show its marked strong-form pronunciation.
Figure 4 Listening comprehension exercises relating to weak form and strong form of *that*

As shows above in Figure 4, the sentence *John thinks that man is evil.* can be interpreted differently depending on the pronunciations of *that*. In the listening comprehension exercises, two audio recordings of the sentence are provided, one with the weak-form pronunciation of *that* while the other with strong-form pronunciation. Then the learners are required to match the appropriate meaning with the recording. When *that* pronounces its weak form /ðæt/ as a conjunction, the sentence means “John thinks all humans are evil”; yet when it pronounces the strong form /ðæt/ as a pronoun in deliberate speech, the sentence means “John thinks a specific (male) individual is evil.” For the EFL learners, this kind of receptive exercises can be used to work towards recognition of the difference between weak form and strong form, and help to raise awareness of weak-form pronunciations.

Productive exercises can also be used to help learners towards their target of weak-form pronunciations. Both weak forms and strong forms can be isolated and drilled on their own, before being put back into the sentence or utterance. In order to enhance the contrast between the weak form and strong form, a short paragraph sample relating to *that* is written by the author in Figure 5 with a mixture of usages of weak form and strong form.
As shown in Figure 5, the capitalized *THAT* represents the strong-form pronunciation while the lowercase *that* stands for the unmarked weak-form pronunciation. The EFL learners can practice their pronunciation by operating the buttons on the multimedia control panel, on which the four function buttons are *Play*, *Record*, *Compare & Visualize*, and *Save*. The learners can first choose to hit the *Play* button, by listening to the audio recording they would become familiar with the pronunciation features of *that*; then hit the *Record* button, just recording their own readings, the system would be devised to make an automatic comparison between the original reading and the learners’ self-recordings. What the learners need to do is press the button of *Compare & Visualize*, with the comparison results being visualized. Any feedbacks from the comparison could be saved by the learners as long as they would like to. Nowadays, the visual representation of the rhythm has been realized by various acoustics software, yet it has not been applied into the field of internet lexicography. Although the above-mentioned remains an assumption, which may be a dream for the lexicographers, it is still meaningful because of the human-oriented spirit it has reflected. Sobkowiak (2007:147) put it well when he said “this (‘this’ refers to pronunciation practice) limitation, however, is not one of technology or pedagogy, but of imagination”. I wish that with this imagination in mind, the above assumption could be realized soon in the near future.

### 4.3 Suggestions for the consistency of weak-form representation

Figuratively speaking, if the dictionary is compared to a work of art, any minor inconsistency found in it would be a disappointing flaw. Considering the sequential consistency between weak forms and strong forms, the present study sides with the practice that the weak forms should be given the first place based on their frequency of usage in connected speeches. Admittedly, there are a small number of common
weak-form words, such as *by, any, or*, etc., of which the weak forms are occasionally used. As for these exceptions, an extra specification is necessary. Taking *by* for instance, its weak-form representation can be represented as follows: **by Weak forms (occasional use): /ba/ /bæ/; Strong form: /ba/ /bæ/**. Thus, the sequential relation between weak forms and strong forms can be kept consistent in the practice of dictionary-making.

Next, in terms of the sequential consistency among different weak-form variants, two factors also needs to be considered: one is the shape of transcription symbol, as the two weak-form variants of *to* /tə/ and /tu/, which are in a complementary distribution according to the phonological rules. If the usage frequency cannot help, then the shape of transcription symbol may come to a good choice in deciding which weak-form variant should come first to the availability of the users. The order of /tə−tu−tu:/ seems to be more natural than the order of /tu−tə−tu:/ since /tu/ might play a role of transition from /tə/ to /tu:/ This method could also be applied to the weak-form variants of *into*, *onto*, and *the*. The other factor is the length of written transcription as in the case of *and*: the arrangement of /ænd−ən−n/ appears to be acceptable than that of /n−ən−ænd/, though “in slow and careful speech /ən/ is more often than /ænd/” (Jones, 1997: 20). The reason lies in that the order of /ænd−ən−n/ displayed directly the relationship between the length of transcription and the speed of speech, i.e., /ænd/ usually in a slow and careful speech, /ən/ in a rapid speech and /n/ in a more rapid speech. Similarly, ordering the availability precedence of different weak-form variants, such as that of *have, had, has, must, her, him*, etc., might be the same as this way.

Lastly, as to the consistency between transcriptions and recordings, although the phonological inconsistency problem is a much wider issue, not specifically confined to electronic versions of EFL dictionaries, the mismatches between weak-form transcriptions and their accompanying recordings will unavoidably produce a lot of confusion in the unsuspecting user. “The cause must clearly be sloppy proofreading, or rather proof listening, a job requiring advanced phonetic skills, a good ear, plenty of concentration and patience” (Sobkowiak 2008:67). The compliers shoulder great responsibility to provide the right audio pronunciations for the learners.

To sum up, among different dictionaries, there may exist some inconsistencies due to different theoretical or practical considerations. However, when it comes to one specific dictionary, such inconsistencies should be avoided in order to keep the stylistic rules of weak-form representation consistent. Therefore, “it goes without saying that, whatever method of indicating pronunciation is adopted, it should be consistently and correctly used in the body of the work” (Abercrombie 1978:124).

5. Conclusions

If this picture of the weak-form representation in the OELD, sketched on the basis of the most common weak forms from the Big five, is at all close to empirical truth, a number of conclusions seem to follow. First, the incompleteness of the weak-form representation in the OELD mainly reflects in two aspects: one is the incompleteness
of the written transcriptions, and the other is the incompleteness of the audio recordings. Second, the weak-form representation in the OELD is oversimplified while its learnability is not highlighted. Third, the treatment of phonological consistency relating to the weak-form representation seems haphazard. The deficiencies pointed out in this study fall into two categories: some are simply due to careless errors of the compliers; others may ultimately result from the lack of unified standards for representing weak-form words in the OELD, leaving open a series of issues, such as the scope of weak forms with written transcriptions and their accompanying audio recordings, the number of weak-form variants, the arrangement of weak-form variants, the precedence relationship between weak forms and strong forms, etc., in the big five.

As has been suggested throughout the paper so far, the weak-form representation should be seen as an integral part of what goes on in the OELD, and it is important that lexicographers treat it as such. With this in mind, it is urgent for lexicographers to grapple with the above issues by laying more emphasis on the so-called ‘secondary business’. However, the present study should not be too good to be left on the theoretical shelf. For the future research, more priorities should be placed on related empirical studies, such as the difficulties of acquiring weak forms that EFL learners may encounter, the effectiveness of weak-form representation in the OELD, the technical realization of weak-form representation in the OELD, etc. Although some suggestions aiming at improving the learnability of the weak-form representation in the OELD, appears to be a lexicographic dream currently, the dream would soon come true under a joint effort by “lexicographers, linguists, phoneticians, pedagogues, copy editors, programmers, database and sound engineers, recording speakers, and a host of other specialists” (Sobkowiak 2003:439).

References

A. Online dictionaries.

OALD: http://www.oxfordlearnersdictionaries.com/
LDOCE: http://www.ldoceonline.com/
CLD: http://dictionary.cambridge.org/
MD: http://www.macmillandictionary.com/
COBUILD: https://www.collinsdictionary.com/

B. Other literature.


Dictionary User Research in the Digital Era: Value of User-generated Content

Yuyan He
Center for Lexicographical Studies, Guangdong University of Foreign Studies
yuyan_he12@163.com

Abstract

This paper explores what user-generated content (UGC) can contribute to dictionary user research, with analysis of data collected from digital platforms (WordReference.com, UsingEnglish.com, amazon.com, dangdang.com, zhihu.com and douban.com) where different groups of Chinese EFL learners gather. In recent decades, ‘user presupposition’ (Wiegand et al. 2010: 680, cited after Müller-Spitzer (2014: 1)) seems to be the central point of every lexicographical process as investigations into dictionary use are gaining increasing interest of researchers in this field. Moreover, the overwhelming digital revolution inspires the emergence of ‘prosumer’ (Lew 2013). In this view, UGC is definitely worth observing for modern dictionary user research and although there exist some UGC-based studies, that of UGC by Chinese EFL learners proves to have a gap. In this paper, UGC includes language-related questions and dictionary-concerned comments, generated by distinct Chinese advanced EFL groups. Data are collected from six digital platforms (as mentioned above) and coded into corresponding concepts on the basis of grounded theory. Then coded data is subjected to content analysis and correlation analysis to find out what kind of language problems Chinese advanced EFL learners tend to encounter, what aspects of the mainstream learner’s dictionaries they think are worth being appreciated, and finally to figure out the connection of this two types of UGC. With these endeavor, this paper aims to test the value of UGC research by means of qualitative methods and discover the national distinct features of Chinese advanced EFL learners.

Keywords: user research, digital platform, user-generated content (UGC), Chinese advanced EFL learners

1. Introduction

With the dictionary compilation perspective shifting from editors to users, dictionary user research has been conducted since decades ago and seems inclined to remain to be a spotlight during the recent digital revolution. In addition, as a silent teacher, learner’s dictionaries keep playing an essential part in second language acquisition and discussion on learner’s dictionary improvement has never been intermitted since the first learner’s dictionary was born, especially in the field of EFL (English as a Foreign Language) dictionaries. To make a panoramic view of user research on English
learner’s dictionaries, positive aspects could of course be found in the previous studies, but research gaps, which has been pointed out by Lew (2011, 2016), Nesi (2013), Tarp (2009), Zhang Yihua (2015: preface), etc., remain extensive, such as indifference to qualitative studies and learner’s national distinctive features. Simultaneously, as the digital revolution emerges and continues thriving, we have been taken from one universe to another – from printed dictionaries to digital ones, from passive message recipients to compilers’ focus and from traditional surveys to cutting-edge research methods like log file analysis (e.g. Lew 2015; Lew and de Schryver 2014; Müller-Spitzer and Koplenig 2014). The digital upheavals in lexicography is evidently overwhelming.

As Lew (2013) puts it, the digital revolution inspires the emergence of ‘prosumer’, a blend of producer and consumer, which means dictionary users in the digital era are no longer passive recipients but active contributors to the creation and provision of dictionary contexts. In this view, UGC is definitely worth observing for modern dictionary user research and although there exist some UGC-based studies, that of UGC by Chinese EFL learners proves to have a gap.

Based on Tarp (2008: 44-53)’s lexicographical function theory that more attention should be paid to extra-lexicographical situations but not only to dictionary use circumstances, and drawing upon the previous study by Holdt et al. (2016), the present study intends to study learners as dictionary users about their authentic language dilemma they encounter in the procedure of language comprehension and production, as well as the learners’ impression on the existing dictionaries they use, with a combining method of qualitative and quantitative analysis. Three research questions for the current study are formulated as follows:

1) What language learning questions trouble Chinese advanced EFL learners and how much do they count?
2) What do Chinese advanced EFL learners think of the mainstream learner’s dictionaries they use and how much do they count?
3) This article assumes that dictionary components appreciated by the learners will lead to less related language questions. Hence, the last question is to figure out whether this assumption works or not in current lexicographical situations.

2. Literature Review

Dictionary user research have a history of more than half a century. Reviewing first the early articles and then the later monographs, three recurring topics of user research can be concluded as follows:

2.1 Pragmatism Turning of Dictionary Compilation

Philosophy of dictionary compilation has been turned over twice: firstly from prescriptivism to descriptivism, and secondly from descriptivism to pragmatism.
centered on the users (Xie Haijiang and Zhang Liping 2010). Therefore, relations between users, dictionaries and editors has been changed, as a result of which, dictionary user research appeared and has grown up to a certain extent.

With this pragmatism turning, dictionary user research was first conducted in the 1960s by Barnhart (1962) and this attempt was further appreciated by Householder (1967: 279) that ‘dictionaries should be designed with a special set of users in mind and for their specific needs’. Later in the 1970s, some more scholars pay attention to this field and user research progressed to some extent (e.g. Barone 1979; Tomaszczyk 1979), and subsequently, it developed quickly in 1980s (e.g. Hartmann 1987; Meara and English 1987; Nesi 1987; Tono 1989). With the seminar hosted by Euralex held in Oxford in 1991 discussing the direct observation of how a user consult a dictionary, user research started to fasten its international academic status. Moreover, the numerous monographs in 1990s (e.g. Benbow et al. 1990; Hulstijn and Atkins 1998; Nuccorini 1994; Tono 1998) witnessed the firm establishment of user research, especially in the prosperous field of lexicography for EFL learners. As the argument that dictionaries for foreign learners should not carry on with the compiling traditions of that for native-speak- ers has gained ground in L2 lexicography, the early monolingual learner’s dictionaries (e.g. Oxford Advanced Learner's Dictionary of Current English in 1974, Collins English Learner's Dictionary in 1974, Longman Dictionary of Contemporary English in 1978, etc.) specially emphasized the representation of morphology, grammar, meaning, examples, etc. for the learner’s sake both in language comprehension and production (e.g. Jain 1981; Rundell 1988).

Since dictionary user research begins to step towards a new era, both theoretical and practical lexicographers have been fully aware of the importance of user’s needs and expectation before, during and after dictionary compilation (Gouws, 2006). Therefore, although there hasn’t been a systematic theory of dictionary user research yet, scholars are endeavoring to solid its right of speech. For example, some classified user contribution into three major types: direct, indirect and accessory contribution, which have their multiple forms and implementation (Abel and Meyer 2013), and some appealed to distinguish user needs regarding daily communication from standardized language learning (Gouws 2006). Especially on methodology, a vast array of updated approaches (e.g. Bergenholtz and Johnsen, 2007; Holdt et al. 2016; Lew et al. 2017; Mēchura 2008) as well as stricter empirical samples (Benbow et al. 1990; Blachowicz and Others 1990; Tono 1989, 1998) can be found.

2.2 Interaction between User Research and Learner’s Dictionaries

In the recent 20 years, investigations of learner’s dictionaries have been thriving, especially those for English language learning, together with which, a considerable number of scholars have studied language learners’ needs and their dictionary use, as user research develops.

A number of studies show that, due to learners’ deficiency in dictionary reference skills, learner’s dictionaries are usually used insufficiently. Through a questionnaire investigation, Béjoint (1981) concluded that certain number of advanced EFL learners
tend to use dictionaries for decoding purpose and seldom notice the plentiful encoding information. Further studies confirmed that learners are uncertain about what a dictionary entry contain (Chan 2012; Cubillo 2002) and confused about how to deal with multiple entries (Nesi and Haill 2002; Neubach and Cohen 1988), and lexicographically relevant errors the learners make exhibit high frequency (Nuccorini 1994; Tseng 2009). Despite the learners’ own problems, scholars suggest language teachers systematically instruct students about how to consult a dictionary, and also call for adjustment or even innovation in dictionary design for the enhance of the full use of learner’s dictionaries (Lombard 1994; McKean 2000).

As the mainstream English language learner’s monolingual dictionaries take up a major part of dictionary market, which means a number of English language learners become users of these dictionaries, investigators have attached great significance to them since the first English learner’s dictionary was born. In the early research of Barone (1979), the monolingual learner’s dictionary, the 3rd edition of Advanced Learner’s Dictionary published in 1974, played a limited role in effective language learning. Cowie (1979) studied the treatment of collocations, idioms and polysemy in the design of English learner’s dictionaries and found that such information can’t fully satisfy the foreign learners’ expectation. Controlled vocabulary in learner’s dictionary definition, arrangement of the headwords for better accessing and appropriate example policies were also discussed (Cowie 1989; Herbst and Popp 1999; Marello 1987), which proves the importance of them in both language comprehension and production. However, influence of the former of the mainstream monolingual learner’s dictionaries seems superior to the later on learners’ language learning (Laufer 2011; Laufer and Hadar 1997; Nesi 1987). User research concerning monolingual learner’s dictionaries of other languages can sparsely be found, such as Slovene (Holdt et al. 2016).

With the influence of SLA theory, learner’s dictionary research with regard to learners’ SLA features currently focus a lot on learners’ production errors of vocabulary and grammar (mainly in English). Mainly based on learners’ writing material or error corpus like Chinese Learner English Corpus (CLEC), researchers combine error analysis with information representation in learner’s dictionaries, summarizing that dictionary should not simply provide grammatical notes for verbs, but also for nouns, adjectives and adverbs (Christianson 1997; Huang 1985; Koeze 1990). Moreover, the listing of synonyms is of great demand for language learners (Huang 1985; Laufer 2011; Ng 2016). Some scholars notice the effect of learners with different mother tongue demonstrating different characters of first language transferring on learner’s dictionary compilation, and appeal the innovation of dictionaries with national distinctive features (Zhang Yihua 2015: 111-128). Recently, paralleling the development of cognitive science, lexicographers also ask for information arrangement that can fit learners’ cognitive ability (Gao Jun, 2012).

2.3 Influence of Digital Revolution on Dictionary Research

Thanks to the galvanizing of digital revolution, it is much more difficult than ever to meet the expectations of dictionary users as dictionaries migrate to digital forms and
users can make much more active contributions to the compilation of dictionaries (Hult 2014). Three types of dictionaries influenced mostly by the digital technology are general monolingual dictionaries aimed at adult mother-tongue speakers; bilingual dictionaries for ‘big’ language pairs; and monolingual learner's dictionaries (Hanks 2010; Rundell 2015), which are the main targets of dictionary research. Interestingly, Hartmann (2005) highlighted that the mixing of two or more ‘pure’ dictionary types has become a growing trend with the development of electronic lexicography.

From a different angle, the identities or characteristics of the traditional dictionary users no longer remain the same in the digital era. Strictly speaking, many scholars (Hult 2014; Lew and de Schryver 2014; Worlfer and Müller-Spitzer 2016) held that the term of ‘dictionary user’ refers to the recipients of a dictionary. However, Lew and de Schryver (2014) argued that users of Web 2.0 are no longer ‘passive recipients of packaged content’, but can be redefined with the neologism ‘prosumer’ (a blend of producer and consumer), who are active contributors of the creation and provision of self-made content. Accordingly, research tools of log files analysis (Bergenholtz and Johnsen 2005 2007; Müller-Spitzer et al. 2015), user-generated contents mining (Muzny and Zettelmooyer 2013; Navarro et al. 2009; Zesch et al. 2008), simultaneous feedback (de Schryver and Prinsloo 2001) as well as real-time comments (Welker 2013) are rising, as promising supplement to the early methods of questionnaire, examination, and so forth.

2.4 Reflections on Previous Studies

From the above review, we can obviously see the promising future in the field of dictionary user research, especially studies on EFL dictionaries and learners, under the combinative assistance of digital technology. However, this article has also found some deficiencies in this field of studies: first, depth and width of quantitative research on dictionary users are far from enough in contrast with those of quantitative research, which means underlying causes behind the quantized data or visible phenomena have more or less been neglected; second, numerous monographs seldom pay attention to the linguistic and cultural background differences of learners as dictionary users, although some scholars (eg. Zhang Yihua 2015: 111-128) have noticed this problem. In addition, although user research based on Chinese or Chinese EFL learners can sparsely be found, the methodological approaches and research objects have not yet entered the digital era, limiting in traditional forms of test and questionnaires and in printed dictionaries.

3. Research Method

3.1 Material Selection and Data Crawling

Raw material for the coming steps of data settlement all comes from the internet.
3.1.1 Language-related questions.

This is the extra-lexicographical situations. This part of data is selected from two popular language forums, WordReference.com and UsingEnglish.com, where language learners from all around the world gather to seek opportunities to interact with native speakers about special learning difficulties. As to this kind of data, sample typicality and learner’s background should be taken into consideration. According to incomplete statistics, about 100 and 60 questions on average can respectively be seen on this two forums, which more or less proves the typicality of these online language communities. All the website instructions and interpersonal communication are in English. This assures that the language educational level of the questioners can be defined as advanced or at least pre-advanced. And then, information of the nationality and native language posed on the questioners’ personal profile can help us to identify Chinese learners among the global pool. In addition, there is no need to challenge the completeness and truthfulness of profile information, because, as to completeness, nationality and native language are two of the necessary messages the forums require a new register to fill out; and as to truthfulness, recognition of cultural identity could emerge when one locates in a multi-cultural crowd, and moreover, contents in the communicative areas generated by them show their national distinctive features directly or indirectly.

Selected questions are posed from February 1, 2017 to April 30, 2017, with a total of 1622. During this period, 387 Chinese learners were seen active. In order not to involve too much questions posed by the same learner – which may result in lower reliability for the study – the maximum questions asked by a single learner is controlled at the level of 5, and all these 5 queries are taken randomly.

3.1.2 Comments on mainstream dictionaries.

As for the intra-lexicographical situations for actual users, book comment areas of two top book e-commerce giants in China (amazon.com and dangdang.com), one Q&A community where a lot of educated Chinese gather (zhihu.com) and one most active online reading club in China (douban.com) are chosen for monitoring Chinese advanced EFL learners’ real-time comments on three mainstream advanced EFL learners’ dictionaries: Oxford Advanced Learner’s Dictionary (OALD-8), Longman Dictionary of Contemporary English (LDOCE-5), Collins COBUILD Advanced Learner’s English Dictionary (COBUILD-7). Reasons for choosing this three dictionaries are (1) the amount of comments about these three types of learner’s dictionaries rank at top three in China, and from the contents posted at the comment bars one can find a large number of Chinese advanced EFL learners own or have used two or three of them, which exhibits the representativeness of OALD, LDOCE and COBUILD among Chinese learner’s dictionary market, and moreover, the overlapping use of these three dictionaries contributes to the objectivity and comprehensiveness of their comments; (2) depending first on the quantity of the comments, this study tries to select the latest version of the three types of dictionaries so as to follow the most

681
current views of the actual users on the dictionaries. In this part, this study only focuses on users’ praise for the dictionaries with the purpose to set a foundation for the solution of the third reach question. Admittedly, the number of all the praise in the above four digital media is 6576, but those worth analyzing (detailed comments rather than general appreciation like ‘What a wonderful dictionary! It helps a lot.’) takes less than a half amount, and this is why up to four related websites are included – to ensure the sample size. After selection, the total number of praise for the three dictionaries is 2122.

3.2 Data Coding

A qualitative method guided by grounded theory (Strauss and Corbin 1998) is applied here. With an open attitude to the raw material – no control from prior theories – data should be coded into three levels, from open codes to axial codes and lastly to selective codes. Thus, language-related questions and dictionary-concerned comments are to be coded into abstract concepts in certain sequence. For the convenience of further analysis, this two kinds of material will share the same concepts in the level of axial coding and selective coding. And since the present study will finally end in dictionary-related issues, coding of comments about dictionaries is better go first so that the coded concepts can be straightly utilized for the categorization of language-related questions.

With the help of Microsoft Word2016 and Excel2016, 1673 open codes are figured out from the raw material, until no more codes can be further found. Moreover, for the purpose of ensuring reach validity, a group of comments (200 pieces) on *Macmillan English-Chinese Dictionary for Advanced Learners* are employed to check the saturability of open codes, and the result shows no more codes again. Due to the space limitation, only a few open codes and corresponding examples from the raw material are demonstrated in Table 1. As we can see, 12 axial codes and 3 selective codes have been abstracted from the initial open codes.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Comprehension</td>
<td></td>
<td>Definition</td>
<td>exhaustive definitions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>The full sentence definitions help me a lot in reading. The multi-meanings of a word are arranged in a user-friendly way. [COBUILD-7]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Collocation</td>
<td></td>
<td>I think the rich amount of collocations is the shining point of this dictionary, which is better than OALD. This information shows me many idiomatic expressions used by native speakers, which reduces the comprehensive dilemma I come across when communicate with foreigners</td>
</tr>
</tbody>
</table>
or watch American films. [LDOCE-5]
The colorized pictures in the back matter really satisfy me. I remember the detailed illustration of ‘sports’ helps me much when I have a free talk with my foreign basketball partners. [OALD-8]

2 Production Thesaurus beneficial discrimination of synonyms

It surprises me that this learner’s dictionary offer information similar to that of a thesaurus. The clear discrimination of synonyms gives me a great favor when I write something. [LDOCE-5]

(To be continued)

(Continued)

Example typical examples helping grammatically correct production

The abundant examples, including phrases, collocations and sentences offer an authentic context for EFL learners to intake the specific rules of how a word can be used. [COBUILD-7]

2 Production Pragmatic information useful pragmatic information for writing in appendixes

The exhaustive pragmatic information supplied in the appendixes benefits my writing a lot. [LDOCE-5]

Usage notes detailed lexical usage notes

Thanks to the usage notes, I realized what words to be used in what situation. [OALD-8]

Morphology wordfamily information

Detailed information of prefixes, suffixes and infixes can help me with vocabulary learning. [OALD-8]

Headword selection extensive coverage of headword selection

The headword selection covers an enough range. I can even find some rarely-used words in it, which can hardly be seen in other kinds of learner’s dictionaries. [LDOCE-5]

3 Information accessibility Neologism enough inclusion of neologisms

Great dictionary keeping pace with times! Neologisms like googlewhack, interactive whiteboard can surprisingly be found. [LDOCE-5]

Pronunciation multiple types of pronunciation

Variants of pronunciations are supplied here, which meets
Based on the second and third level of concepts coded from dictionary-concerned comments, 1386 language-related open codes are also abstracted into the same 12 axial codes and 3 selective codes. For the concept coding is space consuming and no new axial codes and selective codes are to be displayed, here just several examples and their corresponding open codes are showed as follows:

(1) [WordReference.com] I haven't been studying Spanish for a long time. <-----Example sentence added to post by moderator (Florentia52)----->
Is it correct to say so? Does it mean I have just started studying Spanish not long ago?
[open code: understanding of present perfect progressive tense → axial code: usage notes]
(2) [WordReference.com] To describe the performance of reading a piece of word with background music, what word should I use? The performers are adult artists. On some occasion, it's only their sound tracks, while other time it's the video of their actual performance on stage.

What's the difference between these two words?
[open code: which word to choose to describe? → axial code: thesaurus]
(3) [UsingEnglsih.com] I wish to know which of the following pictures can be called a memory disc. Can anyone tell me please?
(pictures are omitted by this article)
[open code: what a word refers to in the physical word exactly? → axial code: illustration]

3.3 Data Analysis

According to Table 1, Nvivo10 is employed to conduct content analysis to the language-related questions and dictionary comments separately. The procedure aims to calculate the frequency of these UGC on the basis of the axial codes above. Calculating results are as follows:
Table 2 Content Frequency Distribution of Chinese Advanced EFL Learners’ UGC

<table>
<thead>
<tr>
<th>Selective code</th>
<th>Axial code</th>
<th>Questions frequency</th>
<th>Proportion (%)</th>
<th>Comment frequency</th>
<th>Proportion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comprehension</td>
<td>Definition</td>
<td>153</td>
<td>11.04</td>
<td>427</td>
<td>25.52</td>
</tr>
<tr>
<td></td>
<td>Collocation</td>
<td>189</td>
<td>13.64</td>
<td>133</td>
<td>7.95</td>
</tr>
<tr>
<td></td>
<td>Illustration</td>
<td>81</td>
<td>5.84</td>
<td>70</td>
<td>4.18</td>
</tr>
<tr>
<td></td>
<td>subtotal</td>
<td>423</td>
<td>30.52</td>
<td>630</td>
<td>37.65</td>
</tr>
<tr>
<td>Production</td>
<td>Thesaurus</td>
<td>117</td>
<td>8.44</td>
<td>168</td>
<td>10.04</td>
</tr>
<tr>
<td></td>
<td>Example</td>
<td>63</td>
<td>4.55</td>
<td>161</td>
<td>9.62</td>
</tr>
<tr>
<td></td>
<td>Pragmatic</td>
<td>225</td>
<td>16.23</td>
<td>119</td>
<td>7.11</td>
</tr>
<tr>
<td></td>
<td>information</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Usage notes</td>
<td>351</td>
<td>25.32</td>
<td>147</td>
<td>8.79</td>
</tr>
<tr>
<td></td>
<td>Morphology</td>
<td>27</td>
<td>1.95</td>
<td>7</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>subtotal</td>
<td>783</td>
<td>56.49</td>
<td>602</td>
<td>35.98</td>
</tr>
<tr>
<td>Information accessibility</td>
<td>Headword selection</td>
<td>36</td>
<td>2.60</td>
<td>196</td>
<td>11.72</td>
</tr>
<tr>
<td></td>
<td>Neologism</td>
<td>54</td>
<td>3.90</td>
<td>77</td>
<td>4.60</td>
</tr>
<tr>
<td></td>
<td>Pronunciation</td>
<td>9</td>
<td>0.65</td>
<td>14</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>Encyclopedic</td>
<td>81</td>
<td>5.84</td>
<td>154</td>
<td>9.21</td>
</tr>
<tr>
<td></td>
<td>information</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>subtotal</td>
<td>180</td>
<td>12.99</td>
<td>441</td>
<td>26.37</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>1386</td>
<td>100</td>
<td>1673</td>
<td>100</td>
</tr>
</tbody>
</table>

For a clearer imagination of the frequency distribution connection between the two types of UGC, a curve graph can be drawn accordingly as Figure 1.

![Figure 1 Content Frequency Percentile Curve of Chinese Advanced EFL Learners’ UGC](image-url)
Figure 1 gives a visualized picture for the characteristics of UGC frequency distribution.

Besides some discrepant points – ‘definition’, ‘usage notes’, ‘pragmatic information’ and ‘headword selection’ – other frequency distribution points exhibit a relatively consistent tendency, which indicates that there might be a certain correlation between language-related questions and dictionary comments after the data coding and content analysis of the raw UGC. A correlation analysis is thus implemented for the above content analysis result with the help of SPSS23. The analysis result is displayed in Table 3.

**Table 3 Pearson Correlation of the Content Analysis Result of UGC**

<table>
<thead>
<tr>
<th></th>
<th>questions</th>
<th>comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>questions</td>
<td>Pearson Correlation</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>0.316**</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>78</td>
</tr>
<tr>
<td>comments</td>
<td>Pearson Correlation</td>
<td>0.316**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>78</td>
</tr>
</tbody>
</table>

notes: **. Correlation is significant at the 0.01 level (2-tailed).

Table 3 illustrates the result of correlation analysis ($r=0.316$, $p=0.005<0.05$), that after content analysis, questions and comments exhibits significant positive correlation at the 0.01 level.

4. Discussion

4.1 Features of Dictionary-concerned Comments

From Table 2, it is evident that proportion of appreciations for definitions rank first (25.52%) among the whole sample. On the one hand, this may delight the lexicographers because the very dictionary component that they endeavor to improve is exactly the definition. On the other, it to some extent indicates the main function of the mainstream learner’s dictionaries for Chinese EFL learners is comprehension assistance, which implies a certain gap between the lexicographers’ original intension (i.e. a learner’s dictionary should help learners with their language comprehension and production) and the users’ actual impression of the dictionaries. Further statistics of the comprehension proportion (37.65%) displays a higher level than that of production (35.98%), not to mention the overlapping of the function of some dictionary components, such as ‘examples’ which can also help with the comprehension.

Next, headword selection and thesaurus follows to account for a relatively higher proportion within the comments. Chinese EFL Learners’ expectations are satisfied by this two factors, which proves one of their very need for learner’s dictionary is about
vocabulary acquisition. However, dictionary information supplied for production draws relatively less attention of Chinese learners, such as examples, pragmatic information and usages note, which accounts for proportions of 9.62%, 7.11% and 8.79%.

Unfortunately, praise for the collocations – both lexical and grammatical ones which are abundantly used by native speakers and should better be used in chunks in order to change input into intake – doesn’t rank high. The reason may be: the arrangement of the collocations in the dictionaries is not that apparent, or Chinese learners’ attention to this information is not that enough. This article tends to stand by the latter on the basis of previous studies. Moreover, Chinese learners’ particular comments on collocations in the present sample focus more on their assistant for language input, rather than output.

Interestingly, Chinese learners’ relatively great needs for illustrations (4.18%) and encyclopedic information (9.21%) are beyond this article’s expectation, which may owe to the growing cross-cultural communication between China and the globe.

4.2 Discovery from Language-related Questions

It is obviously that what dilemma Chinese advanced EFL learners encounter most is language production, which accounts for a proportion of 56.49% among the current sample – much higher than that of comprehension (30.52%). In addition, if we have a close and concrete eye on the authentic questions posed by Chinese EFL learners, numerous problems on grammar will be discovered, especially those on preposition, articles, singular or plural form of abstract nouns, singular or plural for of be in complex sentences as well as perfect tenses. This may again be an alarm for the previous ways of second language teaching and learner’s dictionary compilation which take language forms and meanings as separate parts and spend a great deal of time for grammar guidance independently, and even so, language learners still come across plentiful grammar obstacles. This research results can to some degree inspire the further endeavor for studies from the perspective of cognition. Furthermore, spelling mistakes can often be seen with the UGC by Chinese learners, the reason for which can be attribute to the digital revolution – students especially advanced learners seldom write with a pen and depend too much on the automatic spelling checker on the computer.

In 4.1, discussion about collocation in the intra-lexicographical situations emphasizes language comprehension and it is also true in extra-lexicographical circumstances. As shown in Table 2, Chinese advanced EFL learners have a lot of language learning difficulties concerning collocation comprehension and hence decoding of a whole sentence or context is delayed. In fact, few Chinese learners pose questions concerning idiomatic collocation production, which raises a challenge for dictionary compilers about collocation representation.
4.3 Explanation of the Correlation of the Two Types of UGC

Here comes the interesting part. In order to seek out whether there is a certain relation between the intra-lexicographical expectations and extra-lexicographical needs of Chinese advanced EFL learners, a frequency percentile curve is drawn (cf. Figure 1) to set the foundation for further observation. As mentioned in the introduction, the assumption of the present study is that, there exists a correlation between language-related questions and praise for dictionaries, and it is a negative correlation, because it is common sense that the good quality of a product seldom leads to related problems.

At the first stage, explanations should be made about the processing of the data in Figure 1 and Table 3. Percentages of frequency rather than the original frequency are used for analysis, because (1) the total code numbers of the two samples are different and original frequencies of different axial codes display a great discrepancy which would have an negative effect on the data analysis, and hence the percentage processing of original data will contribute to the research validity; (2) this article initially explains the research results with proportion as in 4.1 and 4.2 and for interpretation consistency, it is reasonable that the following account goes the same way.

As shown in Figure 1, only the points of ‘definition’, ‘pragmatic information’ and ‘usage notes’ seen to accord with the assumption of negative correlation. The result of the Pearson correlation analysis confirms the assumption of correlation, but the direction is an opposite one – positive correlation with the coefficient of 0.316. From this result, we could draw the following inference. Firstly, definitions supplied in the mainstream learner’s dictionaries offer relatively enough assistance for the text decoding procedure of Chinese advanced EFL learners. However, their assistance for text encoding is far from enough. Secondly, more appreciation for the quality of a product doesn’t mean the less related problems of it, at least for the circumstance of a learner’s dictionary. Thirdly, praise and criticism of a dictionary are better taken into equal consideration together, for lexicographers can discover what the users emphasize and further ‘mine’ their needs.

5. Conclusion

From qualitative analysis to quantitative methods, the current study has discovered something new in the intra-lexicographical and extra-lexicographical situations, demonstrating the national distinctive features of Chinese advanced EFL learners (cf. section 4). Furthermore, this study has made an attempt to see if any value of the UGC can be found for dictionary user research, with the inspiration of previous investigations, and gives a positive answer to this question (ibid). However, limitations of the research also exist. To begin with, different from language-related questions, dictionary comments are not that specific, which could more or less confuse the researchers when coding them. Furthermore, due to time...
limitation, size of the sample could not be large enough. Finally, owing to the anonymity of online communication, detailed profiles of the learners who produce the UGC are difficult to know.

For future investigations, this article suggests more implementation of qualitative research although it is time-consuming and expensive, something unexpected will be possible to gain. Or, further exploration on the finding of this study that Chinese advanced EFL learners tend to have a demand for encyclopedic information and illustration in their dictionaries is worth a try.

References


WordReference.com https://forum.wordreference.com/forums/english-only.6/
UsingEnglish.com https://www.usingenglish.com/forum/
amazon.com https://www.amazon.cn/%E5%9B%BE%E4%B9%A6/dp/B00JKOL6KU/ref=sr_1_1?ie=UTF8&qid=1494852903&sr=8-1&keywords=%E6%9C%97%E6%96%87
dangdang.com http://product.dangdang.com/22489058.html

B. Other literature


117-142.
Bergenholtz, H., and M. Johnsen. 2007. Log files can and should be prepared for a functionalistic approach. Lexikos17(1), 1-20.
Gouws, R. H. 2006. Discipline, dilemmas, decisions and data distribution in the planning and compilation of monolingual dictionaries. Lexikos16(1), 84-94.
Huang, G. F. 1985. The productive use of EFL dictionaries. RELC journal16(2), 54-71.


Ng, C. W. 2016. Impacts of the monolingual and bilingual dictionaries on the lexical errors committed by EFL learners in Hong Kong: a semantic analysis. *Lexicography* 2(2), 143-173.


Abstract

As a major resource and tool of language learning, learner’s dictionaries have provided sufficient information for L2 writing, which serves as a good guide of peer feedback. Hence, learner’s dictionaries are an indispensable part of scaffolding in the L2 writing feedback system. However, the effects of dictionary use in L2 writing have long been ignored either in L2 writing pedagogy or in learner lexicography. By applying the concept of “scaffolding” to peer feedback as the theoretical framework, this study first clarifies three distinct types of scaffolding information presented in current English learner’s dictionaries, and then makes an investigation into EFL learners’ perception and practical use of scaffolding information in their English writing. Results show that most EFL learners have positive attitudes towards scaffolding information and its role in motivating effective feedback in English writing. But their practical use of such information is not satisfactory owing to their inadequate skills and knowledge of dictionary use. This reflects a high demand of a dictionary use course in universities, which will help to raise EFL learners’ dictionary use efficiency as well as improve English teachers’ lexicographical expertise in English writing pedagogy.

Keywords: scaffolding information, learner’s dictionaries, peer feedback

1. Introduction

Providing effective feedback is quite essential in L2 writing pedagogy. So far there have been three general categories of feedback: teacher feedback, peer feedback, and online feedback (Zhou 2013)[1]. Among them, peer feedback is quite controversial in assisting L2 writing. Although some scholars have found that peer feedback has a

---

[1] Apart from “peer feedback”, there are some other popular terms used in research on L2 writing, for example, “peer response”, “peer comments”, “peer review”, “peer evaluation”, “peer revision”, “peer enquiries”, etc. Owing to its higher use frequency, this study uses “peer feedback” as a general term to achieve consistency in terminology. All the others are seen as its synonymous terms.
beneficial effect on writing improvement, encourages critical reasoning, and even leads to more learner autonomy (Villamil and De Guerrero 1998; Berg 1999; Rollinson 2005; Rouhi and Azizian 2013; Yu and Lee 2014), it turns out to have limited use in L2 writing due to a number of factors (Leki 1990; Carson and Nelson 1996; Connor and Asenavage 1994; Zhang 1995; Sengupta 1998). According to Rouhi and Azizian (2013: 1350), these factors fall into four categories. They include limited L2 proficiency and inadequate mastery of rhetorical rules, too much attendance to surface errors, negative attitudes towards peer feedback, and different sociolinguistic norms of interaction and beliefs in writing evaluation (Leki 1990; Nelson and Murphy 1992; Villamil and De Guerrero 1994; Tsui and Ng 2000). Therefore, to fully demonstrate the efficacy of peer feedback, sufficient guidance from teachers and necessary outside resources are needed to motivate L2 learners (Chang 2015). As an indispensable language learning tool, learner’s dictionaries have provided abundant information for L2 writing, which serves as part of scaffolding in the L2 writing feedback system (Yu 2013). However, owing to the long-time marginal role of dictionary use in pedagogical practice, the effects of learner’s dictionary use as the medium of feedback have long been ignored either in L2 writing pedagogy or in learner lexicography.

This study, by applying the concept of “scaffolding” to peer feedback as the theoretical framework, intends to investigate into scaffolding information presented in English learner’s dictionaries from users’ perspective with a focus on their perception and practical use in English writing process. On the basis of the results, some suggestions are given on lexicographical practice so as to enhance the role of dictionary use in helping to facilitate peer feedback efficacy in English writing pedagogy. Hopefully the study will also provide evidence in support of designing a dictionary use course for EFL learners.

2. Literature Review

2.1 Scaffolding theories

Scaffolding theories used in language pedagogy originated from the construction field, in which scaffolding serves as a crucial building aid. It was introduced into the teaching arena in 1976 with an emphasis on the belief that by positive guidance or collaboration, poor learners will receive assistance from teachers or good learners. This will, in turn, benefit their language learning to a large extent. By incorporating Vygotsky’s theory of the Zone of Proximal Development (1978: 86), scaffolding instruction is aimed at building a conceptual framework for learners so as to enhance their perception of certain language points. With such a scaffolding framework, complicated learning tasks could be disintegrated, and learning efficiency is very likely to be escalated accordingly. (Wei 2016).

As to the effects of the application of scaffolding theories in language pedagogy, researchers, both home and abroad, have made many investigations in different contexts. For example, Cazden (1979) found that scaffolding instruction helped
learners to achieve better cognitive development when applied to specific teaching activities. Richards (1996) testified that new linguistic structures were better acquired through incorporating scaffolding into language teaching. Zhang (2004) elaborated on the significance of scaffolding theory in high school English writing instruction. Bai (2007) explored the influence of scaffolding theory on English learners’ listening and speaking in universities.

When it comes to the connection between scaffolding theories and writing pedagogy, there are two research findings which are worth mentioning. One is the L2 writing feedback model (Yu 2013), which involves Subject, Object, Medium, Rules, Learning Community, Separation of Work, and Result concerning feedback. Among them, Media serve as necessary scaffolding because of their direct interaction with all the other six components. With dictionaries as a significant medium, their practical use will affect the effectiveness of feedback, especially peer feedback, to a large extent. The other is the detailed classification of three types of scaffolding in helping to facilitate effective peer feedback in L2 writing pedagogy by using English learner’s dictionaries as an example (Wei 2016). It has been justified by analyzing general dictionaries and specialized dictionaries that the scaffolding effects of learner’s dictionaries are found in three aspects, i.e. language, structure, and content respectively. However, these two research findings are both based on theoretical analysis without statistical evidence. Therefore, this study intends to investigate the availability and feasibility of scaffolding information in English learner’s dictionaries. To achieve the goal, the scaffolding information presented in dictionary texts needs to be analyzed in depth first.

2.2 Scaffolding information in learner’s dictionaries

Based on large-scale empirical studies, EFL learners usually have difficulty with the English language, content, and structure in composing an essay (Wang and Yu 2008). In other words, inappropriate vocabulary use, lack of content, and poor organization are what EFL learners fail to excel in English writing. On the basis of Wei’s elaboration of scaffolding effects of English learner’s dictionaries, scaffolding information in English learner’s dictionaries can be classified into three aspects as shown in Figure 1.

---

2 Different from the study conducted by Wei (2016), the focus of the current research is on general-purpose dictionaries. Therefore, the dictionary texts chosen for the following analysis are the Big Five as well as MWALD which are common choices for EFL learners. Accordingly, learner’s dictionaries or English learner’s dictionaries in this study mainly refer to English-English dictionaries for general purposes.
First, Writing Basics (WB) refers to fundamental grammatical knowledge required in English writing for general purposes. It is the most essential information provided in learner’s dictionaries, including “word classes”, “tenses”, “sentence structures”, “collocations”, “affixes”, “common grammatical mistakes and errors”, etc. Such information serves as good scaffolding for L2 learners to observe basic grammatical rules concerning language structure, which are thus too simple to be ignored by EFL learners. Both English-English dictionaries and English-Chinese dictionaries which are now commonly found in Chinese market have rich resources in this aspect.

Second, Writing Expertise (WE) involves the professional knowledge to assist English writing. It touches upon “registers” (e.g. word choice for different themes), “styles” (e.g. formal and informal), “genres” (e.g. writing for general purposes, writing for academic purposes, and writing for practical use), and “writing strategies” (e.g. cohesion, logic, and transition). Unlike WB, WE focuses on content, which is aimed at introducing some central concepts which L2 learners need to master before, during and after the writing process. With such information, they will understand the nature of English writing holistically, learn about how to cope with content properly, and become professional writers. Unfortunately, WE is now only available in English-English dictionaries such as Oxford Advanced Learner’s Dictionary of English (OALD, 2010), Longman Dictionary of Contemporary English (LDOCE, 2009), Cambridge Advanced Learner’s Dictionary of English (CALD, 2013), Collins English Dictionary for Advanced Learners (COBUILD, 2012), Macmillan English Dictionary for Advanced Learners (MEDAL, 2005), Meriam-Webster Advanced Learner’s Dictionary of English (MWALD, 2008). To achieve different goals in dictionary compilation, each one has its own features in textual design. Therefore, EFL learners need to know about their features before making a choice to help with their offering of peer feedback.

Compared with the previous two, Writing Instruction (WI) offers workable guidance and suggestions in coping with potential problems that L2 learners might encounter in their writing. It covers the design of the overall textual structure, steps in general writing procedure, guidance on varied writing genres, and draft revision. It is mainly used to familiarize L2 learners with the writing process, and improve their writing efficiency eventually. As WI is quite a new design feature in dictionary texts, even among the Big Five monolingual English dictionaries, only two, i.e. OALD and
CALD, have made this innovative attempt, let alone English-Chinese dictionaries.

3. Methodology

To investigate users’ perception and use of scaffolding information available in English learner’s dictionaries, the study made a survey among EFL learners by using an online questionnaire. Details are given as follows.

3.1 Participants

Participants in the investigation are undergraduate students from more than 20 universities in China. Among them, 32.6% are male, and 67.4% are female. 20.3% are in the first year, 51.7% in the second year, 18.3% in the third year, and 9.7% in the fourth year. Furthermore, 26.2% are English majors, while 73.8% are non-English majors. They fall into three groups with regard to their English proficiency: advanced learners (AL, 31.4%), intermediate learners (IL, 44%), and poor learners (PL, 24.6%). Generally the distribution of participants concerning gender, age, major, and English proficiency is reasonable.

3.2 Procedure

A questionnaire was designed by using a professional online model, and then distributed on the Internet. More than 700 questionnaire were collected, in which 629 turned out to be effective. After some fundamental calculation done online, further detailed data analysis was made so as to help in-depth discussion that followed. For the sake of clearer demonstration of the results, percentages were used to explain the data collected for each question item under investigation.

3.3 Questionnaire

The self-designed questionnaire used for the survey is composed of 13 items, which are grouped into four sections. They are learners’ perception of scaffolding information (Q1 to Q5), learners’ practical use of scaffolding information (Q6 to Q9), learners’ needs of scaffolding information (Q10 to Q12), and learners’ viewpoints of scaffolding information presentation (Q13). All of them are multiple-choice questions except Q13 which invites participants to write down their opinions voluntarily.

---

3 The criteria of participants’ English proficiency are their performance in TEM4 (Test of English Majors, Band 4) and CET4 (College English Test, Band 4) respectively. English majors with TEM4 above 71 are considered AL, those between 61-70 are IL, and those below 60 are PL. Non-English majors with CET4 above 498 belong to AL, those between 425-497 are IL, and those below 424 are PL.

4 The website is www.sojump.com for reference.
4. Results & Discussion

4.1 Learners’ perception of scaffolding information

Q1-Q5 are designed to investigate learners’ perception of the effects of scaffolding information presented in learner’s dictionaries as an aid for English writing. Results show that 77.78% believe in the positive role that learner’s dictionaries play in assisting writing, and 70.27% consider it necessary for dictionaries to include a special writing section like “Writing Tutor” in Oxford Advanced Learner’s Dictionary to improve their writing as shown in Table 1. This proves that incorporating dictionary use into English writing is well accepted by EFL learners (Harvey and Yuill 1997). To be specific, WB is the most popular scaffolding information with the highest percentage (93.58%), followed by WE (85.92%) and WI (84.82%) respectively. It reflects EFL learners’ heavy reliance on grammar or language accuracy in English writing, and their relatively less attention to content and structure in composing an English essay.

As for different learners’ perceptions to scaffolding information, it has been found that generally there are more SI supporters with the rise of English proficiency as shown in Table 1. In other words, the better English they have, the more important they think scaffolding information is in assisting English writing, and the higher expectancy they have in the innovative design of such information design in learner’s dictionaries. This finding is consistent with the previous study (e.g. Liu 2010). But the highest percentage of Q1, Q2, Q4, and Q5 all rests with IL, not AL. The possible reason for this phenomenon might be the fact that with strong self-learning skills, AL doesn’t have to rely so much on dictionaries for good writing as PL and IL do. As Liu (2011: 121) proposed, EFL learners with different English proficiency require different dictionary use guidance to improve their English writing. It also demonstrates that the target users of scaffolding information in learner’s dictionaries must be those whose English is at or below intermediate level. This finding is quite meaningful from a lexicographical perspective, especially for the innovation of English-Chinese dictionary compilation.

Table 1 Contrast of learners’ perceptions of scaffolding information

<table>
<thead>
<tr>
<th>Group</th>
<th>Q1: Usefulness of SI</th>
<th>Q2: Special writing section</th>
<th>Q3: WB</th>
<th>Q4: WE</th>
<th>Q5: WI</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL</td>
<td>75.16%</td>
<td>69.43%</td>
<td>91.72%</td>
<td>78.34%</td>
<td>76.43%</td>
</tr>
<tr>
<td>IL</td>
<td>77.94%</td>
<td>70.82%</td>
<td>93.59%</td>
<td>89.32%</td>
<td>88.26%</td>
</tr>
<tr>
<td>AL</td>
<td>77.47%</td>
<td>68.78%</td>
<td>94.45%</td>
<td>83.74%</td>
<td>83.39%</td>
</tr>
<tr>
<td>Total</td>
<td>77.78%</td>
<td>70.27%</td>
<td>93.58%</td>
<td>85.92%</td>
<td>84.82%</td>
</tr>
</tbody>
</table>

*SI: scaffolding information, WB: writing basics, WE: writing expertise, WI: writing instruction
4.2 Learners’ practical use of scaffolding information

Q6-Q9 are designed to investigate learners’ practical use of scaffolding information, i.e. special writing sections available in the four English-English dictionaries—OALD, COBUILD, CALD, and MEDAL. Generally speaking, SI’s overall use is rather unsatisfactory as shown in Table 2.

Table 2 Learner’s use of special writing sections in learner’s dictionaries

<table>
<thead>
<tr>
<th>Group</th>
<th>Q6: OALD</th>
<th>Q7: COBUILD</th>
<th>Q8: CALD</th>
<th>Q9: MEDAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL</td>
<td>15.92%</td>
<td>7.64%</td>
<td>5.1%</td>
<td>3.82%</td>
</tr>
<tr>
<td>IL</td>
<td>20.28%</td>
<td>10.32%</td>
<td>8.19%</td>
<td>4.98%</td>
</tr>
<tr>
<td>AL</td>
<td>24.14%</td>
<td>13.03%</td>
<td>9.21%</td>
<td>7.11%</td>
</tr>
<tr>
<td>Total</td>
<td>18.6%</td>
<td>9.86%</td>
<td>7%</td>
<td>4.45%</td>
</tr>
</tbody>
</table>

First, EFL learners are not so familiar with the specially designed writing sections which provide abundant SI. Among the four learner’s dictionaries, participants’ familiarity of “Oxford Writing Tutor” in OALD enjoys the highest percentage (18.6%), followed by “Brief Writer’s Handbook” in COBUILD (9.86%), “Focus on Writing” in CALD (7%), and “Improving Your Writing Skills” in MEDAL (4.45%). Consequently, it is very likely that SI conveyed in learner’s dictionaries cannot be widely used. This result can be attributed to the arrival time of these dictionaries at the Chinese market, their sales, and their reputations among EFL learners. In this case, English teachers are responsible to introduce the new textual design of these dictionaries to EFL learners so as to make better use of SI in writing pedagogy (Wu 2007; Yang and Wei 2016).

With regard to different learners’ actual use of the special writing sections, it has been found that AL reported the highest use frequency among all the four English dictionaries under investigation than IL and PL as shown in Table 2. This indicated the positive correlation between English proficiency and SI’s practical use as found in. Of course, more statistical evidence is needed to confirm this hunch in the follow-up large-scale empirical study.

4.3 Learners’ needs of scaffolding information

Q10-Q12 are designed to investigate learners’ needs of scaffolding information in helping with their English writing, each of which has a different focus. Participants are required to report their needs in their English writing among the 23 items of WB in Q10, 9 items of WE in Q11, and 9 items of WI in Q12.

Table 3 Learner’s needs of scaffolding information

<table>
<thead>
<tr>
<th>Rank</th>
<th>Q10: WB</th>
<th>Q11: WE</th>
<th>Q12: WI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English spelling</td>
<td>Diction</td>
<td>Writing for practical</td>
</tr>
<tr>
<td></td>
<td>Percentage</td>
<td>Description</td>
<td>Use</td>
</tr>
<tr>
<td>---</td>
<td>------------</td>
<td>--------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>2</td>
<td>77.74%</td>
<td>Tenses</td>
<td>Writing for general purposes 80.44%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Textual structure 81.24%</td>
</tr>
<tr>
<td>3</td>
<td>67.73%</td>
<td>Phrasal verbs</td>
<td>Thesis writing 79.81%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Writing strategies 75.99%</td>
</tr>
<tr>
<td>4</td>
<td>66.14%</td>
<td>Word classes</td>
<td>Writing for practical use 75.36%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Writing for general purposes 78.22%</td>
</tr>
<tr>
<td>5</td>
<td>65.5%</td>
<td>Lexical collocations</td>
<td>Thesis writing 75.03%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Writing procedure 75.04%</td>
</tr>
</tbody>
</table>

As shown in Table 3, the first five highly demanded items of WB are “English spelling” (77.74%), “tenses” (75.04%), “phrasal verbs” (67.73%), “word classes” (66.14%), and “lexical collocations” (65.5%). The first five items of WE with the highest percentage are “diction” (87.92%), “writing for general purposes” (80.44%), “writing strategies” (75.99%), “writing for practical use” (75.36%), and “thesis writing” (75.03%). The first five items of WI which are needed most are “instruction on writing for practical use” (82.03%), “instruction on textual structure” (81.24%), “instruction on thesis writing” (79.81%), “instruction on writing for general purposes” (78.22%), and “instruction on writing procedure” (75.04%).

Comparatively, WE and WI receive higher percentage than WB. As the current SI presentation of learner’s dictionaries shows, these two types of SI are not so common as does WB, more innovations in WE and WI are required to meet learners’ needs. In this case, the pedagogical effects of learner’s dictionaries will be brought into full play in English writing.

### 4.4 Learners’ viewpoints of scaffolding information presentation

Q13 invites EFL learners’ viewpoints concerning scaffolding information presentation in learner’s dictionaries as well as dictionary use in English writing on a voluntary basis. Altogether 66 participants’ viewpoints have been collected. Among them, 35 participants (53.03%) show positive attitudes towards the effects of dictionary use in English writing. They believe that SI presentation in English-English dictionaries is “highly practical”, “extremely important”, and “very good”, hoping that this innovation could be applied to English-Chinese dictionaries as well.

From their perspective, SI is useful in the following aspects: “helping to raise accuracy in expression”, “enriching vocabulary”, “guiding learners through different stages in the whole writing process systematically and scientifically”, “helping to tackle problems we encounter in writing”, and “improving our writing skills to a large extent”. Obviously, the feedback covers all the three types of SI, i.e. WB, WE, and WI. They also recommend adding some new items of SI such as “well-written English
sentences”, “notes to the writing section of TEM/CET”, “frequently used phrases”, “writing samples”, and so on. These opinions reflect EFL learners’ high demands of SI, and some of them really deserve attention from lexicographers and English teachers.

25 participants (37.88%) express doubts about the effects of SI in English writing. They think that writing depends more on self-practice and teachers’ guidance than dictionary use, and that more dictionary use will result in learners’ dependency, thus interfering with writing improvement. 2 participants (3.03%) state that “the rules and formulas provided in dictionaries will deprive EFL learners of their creativity in writing”. 1 participant (1.52%) expresses clear opposition to the relation between dictionary use and English writing because “dictionaries are designed for consulting words”, and “if detailed scaffolding information is provided in dictionaries, what are writing manuals used for?” Another 3 participants (4.5%) are unclear about SI’s role in English writing. It is clear that some viewpoints reflect learners’ lack of understanding concerning learner’s dictionaries and their effectiveness in English writing. Therefore, a dictionary course design is quite necessary to familiarize EFL learners with effective dictionary use in English writing (Wan and Ling 2005; Wu 2007).

5. Conclusion

Scaffolding information in current learner’s dictionaries falls into three types: WB, WE, and WI. Their role in motivating effective feedback in English writing has gained wide acceptance among EFL learners.

Results of the survey show that the higher English proficiency EFL learners have, the more important they consider SI is in assisting English writing, and the higher expectancy they tend to possess in SI presentation in learner’s dictionaries. However, the overall use frequency of SI is rather low due to EFL learners’ lack of dictionary skills and knowledge. Comparatively, the demands of WE and WI are higher than that of WB, which calls for more lexicographical innovations in these two aspects in both English-English dictionaries and English-Chinese dictionaries. As for EFL learners’ viewpoints of scaffolding information presentation, most are positive, but still quite a number of learners are in need of professional knowledge of learner’s dictionaries. This requires the systematic and scientific design of a dictionary use course in universities (Wan and Ling 2005; Wu 2007; Chen 2007, 2008; Wei 2016). That will not only help EFL learners to raise their dictionary use efficiency, but also improve English teachers’ lexicographical expertise in English writing pedagogy.
Acknowledgments

This study is financially supported by the Education Department of Shaanxi Provincial Government on the topic “Introducing Learner’s Dictionary Use to the Development of L2 Writing Autonomy” (NO. 16JK1746).

References


Dictionary Use and Peer Feedback in EFL Writing

Ying Cai
School of Foreign Languages, Wuhan University of Technology
cathycaiy@163.com

Abstract

Recent years have witnessed the fast and steady growth in the research into dictionary use, especially the use of learner’s dictionary. 70 students at a university participated in an eight-week instructional experiment which was designed to explore whether dictionary use can be adopted as an effective strategy to improve the performance of peer feedback in EFL (English as a foreign language) writing class. All the participants had a training course on peer feedback, and then group A (experiment class) had another training course on dictionary use. A post-task survey was also conducted to investigate students’ lookup behavior and in what ways dictionary use and guidance on it affect the quality of peer feedback. Multiple sources of data were collected, including essays, peer feedback forms, and questionnaires. Results of the study revealed dictionary use plays a significant role in improving the performance of peer feedback. The specific findings were also indicated: 1) dictionary users can give more appropriate and accurate suggestions in peer feedback than those who do not use a dictionary; 2) students receiving instruction on dictionary use tend to consult dictionary more frequently than those who do not; 3) instruction on dictionary use can direct students’ lookup behavior, making them aware of various methods they can adopt to obtain information from dictionaries, enhancing their ability to give feedback; 4) students are more likely to consult dictionary for information about spelling, part of speech, collocation, usage, and example in peer feedback based on their English language proficiency; 5) dictionary use and guidance on it motivate students to become more independent FFL learners, thus improving their autonomous learning ability. However, the instruction on dictionary use needs to be optimized in terms of content, method and time length. It is suggested that EFL teachers and professionals incorporate lexicographical achievements into EFL writing pedagogical programs, which in turn may promote the development of dictionary making.

Keywords: dictionary use, learner’s dictionary, peer feedback, EFL writing

1. Introduction

An increasing number of studies have been carried out to explore dictionary use from
different perspectives, including users’ reference needs, reference skills, guidance on
dictionary use, role of dictionary use in language learning and teaching. Scholars
(Hartmann 1979; Tono 2001; Wu 2001; Yong 2003; Luo and Zhao 2005; He 2008;
Wei 2009; Xu 2009; Chen 2012; Zhang 2015) have focused on dictionary users’
perspective and contributed pertinent suggestions on how to guide EFL learners to get
the utmost out of dictionary, especially learner’s dictionary. Learner’s dictionary is
defined as “a pedagogical dictionary aimed primarily at non-native learners of a
language. The degree to which dictionaries have been integrated into the learning
process varies from culture to culture…” (Hartmann and James 1998) Although they are
designed to fulfill dictionary users’ needs and pedagogical assumptions, these exactly
learner-targeted dictionaries haven’t received attention from teachers and advanced EFL
learners as much as lexicographers expected. The attempt to apply the findings of
dictionary use studies to EFL teaching and learning has been a compelling challenge.
The present research intends to demonstrate whether dictionary use can be adopted as
an effective strategy to improve EFL learners’ performance of peer feedback, thus
improving their writing. If it is true, in what specific ways does dictionary motivate
feedback givers to offer more constructive and useful suggestions? It tries to figure out a
way to integrate lexicographical research results into EFL teaching and learning.

2. Literature review

Feedback, either from the teacher or from other learners, serves as a common practice to
improve EFL writing. According to Longman Dictionary of Language Teaching and
Applied Linguistics, peer feedback is an activity in the revising stage of writing in which
students receive feedback about their writing from other students – their peers. Typically
students work in pairs or small groups, read each other’s compositions and ask
questions or give comments or suggestions. Foreign and Chinese scholars conducted an
enormous number of systematic studies on peer review. Some scholars (Caulk 1994; Mo
2007; Topping 2009; Cai 2011) emphasized the significant benefits of peer feedback.
They argued that peer feedback is more practical and constructive than teacher feedback,
and it can sharpen students’ awareness of writing and help build an English learning
community. On the contrary, for other scholars (Nelson and Murphy 1992; Sengupta
1998), peer feedback seems to have little or no impact, especially when feedback-givers
are learners with low English proficiency or giving critical and negative comments on
peers’ essays. However, most of the scholars (Zamel 1985; Keh 1990; Yang 2006; Yu
2013; Zhou 2013) held that peer feedback can be used as a useful supplement to teacher
feedback in EFL writing classroom since teacher feedback is considered more
authoritative while peer feedback is more effective in reducing students’ anxiety, and
significantly improving their learning autonomy.

More significantly, several previous studies have addressed strategy use in peer
feedback. Peer feedback training could be used as an effective strategy to improve the
quality of comments and suggestions. (Min 2006). Tono, Satake, and Miura (2014)
discussed using corpora on revision tasks in L2 writing. They found that information from corpus data actually helped in revising different types of errors in the essays. Yu and Lee (2016) concluded five major strategies language learners adopted in peer feedback: using L1, employing L2 writing criteria, adopting rules of group activity, seeking help from teachers, and playing different roles. With a view to increasing the effectiveness of peer feedback, Wei (2016) proposed using learner’s dictionary as the scaffolding of peer feedback in language teaching from lexicographical perspective.

3. Research Methodology

3.1 Research questions

The experiment is designed to explore whether dictionary use can improve learners’ performance of peer feedback in EFL writing. The specific research questions are as follows: 1) Do students using electronic advanced learner’s English-Chinese dictionary give more useful and constructive suggestions than those who did not use a dictionary? 2) Can students having received instruction on dictionary use perform better in peer feedback than those who have not? 3) Which parts of learner’s dictionary work best for peer feedback in EFL writing?

3.2 Participants

The experiment is conducted in two EFL regular classes from a university in south central China. Class 1 (including 35 students) is called group A (experiment group) having training courses on peer feedback and dictionary use while class 2 (consisting of 35 students) is named group B (control group) only receiving instruction on peer feedback. 70 participants are first year science and engineering postgraduates. They are taught by the same English teacher who has taught postgraduates for more than ten years and they have two regular classes per week (90 minutes per week). In addition, they have a similar language background since they have been learning English for 8-12 years, and 70% of them have passed College English Test Band 6.

3.3 Peer feedback instruction

All subjects in group A and group B were asked to have a 45-minute training class in which the process of peer feedback and evaluating system were introduced. A peer feedback form, which was made based on IELTS scoring system, was handed out to them. The researcher explained each item of the form to the students with detailed examples, guided students to read a sample, and showed them how to make a judgment, mark errors in the essay and fill out the form. All subjects were asked to fill out form in red ink for information obtained from dictionary, but in pencil for information from other resources.
3.4 Dictionary use guidance

The dictionary use guidance only targeted at group A (experiment group) in two 45-minute regular classes. The researcher introduced some influential and authoritative learner’s dictionaries and explained how dictionaries are designed to help users get what they need. These participants were expected to get familiar with dictionary design and reference skills through a series of carefully designed activities. To make dictionary use possible and convenient in class, the researcher asked the participants to download the pre-built dictionary data, such as the 7th edition of *Oxford Advanced Learner’s English-Chinese Dictionary*, the 4th edition of *Longman English-Chinese Dictionary of Contemporary English*, the 3rd edition of *Cambridge Advanced Learner’s English-Chinese Dictionary*, and *Collins Cobuild Advanced Learner’s English-Chinese Dictionary* and install them in Eudic, an open dictionary platform, which has been installed on their smart phones.

3.5 Procedure

The participants completed the whole experimental cycle in approximately eight weeks. 1) They were demanded to write four essays, involving elite education, cosmetic surgery, environment and tourism. Data was collected when the students finished their fourth essay task, which required them to write an argumentative essay with at least 200 words on the topic “Tourism brings about some problems for the local area. What are the problems and what can we do to solve the problems? ” They were supposed to give reasons and use relevant examples from their own knowledge or experience within 30 minutes without any reference book in class. The last four numbers of their student numbers, instead of their names, were written on the answer sheet. The data extracted from the last task was applied to the research since the participants had finished the first three peer feedback activities and become familiar with essay revising rules and feedback giving. In addition, students in group A (experiment class) had enough time to practice dictionary reference skills through essay correction. 2) The researcher invited group A to correct group B’s essays, and asked group B to correct group A’s essays. The whole anonymous revising process lasted for 35 minutes. Students were encouraged to use their smartphones, or discuss problems with their partner who sat next to them. They marked different error types in the essay and filled out the feedback forms in red ink (information from dictionary) or in pencil (information from other resources). If they could correct the errors, they corrected them. Finally, the written feedback forms were gathered.

3.6 Survey

An after-peer feedback survey was conducted. 70 participants were asked to fill out a questionnaire with 10 questions which emphasizes the following aspects: 1) What did you choose to be a revising strategy when you encountered language difficulties in peer
feedback? What were your reasons? 2) If you have used a dictionary, what was it? 3) Which parts in the dictionary were the most useful for essay correction? Why? After 10 minutes, questionnaires were collected. The researcher stapled the essay, feedback form and questionnaire together for each participant.

4. Results and discussion

4.1 Revising strategy choice

![Figure 1 revising strategies](image)

Figure 1 shows different strategies group A and group B chose to correct their peers’ essays. Although it is common that both of the groups set dictionary use as their priority, the students of group A having chosen to use dictionary are obviously more than those of group B. Interestingly, compared with group A, nearly half of group B chose pair discussion as a strategy to improve the quality of their feedback. It indicates that the students of group A who are skillful in dictionary use have become more independent when it comes to giving comments on peers’ essays. Students of group B explained in the questionnaire they chose discussing with partners for it is more efficient and convenient than consulting a dictionary even though they knew answers from partners might not be 100 percent accurate. Another reason is that they couldn’t understand some sentences of the essay they got, so they wanted to discuss with their partners. It seems that using the Internet was a less frequently-used strategy for them since they failed to realize the advantages of the Internet in essay correction.
4.2 Dictionary choice

As presented in figure 2, the overwhelming majority (77.14%) of group B turned to electronic dictionary Youdao when they needed to look up a word. Although Collins Cobuild Advanced Learner’s English-Chinese Dictionary and The 21st Century English-Chinese Dictionary are included in Youdao, most of the students in group B just ignored them and chose Youdao English-Chinese dictionary. Without guidance on dictionary use, they cannot realize the differences between advanced learner’s dictionary and Youdao English-Chinese dictionary.

Figure 3 displays that group A has been influenced by the instructor and the guidance on dictionary use. 85.71% of the students chose Eudic in which the most popular and influential advanced learner’s English-Chinese dictionaries have been installed.

4.3 How dictionary works for peer feedback

<table>
<thead>
<tr>
<th>Table 1 identified errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary</td>
</tr>
<tr>
<td>Spell</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>GroupA</td>
</tr>
<tr>
<td>Group B</td>
</tr>
</tbody>
</table>

Table 1 reveals the errors identified by students making peer feedback in writing class. Given the other two indicators task response, coherence and cohesion are less relevant to this study, the researcher has not taken them into consideration. From the information described in the table, it can be seen that group A identified more errors about spelling, word choice, part of speech, plural or singular and collocation than group B. However, group A and group B spotted similar number of errors about tense and sentence. They argued in the survey that dictionary use cannot exert much positive influence on identification and correction of the two types of errors, which may be directly related to feedback givers’ English language proficiency. In this case, it implies group A and group B are English learners at a similar level.
From figure 4 above, it can be seen that students of group A consulted dictionary mainly for information about spelling (94.29%), part of speech (85.71%), collocation (80.00%), usage (71.43%), and example (57.14%) in peer feedback. Similarly, students of group B also used dictionary to make sure whether words were spelt or used in the accurate way although the rate of their dictionary use was relatively lower than that of group A. The results from the survey tally with the peer feedback forms. Some typical errors spotted and corrected are as follows:

**spelling:** resource *depletion* (→depletion); *amenities* (→amenities); limitate *people* (→limit); the *over population* can create a great waste of local resources. (→overpopulation);

**part of speech:** water *scare* (→scarcity); reduce the *consume* of water (→consumption); development of *economic* (→economy); The *destroy* of the cosystem made them die out. (→destruction) ; increasing the pressure of *local person* (→the locals);

**collocation:** *release* of the sewage (→discharge); so *far* (→forth); To avoid worse development of local area, three solutions are admitted.(→proposed); have impact to (→on); There’s no simple solution for the problem.(→to);

**usage:** Spending holiday will also make us *stressful*. (→Spending holiday is also stressful.); We could make tips to *avoid tourists to make noise*. (→avoid sb /sth/doing sth);

**example:** There’s *not* doubt.(→no)

The figure also indicates that looking up word and figuring out its meaning was considered least important in feedback giving. The underlying reason is that “the words
their peers used in the essays were so common and simple that it was not necessary for them to look for word meanings in a dictionary.” written by more than 70% of students in the survey. Only when essay writers used common words in an improper and unusual way did feedback givers start to doubt themselves and then decide to use the dictionary. For example: It’s hard to resolve the problem in dry areas. (solve); tourism businessman → travel agencies; The government should take more responsibility and the manager can cut off or lower the financial support… (government officials) ; the waste of water should also be solved. (?);

A small percentage (8.57% from B; 14.29% from A) of students were so careful and rigorous that they found labels and inflections shown in dictionary useful to help them give suggestions and comments. For example: Well, in my opinion (spoken word. Pleas delete well); trekking → trekking).

Compared with group B, it appears that students of group A have learned to view the microstructure and macrostructure of learner’s dictionary from a new perspective and endeavored to use reference skills gotten from the guidance to improve the quality of peer feedback.

5. Conclusion

This study illustrates the role of dictionary use in facilitating peer feedback in EFL writing. Lexicographers and scholars investigate dictionary users’ needs and compile user-targeted dictionaries. However, it is not merely needs of different users that are of importance. An equally important factor is users’ ability to find and make use of the information given. When dictionary use is adopted as a strategy, guidance on dictionary use is crucial to the improvement of students’ revision and writing abilities. The analysis of 70 essays, peer feedback forms, and questionnaires demonstrated that: 1) dictionary users can give more appropriate and accurate suggestions in peer feedback than those who do not use a dictionary; 2) students receiving instruction on dictionary use tend to consult dictionary more frequently than those who do not; 3) instruction on dictionary use can direct students’ lookup behavior, making them aware of various methods they can adopt to obtain information from dictionaries, enhancing their ability to give feedback; 4) students are more likely to consult dictionary for information about spelling, part of speech, collocation, usage, and example in peer feedback based on their English language proficiency; 5) dictionary use and guidance on it motivate students to become more independent EFL learners, thus improving their autonomous learning ability.

On account of the limited size of the subjects, there is still room for improving the design of the whole study. For example, the researcher may invite more subjects or consider using computer-mediated peer feedback system. The instruction on dictionary use, an influential factor, needs to be optimized in terms of content, method and time length. The present study has some implications for integrating lexicographical achievements into EFL teaching and learning.
Supported by “the Fundamental Research Funds for the Central Universities (WUT: 2015VI044)”

References


A Comparative Research on User’s Expectation of Learner’s Dictionary

Shan He
Beijing Normal University
heshan2014@mail.bnu.edu.cn

Abstract

A good dictionary is one which can well serve user’s needs, which is necessary to survey, but not enough. Only when user’s expectation, which connects to the certain cultural, social and educational context, is fully revealed, can the needs be well understood, and that is crucial for the researcher and compiler trying to make better dictionaries to fulfill the needs. The purpose of this research is to argue that user’s expectation cannot be studied isolated because there are a lot of variables, even when some different group of users sharing the same expectation could result in different reasons. The research and compiling of Chinese learner’s dictionary are largely following the lead of English learner’s dictionary which sometimes is just duplicating, in other words, decontextualized. In this case, we surveyed 245 Chinese learners, and 115 Chinese-English learners to compare their needs on the dictionary and to discuss the correlation between their cultural background and their expectation.

Keywords: dictionary user, Chinese learner, Chinese-English learner, expectation, survey

1. Backgrounds

Compared with English learner's Dictionaries, which have over 70 year's history, learner's dictionary in Chinese lexicography is quite young. The immaturity of Chinese learner's dictionary is not adequate to the rapid growth of TCSL—till the end of 2015, over 500 Confucius Institutes had been established in 135 countries, and the number of Confucius Classrooms was more than 1000. However, the user group of Chinese learner’s dictionaries published by Chinese press is surprisingly small. (Cheng, 2005; Du, 2010; Zhang, 2011) The situation might due to the following three reasons: first, the quality of dictionary is not satisfying; second, the research and compiling is divorced from the marketing; the last but not the least, the focus of research and compiling is still on paper dictionaries rather than new medium ones which lags much behind the times. To sum up, user’s needs has yet been well fulfilled. The trickiest part
is that in the research and compiling of dictionaries, user’s expectation can be distorted because what researchers and compilers suppose learners should need may be not what they actually need.

The inferring was largely based on two parts: the research of Chinese lexicography and the research of English dictionary use: firstly, Chinese learner's dictionaries were compiled and researched largely under the paradigm of Chinese traditional lexicography which does not deal with second language learners. The side effect is that there have been a lot of studies on Chinese learner’s dictionary use which provide interpretations and suggestions in the perspective of teacher and linguistics but for dictionaries to better assist Chinese learning, it asks for a more sophisticated understanding of user’s expectation, which can only be revealed in learner’s perspective. The emphasis on learners means not only a transfer of research perspective but also an in-depth understanding of dictionary users whose behavior and cognition was shaped in the certain environment. Secondly, the research of English dictionary use which targets specifically at Chinese-English learners. It is necessary to learn from research on English dictionary, however, it is very dangerous to simply duplicate methodologically without taking the specific context into consideration; since each language is special, as well as the dictionary and its users, for example, the situation of English teaching and learning in mainland China is quite unique which poles apart from what Chinese learners are experiencing, though the Chinese teacher’s attitude to Chinese learner’s dictionary is affected by their second language (mostly English) learning and dictionary use experience. As the pioneer of learner’s dictionaries, it is widely believed that the features English learner’s dictionaries have are no doubt the necessities for other dictionaries, and to compile better Chinese learner’s dictionary means to compile more English learner’s dictionary-like dictionary which is decontextualized and not practical. As Spolsky (2000:131) argued that because language is primarily a social mechanism, languages are learnt in social contexts, learner’s dictionary should be characteristic to show the peculiarity of each language and serve its target users.

Dictionary use, as a part of language learning or at least a strategy used in language learning, should be described and interpreted in certain context as well. Quite a few studies were focusing only on describing numbers without contextual interpretation which could lead to misunderstanding. The purpose of this study is to argue that research on dictionary use is more than doing statistics but interpreting the result contextually. In this case, both Chinese learners and Chinese-English learners were surveyed to show the importance of cultural, social, and educational context, and get some clues of how might it influence different users’ expectation on the dictionary.
2. Users’ needs on the dictionary

Learner's dictionary is written for language learners who are in need, so the user has been and always will be the core of all the work, and that is the reason why one growing field of lexicography research is dictionary use, which is important but complicated as well. Despite the difference between languages, dictionary use is a dynamic process which varies in many ways, for instance, the environment of using, the user's habit of retrieving, the availability of paper/digital dictionary, classroom culture, the type of task, etc. Each of those aspects could be a variable of user's expectation, which decides how satisfied the users could be. In the research of Chinese learner's dictionary, most of which are surveys, the frequently asked questions include what dictionary do they use, what information do they look for, how long does it take to locate the target word. (Xia, 2009; Zhang, 2011; Cen & Shang; 2011; Hao & Wang, 2013; Jin, 2016) On the contrary, who are the users, how do they make choices, what role does dictionary play in their language learning, and the reason why were all the merely asked questions.

It is of great possible that user's cultural background and the cognition of dictionary are closely connected to what they expect from the dictionary as well as the way they make use of it. For instance, the ownership of E-dictionary is unexpectedly high among English learners in Japan which show the uniqueness of Japanese education culture. (Sterling, 2003) Schecter et al. (2001) argued that Asian educational culture emphasizes accuracy far more than encouraging risk-taking, so dictionary for Asian users is not only a tool for second language learning but also a source of emotional security in learning. It reminds us that more importance should be attached to the variations of dictionary users which is exactly the reason why we conducted this study. By comparing the use of dictionary use between Chinese-English learners and Chinese learners, the peculiarity and significance of context could be much clearer.

3. Questionnaires and participants

3.1 Questionnaires

Three surveys were released online individually. Owing to the online survey tools, the questionnaires can be very accessible and wild-spreadable——it can be filled on either tabletop, laptop, tablet, or smart phone; the survey link can be forwarded through E-mail or shared on social network. Furthermore, some other great services are provided for free, such as recording of the finish time, the automatically generated statistical charts are available, etc. which is a great time-saver, thanks to which this study is much more workable.
The questionnaire for the survey of Chinese learners (SCL) and survey of English learners (SEL) comprise 27 and 24 questions by which we are trying to gain participant’s basic information about language learning and dictionary use, as well as, his/her attitude to the dictionary. The questionnaire for Chinese teachers (SCT) consists of 20 questions which first and foremost ask about their attitude to dictionaries and the big picture of dictionary use of Chinese learners from teacher's perspective, which also can verify data collected from the other two surveys.

3.2 Participants

Three surveys were released online individually. SCL has 245 participants who come from over 20 countries and are mainly college students; SEL has 115 participants who are English learners studying in colleges in mainland China; SCT’s participants are 103 Chinese teachers who are working or were used to work either in China or overseas, the total number of collected questionnaires is 463.

3.2.1 Participants of SCL

Participants come from Asia and Europe-America are about half and half, but the former is unevenly distributed, as showed in Figure 1, about 40% of participants are from north America, on the other hand, participants come from anyone of South America, West Europe, and east Europe are more than 5%; while participants come from three Asian areas are more equally distributed which are all between 14% and 20%. 66% of the participants have been to China for language learning and about 89% of them have been taught by Chinese native speakers.

![Figure 1 The distribution of SCL participants](image)

The durations of their Chinese learning are showed in Figure 2: the biggest part, which is near 30% consists of participants who have learnt Chinese for over 5 years, those who have learnt for 1 to 2 years come second place which is about 20%, the rest participants are relatively evenly distributed in each stage.
3.2.2 Participants of SEL

In mainland China, English is compulsory course in secondary schools and colleges, so 96% participants of SEL have learnt English for more than 6 years (Figure 3), but the participants who had been taught by English native speakers, especially when they were beginning learners, are no more than 10%; those who had been to English-speaking countries for language learning is nearly none which differs exceedingly from Chinese learners.

Figure 2 The duration of SCL participant’s Chinese learning

3.2.3 Participants of SCT

Figure 4 shows the areas that participants of SCT are/have been to. About half (51%) of the participants have Chinese teaching experience in mainland China, and the other half(49%) have taught abroad. North America, Southeast Asia, and East Asia are the top 3 areas, but the participants who are in or had been to Africa and South America are just a small amount.
56% of the participants have less-than-3-year Chinese teaching experience, most of them are TCSL majored graduate students, more than half of whom are either teaching/taught abroad as volunteers or teaching/taught Chinese domestically as practice. 44% of participants have been teaching Chinese for more than 3 years can be regarded as experienced teachers. For teachers who teach Chinese in colleges, they must have master degrees at least, and it has been a trend in China that most positions in colleges are only open to those who have doctor degrees, that is the reason why master degree (73%) and doctor (23%) degree holders are more than 95% of the participants. (Figure 5)

**Figure 4** The areas that SCT participants are in/have been to

<table>
<thead>
<tr>
<th>Area</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainland China</td>
<td>80%</td>
</tr>
<tr>
<td>Southeast Asian countries</td>
<td>30%</td>
</tr>
<tr>
<td>European non-English-speaking countries</td>
<td>20%</td>
</tr>
<tr>
<td>European English-speaking countries</td>
<td>10%</td>
</tr>
<tr>
<td>Hong Kong, Macao and Taiwan</td>
<td>5%</td>
</tr>
<tr>
<td>African countries</td>
<td>5%</td>
</tr>
</tbody>
</table>

**Figure 5** SCT participant’s teaching experience

4. Analysis

Data from three surveys were analyzed basing on the basic statistical results provided by the online survey website (www.wenquan.com). The comparison indicates the results of SCL and SEL are different in several aspects.
4.1 Paper dictionary

4.1.1 The ownership of paper dictionary

The ownership of paper dictionary among English learners is at variance with Chinese learners. The portion of participants who own at least 1 (and up to 5) paper English dictionary is 88.7%, which is almost double the number of SCL participants who have ever used paper dictionary which is only 49%. It can be inferred that when it comes to the percentage of ownership the number will be even smaller.

As it was mentioned above, English is a compulsory course in secondary schools in China, which means most SEL participants started English learning when they were teenagers or even younger, and 71% of them purchased their dictionaries during this period as showed in Figure 6. In Figure 7, we can see that 68% of them chose the product of old and well-established brands including Oxford, Longman, and Cambridge, in addition, the dictionary published by Chinese local publishing house was welcomed by learners as well which has the same share of Cambridge, which comes the third most popular British English dictionary brand in the survey.

![Pie chart showing the percentage of English learners by education level](image)

**Figure 6** The beginning of using English learner’s dictionary
Figure 7 Chinese-English learners’ choices of the dictionary

It is quite clear that British English dictionaries are far more influential in mainland China than American English dictionaries which were resulted in the Chinese culture of believing in authority, and it becomes a mindset that the first/oldest one is the best since it represents for tradition. There are still people who believe that British English is more authentic than American English, and the most widely used secondary school English textbooks in China are co-published by Longman which might strengthen the opinion in a way.

89% participants of SEL are users of bilingual learner's dictionaries, on the contrary, the number is only 17% among Chinese dictionary users. Figure 8 shows the most widely used dictionaries which are all monolingual dictionaries, most of which are not even for second language learners. About 50% of participants accepted Chinese teacher's advice in choosing dictionaries, though some suggestions seem unprofessional. Users gave an average score of 7.33 on Chinese dictionaries. About 70% English learners valued their English teacher's opinion the most in the dictionary, and it seems those English teachers were quite trustworthy—at least the dictionaries they recommended are all learner's dictionaries, but not as expected, the average score gave by English learners is 6.5 which is lower than Chinese dictionary.
4.1.2 The cultural context

Why did Chinese-English learners give those reputational dictionaries a score which is barely above the standard? The reason is multiple which need to be comprehended in the certain context. To be specific, the concept of "dictionary" could be different in different cultures. In Chinese, "典" (dictionary) originally refers to the books which record the successful administrative experience of Five Lords, who are legendary rulers of China in ancient times and they are the role models could never be surpassed. Therefore, dictionaries are more than collections of words and phrases, but classics which need modestly prostrating readers themselves in worship. The traditional Chinese lexicography is about finding the exact explanations of classics to tell principles which were put forward by saints, rather than just explaining the meaning of a word or a phrase. This illustrates in a way why modern Chinese lexicography is focusing on semantics so much more than other fields. It also can be proved by the complaints of Chinese dictionary users that dictionaries have little usage information.

Figure 8 What Chinese dictionary do you use/have ever used?
Figure 9 shows the portion of participants of SCL who prefer paper dictionary to non-paper dictionary, and it is quite clear that the percentages among Asian participants are higher than that in European or American countries. Just like what was pointed out by Schecter et al. (2001), owning a dictionary brings Chinese students emotional security they are seeking for, and among all kinds of dictionaries, the thickness of traditional paper dictionary provides emotional security the best, especially before big exams when students are extremely anxious. When being asked in what situations they tend to use a dictionary, 38.9% of Chinese-English learners voted for “preparing for exams” which is twice of the number among English learners, which is only 19.6%. The stress put on accuracy drives language learner away from risk-taking and mistake-making, and this kind of mentality facilitates the belief that the dictionary is the finest language sample and the highest criteria which learners probably couldn't ever achieve. This idea restrains Chinese students from using a dictionary more actively and autonomously.

We asked participants of both SCL and SEL to score how much do they value different people's opinion in choosing a dictionary. Figure 10 shows language teacher's opinion was considered the most important by both Chinese-English learners and Chinese learners. It is quite clear that language teachers were scored way much higher than the others by Chinese-English learners, while, Chinese learner's scores are close to each other. In other words, Chinese-English learner's decision depends largely on their teacher's opinion whereas it's hard to say the suggestion from whom affects crucially on Chinese learner's decision which also means they made decisions after giving it lots of consideration.
Figure 10 How do participants value the advice from different people

The difference of maturity between English learner's dictionary and Chinese learner's dictionary could partially explain the difference—it is high possibility that English teachers offered clear advice on how to choose dictionary since the "Big Four" have already established their reputation worldwide and most English teachers in Chinese secondary schools are Chinese-English learners, who share the similar learning experience with their students which means they might have already thought this question through. On the contrary, there has been no widely acknowledged Chinese learner's dictionary, therefore Chinese learners have to take a wider range of opinions into consideration and their Chinese teachers most of whom are Chinese native speakers who have little experience of using Chinese learner's dictionary and probably cannot provide clear instruction on this and, moreover, it shows that many of the Chinese teacher’s lacking basic lexicographic knowledge which results in dictionaries for native speakers were recommended wrongly to Chinese learners.

There could be another reason that the roles of teachers are different which rooted deeply in cultures. In Chinese tradition, teachers are so respectful professionally and morally thus teachers are more like administrators, supervisors, and instructors rather than facilitators, organizers or enlighteners in western culture. The roles of the teacher in different society decides to what degree their advice affects their students.

4.2 Non-paper dictionary

Both Chinese learners and Chinese-English learners showed strong preference for non-paper dictionary including E-dictionary (Electronic dictionary which is a small device looks like calculator and usually loaded with several famous bilingual dictionaries), online dictionary (which is website based dictionary provided by mainstream search engines such as Google in America, Baidu in China and Naver in Korea) and mobile APPs (such as Pleco).
The inconvenience of using can be regarded as the most important reason which results in user’s preference for non-paper dictionary. The top five unsatisfying things of consulting paper dictionary are showed in Figure 11. Three of the five are about the inconvenience of using. Two groups of participants shared the same problem in consulting paper dictionary speedily and making use of information in the dictionary effectively, as showed in the second and last pair of columns. Locating the target word and understanding the interpretation were much difficult for Chinese learners probably because they mostly use the monolingual or non-learner’s-dictionary. On the other hand, more Chinese-English learners had difficulties in locating the specific sense even though most of them were consulting the bilingual dictionary. Besides that, some external elements should be considered to achieve a better understand of the situation.

4.2.1 The accessibility of non-paper dictionary

Another situation needs to pay attention to is that the accessibilities of the non-paper dictionary are different for two groups of participants. Chinese learners can access non-paper dictionary more easily, while most of the Chinese-English learners starts to use online dictionary/APPs when it was at least the 7th year of learning English, in other words, when they became college students. Their first 6-year English learning experience was gained in the secondary schools where personal cell phones, computers are forbidden, and E-dictionary is not recommended either. Therefore, their awareness of non-paper dictionary was developed much later than Chinese learners together with a strong desire.
4.2.2 The educational context

Compared with Chinese learners, Chinese-English learners have fewer opportunities to make full use of dictionary products, not only because of the accessibility mentioned above, but also the system of education. English weighs as much as Math and Chinese language and literature which are the three most important subjects in both senior high school entrance examination and university entrance examination. This makes most Chinese students learn English with strong external motivations which are much less powerful than internal ones and will reduce the enthusiasm of using a dictionary to assist English learning.

What’s more, Chinese-English learners were suggested buying dictionaries when they are teenagers who are less independent and the energy can be spent in English learning was very limited due to the great pressure of senior high school entrance examination and university entrance examination, in contrast, most participants started Chinese learning in colleges with their free will but without the stress of taking entrance examinations, so that they were motivated to learning Chinese more internally which led the use of Chinese dictionary product more actively.

5. What do users expect from a dictionary?

5.1 Convenient for using

Figure 12 shows Chinese teacher’s evaluation of the different type of dictionaries in which paper dictionary was scored much higher than non-paper dictionaries, despite that, Chinese learners prefer the latter much as showed in the second graph. The teacher’s evaluation is principally based on how authoritative and multi-functional the dictionary is, contrariwise, the data (Figure 9) shows both SEL and SCL participants prefer non-paper dictionary primarily because of the convenience which is one of the rapidly growing social needs in the new era of the internet. The freeness of online dictionary and mobile APPs motivates 53.6% of participants to use. The high-efficiency and low-cost are the primary consideration. Just like what Bejoint (2001:112) indicates that a welcome dictionary is not necessarily a good dictionary from the point of view of the lexicographer, who may have entirely different criteria; but it is certainly a dictionary that responds to a social need.
5.2 Help with tackling the difficulties in the learning

When being asked to vote for the features of a dictionary which are the most necessary and important, two groups of participants’ lists of top 20 contain the same features, but the orders are different which reflects the diverse of their needs. In Figure 13, Chinese-English learner’s list is the blue baseline, and the orange peaks above the blue line are the features that rank lower in Chinese-English learner’s lists, and those below are higher ranked. 6 features were ranked at least 37% higher by English learners and other 5 were 25% lower. The former includes the display of collocation, synonyms, cultural information, misuse warning, whether an expression is out of fashion, and illustrating by pictures which are less valued by Chinese learners. The convenience of retrieving, the number of vocabularies, antonyms, information of pronunciation, and help with memorizing words are involved in the latter which means they are considered more important by Chinese learners and are the weakness of Chinese learner’s dictionary. Despite the differences, participants of both SEL and SCL agreed with each other in the left 9 features, including example sentences, accuracy, readability, the difference between spoken language and written language, signs of high-frequency words, portability, and appendix.
The feature which ranked most distantly by dictionary users is the information of pronunciation. It was in the top 5 on Chinese learner’s list, while in the last 5 on Chinese-English learner’s list which largely results in the difficulty of tones in Chinese. Chinese is one of the few tonal languages in the world. The lacking of clues between Characters and the pronunciations together with the uniqueness of ideography is a great obstacle for Chinese learners, therefore it is reasonable that pronunciation information is vital for Chinese dictionary users.

6. The expectation: to be better learners? Or to better deal with the certain task?

6.1 Chinese learners

To the question “why do you learn English/Chinese?” most Chinese-English learners answered “I have to” because of the educational system left them no other options. On the contrary, Chinese learners gave much more characteristic answers, for instance, “wants to communicate with Chinese people” “wants to do business in China or with Chinese people” “wants to do better in their profession (e.g. Comparative Literature/International Relations)” “wants to challenge myself” and so on. Overall, Chinese learner’s purposes are more communicative and their needs on the dictionary are high on both decoding and encoding tasks.

Both Chinese learners (84.7%) and Chinese-English learners (75.7%) showeda strong preference for the non-paper dictionary, but it was showed that Chinese dictionary APPs are more welcome. The comparison between the most popular Chinese dictionary APP and two English dictionary APPs shows a great difference. Chinese dictionary APP is more multi-functional which helps users in memorizing
meanings and pronunciations, familiarizing the stroke’s order, and more importantly, encouraging users to personalize the APP to make it better fit their Chinese learning and help users to be more active in learning.

In the use of Chinese dictionary APP, the concept of Autonomous Learning/Self-Regulated-Learning is blended with the function of the dictionary which aims at making users better learners. It can be explained culturally that in western education, great importance is attached to the individuality, the development of individuality is the center. (Ding, 2011) Like what Dewey argued in Democracy and Education, educating is growing, a teacher is like a gardener who waters, weeds and provides the proper environment, the development of a person is like the growth of a plant, the key of which is to let the nature play. The meaning of external conditions is to stimulate, protect and guarantee the nature will eventually play nicely.

### 6.2 Chinese-English learner

Different from Chinese learners, Chinese-English learners show stronger needs on decoding tasks which conventionally include reading and listening which weigh totally about half in the university entrance examination, for comparing, speaking is not a part and writing weighs only about 20%, the proportion of which is similar in the curriculum in China. It is, as well, accordance with the data of IELTS, in which Chinese candidates performed better in reading and listening than the two encoding tasks, speaking and writing. Similarly, the Chinese-English learner’s favorite focus on two functions: translating and flashcards for preparing tests (e.g. CET-4/6, IELTS, TOEFL, GRE) both of which are more specific task targeted and keeps the basic, traditional functions of the dictionary.

This distinguishing feature also roots deeply in Chinese culture. Ding (2011) used “the-intellect-centered-type” and “the-morality-centered-type” to differentiate western education and Chinese education. Chinese culture values the hierarchy the most, therefore people believe in authority, obey their parent, respect the teacher, all because we mortals are imperfect who needs to be instructed, supervised and lectured to prevent from making mistake or letting the dark side of ourselves to dominant.(Gu, 2004; Ding, 2011)The uncreativity, dependence, inactiveness and conservativeness of Chinese students could be understood culturally as the respect of morality, so the dictionary product for Chinese users is ought to meet their pursuing of accuracy and authenticity. The language teacher can play a more critical role to show them the possibility that a dictionary can be used more effectively.
7. Conclusion

Dictionary user’s expectations are closely related to the motivation of language learning, the difficulties in their language learning and the problem they tend to solve by consulting a dictionary which roots deeply in the cultural, social, and educational context rather than how detailed or diverse the information is in a dictionary. Dictionary use is not an isolated behavior which should be correlated to other parts of language learning, which means the cultural perspective is essential in the research of dictionary use.

References

Lexicographical Delay

Erdogan BOZ
Eskisehir Osmangazi University, Turkey
dr.erdoganboz@gmail.com

Abstract

The concept of lexicographical delay or lexicographical lag, which is referred to as such several times in the Western literature, is an underdiscussed concept in Turkish. Thus, when we look at the terminology dictionaries of lexicography (Hartmann & James 1998; Burkhanov 1998), we cannot find these terms. In addition, Robinson (1983), Sterkenburg (2003) and Jackson (2013), who also provide us with a list of lexicological terms, also do not mention these terms. There are no such terms in the basic reference sources of lexicography (Zgusta 1971; Jackson 2002; Atkins and Rundell 2008; Svenson 2009, Durkin 2016). In addition, these terms were also not found in searches made in two respectable journals in the field of lexicography (International Journal of Lexicography1, Lexikos2). In this paper, the subject will be discussed in terms of user and dictionary relations after giving a short description of the term lexicographical delay.

Lexicographical delay can simply be described as follows: Delay experienced in the dictionary entry process of the words which fulfill the criteria to enter the general dictionaries and which are candidates to enter the general dictionaries; or the meanings of headwords available in the dictionaries, which are newly emerged but must be included in the dictionary. Of course, a more appropriate terminological description can be provided for this term. The word "meh" in English has been discussed in this context for some time in social media.

The effect of lexicographical delay is significant in terms of user and dictionary relation. Dictionaries, which act as a reference source for dictionary users, are elements where the user temporarily spends a certain time while performing another task. The user wants to get done with the dictionary as soon as possible and return to the original task at hand. For this reason, the user is in a hurry and wants to find out what he seeks immediately. If a word searched is not found in the dictionary and this repeats a couple times, the dictionary user first becomes frustrated and then a problem of trust arises. Accordingly, an indifference towards the dictionary may be created and gradual abandonment of the dictionary may occur.

In this situation, the most important measure that the dictionary creator should take in advance is to avoid causing lexicographical delay. This is one of the important criteria for user-friendly dictionaries.

Keywords: Lexicography, dictionary, lexicographical delay, dictionary user

1 https://academic.oup.com/ijl
2 http://lexikos.journals.ac.za/pub
1. Introduction

The concept of lexicographical delay or lexicographical lag, which is referred to as such several times in the Western literature, is an underdiscussed concept.

Thus, when we look at the terminology dictionaries of lexicography (Hartmann & James 1998; Burkhanov 1998), we cannot find these terms. In addition, Robinson (1983), Sterkenburg (2003) and Jackson (2013), who also provide us with a list of lexicological terms, do not mention these terms. There are no such terms in the basic reference sources of lexicography (Zgusta 1971; Jackson 2002; Atkins and Rundell 2008; Svenson 2009, Durkin 2016). In addition, these terms were also not found in searches made in two respectable journals in the field of lexicography (International Journal of Lexicography3, Lexicos4).

However, the term lexicographical delay is mentioned in several sources: Faber&Gonzales (2002), Lasky (2004), Brewer (2007) and Dix (2016).

In this paper, two attempts will be made to define the term lexicographical delay, the reasons for the occurrence of lexicographical delay will be discussed briefly and the issue will be discussed in terms of the user and dictionary relationship. In the conclusions, the methods to avoid lexicographical delay will be highlighted.

2. What is Lexicographical Delay and Why Does it Occur?

Two definitions will be proposed for the term lexicographical delay. One of these is the narrow definition that everyone can agree upon, and the other is the broad (alternative) definition.

The narrow definition of lexicographical delay can be made as follows: Delay experienced in the dictionary entry process of the words which fulfill the criteria to enter the general dictionaries and which are candidates to enter the general dictionaries; or the meanings of headwords available in the dictionaries, which are newly emerged but must be included in the dictionary.

A broad (alternative) definition of the lexicographical delay can be made as follows. Delay occurring as a result of the failure to follow the developments in the main structures forming the dictionaries; namely the macrostructure, microstructure and outside matter; as well as the subcomponents connected to them; namely the front matter, middle matter and back matter.

Both definitions above can be improved and, of course, better definitions can be made. In our paper, we will discuss the lexicographical delay in its narrow definition.

It is necessary to briefly discuss the reasons for the occurrence of lexicographical delay by following its narrow definition.

We know that European dictionary makers are applying various scientific criteria in the process of incorporating candidate lexeme in the general dictionaries as well as incorporating the new meanings of existing headwords in dictionaries. We can summarize these criteria briefly as follows:

i. Frequency of usage

3 https://academic.oup.com/ijl
4 http://lexikos.journals.ac.za/pub
ii. Dispersion of usage
iii. Time endurance
iv. Diachrony boundary
v. Boundary for dialects and standard language
vi. Genericization boundary
vii. Proper noun boundary (Bozkurt, 2016)

Although the above criteria are applied to a significant extent by the dictionary makers, in some cases, it can be seen that the words as well as the new meanings acquired, which should be incorporated in the general dictionaries, are not included in the dictionaries. The failure to meticulously comply with the above criteria can be shown as the primary reason.

In addition, the scope and the quality of the dictionary compilation, the tools and methods used for screening, certain deviations from the target group and purpose principles, etc. can be shown as additional reasons. Cost and physical features for printed dictionaries can also be mentioned as reasons for lexicographical delays.

3. A Discussion on Lexicographical Delay in Social Media

A discussion on lexicographical delay in a blog in social media in very interesting. In this discussion initiated by Brandon Burt, it was found ironic that the English word "meh", which was started to be used as the internet became widespread in 1990s, was incorporated in The Collins English Dictionary 5 years later. 6 Reviewing the content of the discussion, the author criticizes the failure to incorporate the word "meh" in the dictionary, and the lexicographical delay caused by this reason.

4. Dictionary-User Relationship and Lexicographical Delay

According to Nesi (2013:67), "While dictionaries are used to understand a language or to learn new words, the use of a dictionary is generally classified as receptive and productive. This classification is as follows: reading in a receptive writing environment, listening in a receptive speaking environment, writing in a productive writing environment, and speaking in a productive speaking environment.

Conventional monolingual general dictionaries have a structure that is directed towards the reception of the people speaking its mother tongue. Bilingual dictionaries, on the other hand, have both receptive and productive properties.

The general dictionary user uses the dictionary depending on the need that arises while explaining the meanings of words to children when reading newspapers and magazines, listening to radio and watching TV, making homework and writing letters, playing word games, writing reports, studying and also reading for leisure".

Nesi does not directly mention lexicographical delay in this article, but speaks about lexicographical delay in a very indirect way.

The information sought in the dictionary may vary depending on the user profile and needs. For most non-expert users, information other than the basic meaning of the word is not important and does not attract attention. According to Hartmann and James (2002), the information sought by the expert user is divided into two according to the dictionary structure:

i. "A rich access profile; alphabetical index, contents, user manual, special attachments or indexes on the back."
ii. "Rich content (microstructure); format-related information; spelling, phonetics, linguistic knowledge, knowledge of the type of word; information related to meaning; definition, origin information, usage label, sample, cross-reference, etc."

Summarizing the information given above, the dictionaries (in general) are sources we use to solve the problem that arises while performing another task. The referral period is short and the task is urgent. For this reason, the dictionary user wants to solve the problem quickly.

If the dictionary user's referral results in a failure and this is repeated several times, the user will lose confidence in the said dictionary. If this situation persists, the dictionary user will leave the dictionary. Therefore, the dictionary will start losing users.

5. Ways to Avoid Lexicographical Delay

The dictionary user urgently refers to the dictionary in order in order to look at the meaning of an unknown word or to know an unknown meaning of a known word to solve the problem (by itself) while performing another task.

Such referrals of dictionary users sometimes result in failure. Indeed, in Crystal's (1986: 75-76) article prepared on "The Ideal Dictionary for User and Maker", two of Quirk's questions directed at 220 British university students are directly related to this:

i. Person's failure to find what he seeks for
ii. Person's recommendations to improve the dictionary.

Questionnaires with dictionary users have clearly expressed the failure and their related suggestions for improving the dictionary.

The actions that must be taken to minimize or entirely eliminate the initial loss of trust and subsequent loss of customers as a result of lexicographical delay can be summarized as follows.

i. To adopt the understanding of user-friendly dictionary
ii. To make dictionaries open to user contributions
iii. To determine user requests at regular intervals and with different methods
iv. Use a current, extensive and comprehensive collection
v. To strictly apply the criteria used for candidate lexeme as well as for new meanings, and develop new criteria.
vi. To constantly perform updates.

References


Analysis of the Dictionary Needs of China’s Intermediate Business English Learners

Yumei Cui
Guangdong Industry Polytechic College
Amycui11@hotmail.com

Julia Miller
University of Adelaide
julia.miller@adelaide.edu.au

Abstract

Evaluating dictionary qualities is one subtype of user research in lexicography (Svensén 2009) and helps dictionaries to better meet users’ requirements. However, in China, studies of this type predominantly concern printed dictionaries or general dictionaries. So far, little research has focused on online dictionaries, even though many users prefer online lexicographical resources (Gromann & Schnitzer, 2015). The following study addresses the potential need to compile reliable online bilingual business English dictionaries in China. This study is important because many people working or preparing for work in businesses in China must translate documents and product information into and out of English, even though they are not English specialists. Most people are unaware of specialised dictionaries such as the Oxford Business English Dictionary English-Chinese (2015) and turn to freely available online resources, which are not necessarily dictionaries and are often inadequate for their needs. Better training is needed for students preparing for work in businesses in China so that they will be aware of what to look for in a reliable dictionary.

Keywords: specialized dictionaries, business English, online dictionaries, translations, theory of function

1. Introduction

Lexicography is a field in motion. This is not only reflected in the constant updating of dictionary headwords and their meanings, but also by the rapid transformation of dictionary platforms. Ever since the 1960s, when computers were first employed in dictionary making, platforms have transferred from diskette, floppy disc, CD_ROM and software, to the internet. The mushrooming internet-based dictionaries are an important field of research. In Spain, for example, about 8,240 online Spanish dictionaries can be searched from Google (Sanchez & Cantos 2011). There is not yet a statistic on how many online dictionaries are available in China, but the number of
users of a popular online dictionary, youdao, was estimated to exceed 50 million in 2015\(^1\).

Discussions on the pros and cons of electronic dictionaries (including pocket electronic, CD-based, intranet and internet dictionaries) laid a solid basis for lexicographers to dream of future dictionaries\(^2\). It is quite justified for lexicographers to believe that paper dictionaries might eventually be replaced by electronic dictionaries: 'the advantages of the electronic dictionary and the familiarity of today’s young people with electronic devices will eventually relegate the printed notion of “dictionary” to a secondary sense’\(^2\) (Sharpe 1995).

Electronic dictionaries have only developed for three decades; compiling an electronic dictionary of high quality requires new methods, techniques or approaches, but not a new and different theory (Gouws 2011; Tarp 2012). User types and user needs should be considered in designing, compiling and reviewing electronic dictionaries, just as with paper dictionaries. Reviewing electronic dictionaries in accordance with user types and user needs will surely improve their qualities and help the growing number of users choose their suitable dictionaries and make full use of them.

This research concerns only online dictionaries due to the reason that online dictionaries or their mobile phone versions are the most popular dictionaries for current learners in China. Moreover, there is little research focusing on online dictionaries in China. The research is significant for several reasons. Firstly, the online dictionary has become an indispensable tool in English study for general purposes as well as special purposes. Although some teachers still hold a negative attitude towards electronic dictionaries and even oppose their use by students, investigations show that use of electronic dictionaries accounts for a predominant percentage among students of English majors and non-English majors from universities or vocational colleges in China (Deng 2006; Shi & Chen 2007; Li & Bao 2012; Xie 2014; Li 2015). Secondly, learning English for special purposes (ESP) has become increasingly important since 1985, when the Chinese government adjusted the teaching goal for universities and colleges. Students are required to learn ESP in the third year of their study so that they can understand the English used in their special subject fields. Thirdly, English is the first foreign language in China’s import and export companies. Staff from those companies, especially small and medium sized companies, have to communicate with clients in English orally or in written form, even though their English is often only of an intermediate level. Translating or interpreting from time to time is also one of their job responsibilities although they are not officially qualified translators unless they get certificates. A good business English dictionary can be a big help.

---

\(^1\) http://cidian.youdao.com/feature.html
\(^2\) cited in de Schryver (2003:152)
2. Literature Review

The rapid development of science and technology in modern society has engendered an exponential growth of new technical terms or new meanings of terms, resulting in a large demand for specialised dictionaries. Research on specialised dictionaries has also become a hot topic. Studies have covered almost every aspect of specialised dictionaries ranging from the structures such as grammatical information, definition, and equivalent selection to compiling principles and approaches (e.g. Optiz 1983; Fawley 1988; Nielsen 1994; Bergenholtz & Tarp 1995; Zhang & Guo 2010; Tarp 2012; Mugdan 2015). Studies have pointed to the fact that although the number of specialised dictionaries has never disappointed market demand, their qualities are far from satisfactory and the development of specialised lexicography lags behind general lexicography (Moulin 1983; Optiz 1983; Fawley 1988; Landau 2001; Tarp 2012).

Some scholars believe that the slow development is partly attributable to the tradition that the compiling and study of specialised dictionaries were external to lexicography (e.g. Fawley 1988; Bergenholtz & Kaufmann 1997; Tarp 2000). Fawley (1988) thought that specialised dictionaries are not specific enough of the “contextualization” (Fawley 1988:191), no matter what their purposes or how they are compiled, and this vagueness is a result of vague forms of definition. He argued that this phenomenon was partly due to the tradition that specialised dictionaries were only compiled by experts in specific subjects and problems in these dictionaries were only studied by terminologists and terminographers instead of lexicographers. Tarp (2000) cited text articles from a Biotechnology Glossary as a good example to demonstrate that applying only the principles of terminology instead of lexicography to specialised dictionary compilation, the product is not helpful to users at all.

Others put forward a solution in connection with the research results of general lexicography, especially the results of user research, orienting the compiling of specialised dictionaries to user types and user needs (Moulin 1983; Optiz 1983; Landau 2001 Bergenholtz and Nielsen, 2006; Bergenholtz Tarp 2004; Nielsen and Tarp 2009; Fuertes-Olivera and Tarp, 2011 etc.). Optiz (1983) believed that specialised dictionaries, or “technical dictionaries”, can serve various purposes for different users ranging from experts, students or translators to the general public. The multifunction of specialised dictionaries should be addressed by versatile approaches giving rise to different styles of specialised dictionaries (Optiz 1983:164-167). Moulin’s (1983) teaching experience and Fahtis’s (2014) investigation proved that LSP learners and experts have different needs from specialised dictionaries. The theory of function was proposed to be applied to designing and compiling high-quality specialised dictionaries, taking users’ needs into account. “In order to produce a high quality dictionary, the lexicographer must know the users’ primary characteristics” and secondary characteristics (general experience of dictionary use) as well. Different groups of users have different primary and secondary characteristics, experts vs lay people of a specific subject, experienced users and rookie in using dictionaries, advanced language learners and beginners. So users’ characteristics, users’ needs and
users’ situation are the precondition for planning a high quality dictionary (Tarp 2000:195-196).

Researchers in China drew the same conclusion on specialised dictionaries—improvements are desired. Xu (1981), Qian (1995), Wen (1997) and Yang (2008) analysed problems from a holistic view and advocated that both subject experts and lexicographers should be involved in compiling specialised dictionaries. Moreover, specialised dictionary-makers should learn from general lexicography when compiling microstructures, while using principles and purposes of specialised dictionaries so as to maintain specificity in selecting and defining headwords and their meanings. Investigations revealed that students do not use specialised dictionaries often for either language learning or knowledge acquisition owing to their low competency in using dictionaries and the fact that dictionaries often fail to provide necessary information for them (Xie 2006; Zhang & Guo 2009). Therefore, it is necessary to orient specialised dictionaries to users’ needs and research shows that specialised learners’ dictionaries are in demand in China (Zhang and Guo 2009, Wen 1997; Li and Huang 2010). To compile user-friendly specialised dictionaries, corpora should be utilised as they can provide reliable information on selecting terms, definitions and equivalents, word collocation and examples (Ji 2007; Li 2006; Hu and He 2013; Shi and He 2013).

Recent research in China has started to address the area of specialised lexicography. However, there are still areas untapped. First, although business English dictionaries have been published in great quantities, research on them is limited. Up to 2012, only 16 published articles addressed the topic of English dictionaries in the field of business, finance, management or trade (Li and Huang 2010; Shi and He 2012). Secondly, user research in business English dictionaries is not sufficient. Currently, investigations have only been conducted with college and university students whose major is Business English. Research aimed at other types of users with different language and knowledge proficiency has not been published. Thirdly, existing dictionary reviews only cover printed dictionaries, neglecting the fact that online or other types of electronic dictionaries are preferred by business English major students (Zhang, Xie and He 2012; Xie 2014). For example, Hu and He (2013) reviewed Oxford Business English Dictionary for Learners of English and Xu (2014) introduced his Comprehensive English-Chinese Dictionary of Iron and Steel Metallurgy, but no one has conducted research on online business dictionaries.

3. The study

The research in this study was designed under the guidance of function theory. Function theory is a general theory of modern lexicography that was developed to ‘take users, the user needs and user situations as the starting point for all lexicographical theory and practice’ (Bergenholtz & Tarp 2003: 172). This theory focuses on users’ needs and fundamental dictionary design, irrespective of dictionary medium and types. The following study used this theory to investigate intermediate
level users of English for business in order to establish their dictionary needs and to see whether existing online resources are suitable for their purposes.

3.1 Participants

Fifty third-year students from a vocational college in South China and fifty company staff from South China were invited to participate in the investigation. Forty-six students and 41 staff responded. All participants were under the age of 39, and 98% of them aged from 20 to 29, meaning that they had at least some knowledge of computers and computer-based reference works. They had all learned business English and business English translation at vocational colleges in South China.

Vocational college students have different study focus and aims to university students. For example, they are not required to learn much about translation theories or techniques; what they learn from the class is how to translate documents related to trade processes, including company profiles, product instructions, sales confirmations, emails or advertisements. Most of them work in foreign trade companies after graduation. According to statistics at this college, almost 90% of graduates work in import and export companies. Both students and company staff were invited to take part in the investigation in order to see whether these two groups of users have the same requirements and might thus need the same type of dictionary. Most of the company staff in the study work in foreign trade companies, where their daily job responsibilities mostly involve opening up business opportunities, maintaining business relations with clients and coordinating orders. All of these tasks require proficiency in English.

Second language proficiency is an important factor in user-related research (Hulstijn and Atkins 1998). While 17 participants in this study had done a range of different examinations, the majority had passed China’s College English Test Band 4 (CET 4) and Band 6 (CET 6), which require a command of at least 4000 words and 5500 words respectively. Both tests consist of four sections: listening, reading, translating and writing. Fourteen students had passed only CET 4 and 23 had passed both CET 4 and CET 6. Among the company staff, six had only CET 4 certificates and 27 had both CET 4 and CET 6 certificates. CET 4 and 6 are equivalent to intermediate level English.

3.2 Method

3.2.1

Participants were requested to answer questions concerning their dictionary use habits and their expectations of dictionaries. The habits included frequency of dictionary use, preferred dictionary types, dictionaries in use and information looked up. This information was analysed to identify what situation the participants are usually in when they use dictionaries and what role the dictionary plays in their study and work. Questions on users’ expectations were made up of multiple choice questions and open
questions. They were asked to comment and rank the information in the dictionaries they used and list terms they would like to look up in dictionaries. These questions, posed in Chinese, were designed to find out what information these users needed from dictionaries (see Appendix 1 for English version).

3.2.2 Stage 2

In the second stage of the research, five online tools were selected from answers to the question ‘which online dictionaries do you often use?’ These tools are not necessarily dictionaries, but were perceived as such by participants. Five words were examined in these tools to give a snapshot of how existing resources address the users’ needs. These words were closely connected with participants’ study and work and were selected from three sources: student translation exercises; news reports; and work-related materials (see Table 1). Information provided by these tools was compared with the requirements listed by the participants.

<table>
<thead>
<tr>
<th>Headword</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>不良产品 defective products</td>
<td>Student translation exercises</td>
</tr>
<tr>
<td>总监 director or manager, head</td>
<td>Work-related materials</td>
</tr>
<tr>
<td>私募基金 placing, also placement (AmE) sale of a new issue of stock, bonds, or other securities</td>
<td>News reports</td>
</tr>
<tr>
<td>（投资、项目、公司等）的前期工作 front end: the part of business that meets and deals with customer, the beginning of a project, investment</td>
<td>News reports</td>
</tr>
<tr>
<td>露天广告牌 open-air hoardings (BrE) = billboards (AmE)</td>
<td>Work-related materials</td>
</tr>
</tbody>
</table>

4. Analysis of stage 1

4.1 User characteristics

According to function theory, ‘determining the user characteristics is the first step the lexicographer has to take to determine the user needs’ (Bergenholtz and Tarp 2003: 173). Characteristics of specific user groups are relevant to the particular types of dictionaries. The two types of participants in this study had overlapping characteristics, as mentioned above. Their foreign language proficiency was intermediate and they had all had a college education and had good command of their mother tongue. They all
majored in business English, so they had mastered the necessary skills to complete a business transaction and had basic knowledge in this subject field, both in their mother tongue and a foreign language—in this case, English. However, the student group had only classroom knowledge, while the company staff had already worked in trade companies for periods ranging from one week to seven years. The company staff described their job responsibilities as communicating with foreign clients via email or social networks, producing documents and contacting production department or factories to honour orders, or similar activities. Among them, two people particularly pointed out that they had to translate documents such as quotas and allocations. The company staff appeared understandably to have mastered a higher level in both business and their corresponding LSP in both L1 and L2 than the student participants.

As the research was concerned with specialised electronic dictionaries, questions also related to the use of this type of dictionary. Further questions were around participants’ dictionary use habits. When asked to list the number of times they used dictionaries every week, using a scale from one to 7+, 52% of students and 42% of staff ranked 7+, which illustrates that they use dictionaries or dictionary-type tools frequently. With regard to their preferred dictionary medium, the figures are given in Figure 1.

![Figure 1](image)

**Figure 1** Study participants’ preferred dictionary media (total n=87)

Figure 1 reveals that most students and company staff prefer to use electronic dictionaries, although the types of electronic dictionaries they like best are different. Forty-three percent of students favour portable electronic dictionaries (PEDs) and 35% like mobile phone apps. Company staff also prefer PEDs (39%), but their second choice is online dictionaries (34%). PEDs are attractive to both groups due to their convenience and large amount of information. The reason fewer students chose internet tools may lie in the fact that they were discouraged from using electronic and online dictionaries, as these are commonly disparaged by teachers in China.
In a nutshell, slight differences exist, but both groups share many similar characteristics in language competence, subject knowledge and dictionary experience. They can be categorised as ‘semi-expert’ according to function theory (Tarp 2014: 4), because they have a sizable knowledge of their subject-field but only an intermediate level of general and LSP skills in the target language. They are also a group of non-professional dictionary users, according to Hulstijn and Atkins (1998:10). As Tarp (2008) admits, however, it is complicated to establish a user typology, as the criteria are not decisive and vary from dictionary to dictionary. The users under investigation have all learned basic translation skills and one of their current or future job responsibilities is translation. It is, therefore, unfair to call them lay users, and yet they are not fully qualified translators. They are what they are: a group of non-professional dictionary users of intermediate English ability.

4.2 User situations

The dictionary should be a utility tool compiled for addressing users’ information needs in their particular situations (Tarp 2007, 2008). The user-relevant situations suggested by function theory are usually ‘communicative … (production, reception, translation, text revision and marking) … cognitive … (systematic and sporadic) … [and] operative’ (Tarp 2009: 279). There are basically several types of communication-oriented situations where users seek help from dictionaries, namely, production of texts in either their first or another language; reception of texts in either their first or another language; translation of texts from their first language to another language; or translation of texts from another language to the first language (Bergenholtz and Tarp 2003).

When asked when they used dictionaries, more than half the participants chose situations of translating texts, reading English and writing in English (see Figure 4). Talking and listening were less important. The importance of these situations for staff ranks in a sequence of translating, reading, writing, talking and listening, whereas student participants thought reading and translating were equally important.

---

3 cited in Tarp 2009: 279
Participants were then asked what information they consulted in dictionaries. Both students and staff looked up meanings (definitions) most often (see Figure 3).

For the staff, the second most important information were the equivalents, as translating is one of their routine jobs. The students ranked pronunciation second (70%), but only 37% of staff looked up pronunciation. The area least consulted overall...
was background information (cultural or subject information about the headwords). This result is similar to the investigation conducted by Zhang, Xie and He (2012: 119), who found that business background information provided in dictionaries is neglected by student users. Background information is regarded as a type of information sought mostly in knowledge-related situations. However, this knowledge-related user situation is also ‘a new type of passive communication (time-delayed text reception)’ between users and dictionary compilers (Tarp 2005: 8). In practice, the consultation of this type of information is eventually aimed at accurate use of a term in translation.

4.3 User needs

The last part of the questionnaire comprised multiple-choice questions and open questions. Participants were requested to evaluate eleven types of dictionary information and to list terms they would like to see in future dictionaries. It was obvious that the information important to these participants is not always consistent with the information they can find in existing dictionaries. Information ranked by students as helpful was in the sequence of (1) definitions, (2) pronunciation, (3) usage, (4) grammatical information, (5) examples, (6) phrases and idioms, (7) translation equivalents, (8) synonyms and antonyms, (9) collocations, (10) pictorial illustrations and (11) background information. The staff sequence was (1) definitions, (2) usage, (3) examples, (4) grammatical information, (5) pronunciation, (6) translation equivalents, (7) phrases and idioms, (8) collocations, (9) synonyms and antonyms, (10) background information and (11) pictorial illustrations (see Figure 1).

![Figure 4](image)

**Figure 4** Types of dictionary information consulted by study participants (total n=87)

As for terms to be included in a dictionary, participants suggested that an effective dictionary should include as many terms as possible, but when asked about the type of
entries they expected from a dictionary, they suggested neologisms and colloquialisms, including phrases, idioms and slang, in addition to technical terms. This may be because staff job responsibilities involve communicating with clients, and topics are not limited to products. Communication sometimes relates to introducing local culture and customs, discussing current affairs and advertising new products in order to establish and maintain friendly relations with clients. For student users, learning English is still one of their major tasks, so they are keen on improving their ability to speak English fluently; familiarity with colloquialisms is believed to be an effective way to do this.

![Figure 5](image)

**Figure 5** number of people choosing the type of headwords they expect to look up

### 5. Analysis of stage 2

As electronic dictionaries in some form were the most popular choice for most participants, future dictionaries should certainly be in the electronic medium to attract a growing number of users, and Internet provision may be the best option. In answer to the question ‘which online dictionaries have you used?’, 82% of the total 87 participants claimed to have used youdao; 46% had used translate.google.com; 36% had used fanyi.baidu.com; 21% had used OED.com; 17% had used cn.bing.com=dict/; and 14% had used iciba.com. Although translate.google.com and fanyi.baidu.com are not dictionaries, they are often consulted as such by users doing translations. The next step of this research therefore investigated these five popular online tools to find out whether they can meet this type of users’ needs. Five words were selected.

The first stage of the investigation revealed the information participants expect from dictionaries: definitions, pronunciation, grammatical notes, examples, examples, examples.

---

4 see Table earlier.
equivalents, collocations, synonyms/antonyms, background information and pictorial information. When those types of information are sought from the five most popular online translation tools used by Chinese users, information on pronunciation was the most satisfactory for the study participants. However, the most important information—definitions—is not included in all these resources, except for youdao.com, which provides definitions for some headwords and even invites users to define headwords themselves. Grammatical notes are included in all the resources, but in a simple form, only labelling the part of speech, or listing plural forms of countable nouns. Each resource also includes equivalents and examples, some of which are retrieved from webpages. Hence the number of examples in some of the resources, for example youdao.com and cn.bing.com/dict, is too large to be used easily. There is no detailed information on collocations and synonyms in these five online tools. Some provide phrases and synonyms, but none gives detailed explanation on the usage of the phrases, nor any information to discriminate the synonyms. Background information and pictorial illustrations are rarely seen. (See Table 2 for a summary of information.)

**Table** Information included in five popular online tools used for translation in China

<table>
<thead>
<tr>
<th></th>
<th>Youdao.com</th>
<th>Translate.google.com</th>
<th>Fanyi.baidu.com</th>
<th>Iciba.com</th>
<th>cn.bing.com/dict/</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td>Definition for Chinese headwords can be added by users, definitions for English headwords come from <a href="http://wordnet.princeton.edu/">http://wordnet.princeton.edu/</a> and Collins COBUILD English-Chinese Learner’s Dictionary</td>
<td>Definition for English headwords from unidentified source</td>
<td>Not included</td>
<td>Not included</td>
<td>Definitions for English headwords</td>
</tr>
<tr>
<td><strong>Pronunciation</strong></td>
<td>Phonetic symbols and audio pronunciation included</td>
<td>Audio pronunciation included</td>
<td>Phonetic symbols and Audio pronunciation</td>
<td>Not included</td>
<td>Audio pronunciation</td>
</tr>
<tr>
<td><strong>Usage notes</strong></td>
<td>Part of speech tagging, register tagging and information from Collins dictionary</td>
<td>Part of speech tagging,</td>
<td>Part of speech tagging,</td>
<td>Part of speech tagging,</td>
<td>Part of speech tagging</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td>Included</td>
<td>Included</td>
<td>Included</td>
<td>Included</td>
<td>Included</td>
</tr>
</tbody>
</table>
Meaning is always the primary reason for consultation for users of any level, any major and any type of dictionary (for a detailed description of related research see Cowie 1999:178-180). The participants of this research also put meaning in the first place. As definitions are not common in these five resources, but equivalents are, it is important to assess these equivalents to find out whether they are helpful. After all, the less proficient a user’s language is, the more dependent they are on bilingual dictionaries and correct equivalents (Atkins and Varantola 1997). Table c summarises the results.

Table c Translation equivalents of five terms in five online reference tools used in China

<table>
<thead>
<tr>
<th>Term</th>
<th>Included</th>
<th>Included</th>
<th>Included</th>
<th>Included</th>
<th>Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collocation</td>
<td>Phrases Included</td>
<td>Phrases Included</td>
<td>Not included</td>
<td>Phrase included</td>
<td>Not included</td>
</tr>
<tr>
<td>Synonyms and antonyms</td>
<td>Synonyms Included</td>
<td>Synonyms Included</td>
<td>Not included</td>
<td>Not included</td>
<td>Not included</td>
</tr>
<tr>
<td>Background knowledge</td>
<td>Included</td>
<td>Not included</td>
<td>Not included</td>
<td>Not included</td>
<td>Not included</td>
</tr>
<tr>
<td>Pictorial illustration</td>
<td>Not included</td>
<td>Not included</td>
<td>Not included</td>
<td>Not included</td>
<td>Not included</td>
</tr>
</tbody>
</table>

1. The correct translation should be **defective products**

<table>
<thead>
<tr>
<th>Term</th>
<th>Included</th>
<th>Included</th>
<th>Included</th>
<th>Included</th>
<th>Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>youdao.com</td>
<td>translate.google.com</td>
<td>fanyi.baidu.com</td>
<td>iciba.com</td>
<td>cn.bing.com/dict/</td>
<td></td>
</tr>
<tr>
<td>bad product</td>
<td>bad products</td>
<td>defective product</td>
<td>defective product</td>
<td>effective competition; bad goods</td>
<td></td>
</tr>
<tr>
<td>n. majordomo</td>
<td>director</td>
<td>inspector general; chief inspector; majordomo</td>
<td>inspector general; chief inspector; majordomo</td>
<td>inspector general, (web) director; superintendent; inspector-gene</td>
<td></td>
</tr>
</tbody>
</table>
2. The correct translation should be **director** or **manager, head**

3. The correct translation should be **私募基金**

4. The correct translation should be **（投资、项目、公司等）的前期工作**
The results show that in these resources translations are not always standard; denotations in English and Chinese do not necessarily correlate; labels to specify the subject field are not appropriately used; and information on different varieties of English is lacking. These problems are discussed below with reference to Tables 2 and 3. Although only five words were examined here, they give a clear indication of the types of problem that are widespread in these resources.

5.1 Translations are not always standard (word 1)

This research deals with the use of online dictionaries or translation tools in the business field, and the words under investigation are from business or related practices. Therefore, only meanings related to business are studied here. Standardisation is a cornerstone of technical lexicography and ‘using the technical dictionary as an instrument of standardisation is the practice of major international firms’ (Hartmann 1983:166). However, the resources under investigation do not always provide standard equivalents for users. Word 1, 不良产品, indicates products having a fault or faults or not being perfect or complete; in other words, they are defective. The equivalents ‘bad product’ or ‘bad goods’ in three online resources are not standard, because in real practice, ‘bad products’ usually implies the products are ill designed rather than faulty.

5.2 Denotations of English and Chinese do not correlate (word 2)

Word 2, 总监, denoting the head of a department, usually refers to a director, manager, head or officer (e.g. strategy officer) in a company. However, the equivalent provided by these resources, except www.translate.google, is ‘chief inspector’ or ‘inspector general’. In the Oxford Business English Dictionary for Learners of English
(Parkinson 2006), an inspector is defined as ‘a person whose job is to visit factories, restaurants, etc. to check that laws are being obeyed and that standards are acceptable, such as quality inspector, safety inspector’.

5.3 Labels to specify the subject field are not appropriately used (words 3 and 4)

A sound balance of excluding and including headwords and their meanings is of great importance to a quality specialised dictionary (Hartmann 1983, Wang 2000). Technology and techniques from information science will undoubtedly enable online dictionaries to include as many meanings as possible because of the retrieval system and unlimited space, which can be demonstrated by the overall including of meanings for ‘placing’ (word 3) in the resources, with translate.google.com as an exception that provides only one equivalent. On the other hand, a vast number of non-specialist users will find it hard to choose the needed technical meaning from these resources without the help of labelling. In this case, the label ‘finance’ should be attached to the equivalent of 出售，配售 (the equivalent in these online tools for the meaning of selling or selling shares) to match the meaning of ‘sale of a new issue of stock, bonds or other securities’.

A specialised dictionary will always be expected to include technical terms or headwords with technical meanings. When users consult a specialised dictionary, they should not need to decide whether the meaning or equivalent is within the scope of their subject field. However, if the dictionary is designed to meet a wide range of user types for various purposes, successful consultation will have to depend on ‘a reliable system of differentiating marks that guide the users to the correct equivalent of an unfamiliar item’ (Hartmann 1983:169). Tagging or labelling is the established method to serve this purpose in either technical lexicography or general lexicography. For example, in the Oxford Business English Dictionary for Learners of English, word 4 (‘front end’) is labelled as IT to distinguish it from meanings in the business field. However, equivalents of this word are listed as 前段 (the first end) and 前端 (the beginning phase or stage) in four online resources, users of which may not understand the specificity of the meaning. In youdao.com, ‘front-end’ is labelled in the subject of business, but the equivalent 商店前端 actually means the front part of a shop or store, which is not specific to business at all.

5.4 Information on different variants of English is lacking (word 5)

American English and British English are sometimes different in terms of pronunciation, spelling and lexicon. In almost every paper dictionary, especially learners’ dictionaries, tags and cross-references are effective techniques to introduce this type of information. This should be easy to do in an online dictionary, with the
potential to use hyperlinks between entries. However, word 5, ‘open-air hoardings/billboards’, is not given with both its British and American English variants.

6. Conclusion

The present study investigated two groups of non-professional, semi-specialist users who are studying business English or conducting import and export transactions. These two groups need to consult specialised learners’ dictionaries for communication-related information, including reading business English; writing emails or business documents in English; and translating business documents such as company profiles and product descriptions from Chinese to English or occasionally from English to Chinese. The two groups of users have slight differences in preferred dictionary types and identifying what information is important to them in a dictionary. However, they belong to the same group of user types owing to their similar language and knowledge levels, and user situations. Their dictionary needs often overlap. Besides definitions, equivalents and pronunciation, they also need examples, grammatical notes and collocations to help them understand a headword and use it correctly and appropriately in oral and written communication. The unique advantages of electronic dictionaries, coupled with the popularity of computers and the internet among young users, enable electronic dictionaries, particularly online dictionaries, to be the first choice for this type of user.

However, this group of users don’t know much about dictionaries. They are not familiar with differences between dictionaries and online resources. The major purpose of their consulting dictionaries is to find out equivalents for words they use at work or study. In their opinion, meaning is the most important information in a dictionary. However, because secondary information is frequently lacking or insufficient, and equivalents in these resources are often not standard, precise or subject specific, existing online dictionaries cannot provide individualized information to them. Although research has proved that less proficient language users are more dependent on bilingual or bilingualised dictionaries (Laufer & Kimmel 1997), most current online dictionaries in China cannot meet the needs of beginners and intermediate learners. Zhang and Guo (2010) have already pointed out a demand for specialised learner’s dictionaries in China, but the problem has not yet been well addressed. Users also need to be educated in the comparative value of different online resources, and future research could examine methods for evaluating the effectiveness of such education among vocational college students. Chinese lexicographers need a full picture of LSP learners’ requirements so they can collaborate with experts from different subject fields to provide really helpful tools for China’s LSP dictionary users.
References

Dictionaries


Other Literature


Intermediate University Students’ Motivation to Receive Instruction for English-Japanese Dictionaries

Hiroe Kubota

Shibaura Institute of Technology
kubota.shibaura@gmail.com

Abstract

This paper investigates the degree to which intermediate Japanese university students learning English desire to receive instruction in the use of English-Japanese dictionaries. The present study aims to: 1) compare motivation levels for dictionary instruction in general (asked as one question item) and for individual skills; 2) examine motivation levels for dictionary instruction in each of 30 skills for English-Japanese dictionaries; 3) investigate relationships between five learner factors and motivation levels for dictionary instruction; and 4) compare students’ motivation with the need, as recognized by 19 EFL teachers in Japan, to provide dictionary instruction.

The survey results show a significantly higher motivation to receive instruction in 20 individual dictionary skills than in dictionary instruction in general. The examination in individual skills reveals that the highest motivation to receive dictionary instruction, in which 30.7% of the students marked the highest motivation responses, is seen in finding the definitions of multiple words. The comparison between students’ motivation for and the needs for dictionary instruction that teachers have recognized reveals certain skills in each of the following four categories: 1) high motivation and recognized needs, 2) high motivation and low recognized needs, 3) low motivation and high recognized needs, and 4) low motivation and low recognized needs. The results could offer educational implications for effective and efficient dictionary instruction.

Keywords: Dictionary instruction, English-Japanese dictionary, motivation

1. Introduction

Dictionary use facilitates foreign language learning, and the mastery of dictionary skills promotes autonomous learning (Hartmann 2001). However, dictionary consultation involves cognitively complex tasks (Hartmann 1989; Tono 2001), which necessitates dictionary instruction.

Limiting the scope of the present discussion to Japan, the Japanese national curriculum stipulates the necessity of dictionary instruction in foreign language education throughout the secondary level (MEXT 2008a; MEXT 2008b). The official
guidebook for this curriculum also stresses the importance of dictionary instruction in fostering positive attitudes toward autonomous learning (MEXT 2009). Despite this, several studies have indicated a lack of dictionary skills among Japanese university students, suggesting the necessity of continued dictionary instruction at the university level (Hatakeyama 2001; Tono 2001; Sekiyama 2005; Prichard 2008; Nakayama and Osaki 2009).

In the field of education, motivation is widely acknowledged as one of the key factors for successful learning or skill acquisition. Since instruction is a part of the learning process, it is important for educators to pay attention to learners’ motivation and then to promote it through their teaching. However, very few studies have been conducted on motivation and dictionary skill learning; this paper investigates that topic.

Researchers have proposed many theories on motivation. Dornyei and Ushioda (2011) present three common motivation factors regarding foreign language learning: the selection of a certain activity, the intended effort to achieve a goal, and persistence in pursuing it. Ryan (2008) and Taguchi et al. (2009), investigating the motivation of Japanese learners of English, specify the effort factor as the “intended learning effort,” which is ‘learners’ perceptions of their current efforts to learn and their possible intended future efforts’ (Ryan 2009: 147) in order to facilitate their research in Japan. Both of the surveys by Ryan (2008) and Taguchi et al. (2009) included questions to measure participants’ motivation, one of these was the desire to take classes in the future; ‘If an English course was offered at a university or somewhere else in the future, I would like to take it.' (Ryan 2009: 292) Based on this question, a students’ desire to take classes or instruction can be considered as one factor of motivation in the Japanese setting. Thus, this study investigates Japanese students’ desires to receive dictionary instruction, with the aim of understanding their motivation to improve their dictionary skills.

Japanese learners of English commonly use English-Japanese dictionaries, Japanese-English dictionaries, and English-English dictionaries; however, this paper focuses on English-Japanese dictionaries, as they are used the most commonly and frequently. Further, this paper deals with the motivation for dictionary instruction by university students who have already acquired some dictionary skills and thus are expected to show a complex pattern of motivation.

This paper addresses the following four questions.

1) Does the phrasing of the question affect the reported level of motivation for dictionary instruction? Specifically, does asking about dictionary instruction in general elicit a different answer from questions asking about individual dictionary skills?
2) Which English-Japanese dictionary skills do Japanese students show high and low levels of desire to receive instruction for?
3) Do learner factors show a relationship with motivational level? If so, to which skills as these related?
4) To what extent does students’ motivation to receive dictionary instruction match the need for instruction as perceived by teachers?
2. Previous Research

Despite the importance of motivation in dictionary skill instruction, research on this issue is, to the best of the author’s knowledge, very limited. Huang (2003) interviewed Taiwanese university students on the dictionary skills for which they hoped to receive dictionary instruction. Udagawa and Kubota (2012) investigated the level of students’ desires to acquire English-Japanese dictionary skills and the motivation to receive dictionary instruction among freshmen at a Japanese university. In 22 of 27 skills, the average level of reported wishes was above the midpoint, whereas the motivation to receive dictionary instruction in class was below the midpoint. This gap can be attributed to the obvious difference between the motivational factors investigated. Wishes are at an early stage of motivation (Dornyei & Otto, 1998) and do not require effort; however, receiving instruction necessitates the investment of students’ time and effort. Thus, it is understandable that the motivation to receive dictionary instruction was lower than the wishes to acquire dictionary skills.

There is another factor relevant to this gap. While the students’ wishes were investigated by asking a separate question for each skill, their motivation for dictionary instruction was measured with one general question: “I would like dictionary instruction to be offered in class.” This difference between the questions may affect the ways students thought about the question and thus may have influenced their responses. More specifically, when using a general question, it may be unclear which dictionary skills are being referred to; therefore, the students could reflect their various perceptions of dictionary instruction, including their past experience. Thus this study first investigates whether reported motivation levels differ between the general question and specific questions.

Next, the results in the study by Udagawa and Kubota (2012) show that, while students’ levels of wishes to acquire English-Japanese skills are generally high, there is still a clear variation among wishes for these skills. The study also found statistically significant correlations between the wish levels for certain skills and learner factors. Based on this, it is possible to observe some variances in the motivation level for dictionary instruction and relationships with some learner factors. Thus, the current paper examines: 1) motivation levels for dictionary instruction as an individual skill, focusing on skills with high and low motivation, and 2) the correlation with learner factors.

Lastly, while investigating learners’ motivation levels for dictionary instruction can offer useful suggestions for this practice, the reality is that learners are not necessarily aware of the dictionary skills they need to ensure effective dictionary consultation. Thus, simply reflecting learners’ motivation in dictionary instruction may not be sufficient for effective learning. Further, in terms of actual dictionary consultation, the importance and frequency of use of dictionary skills vary, with some being essential while others are additional. However, teachers have ample experience and knowledge concerning which skills are immediately useful for their students. Therefore, comparing students’ motivation with teachers’ perceptions of the importance of individual skills may lead to more concrete recommendations. For this reason, the
current paper attempts to identify four types of dictionary skills: 1) skills with high motivation and high recognized needs, 2) skills with low motivation and high recognized needs, 3) skills with high motivation and low recognized needs, and 4) skills with low motivation and low recognized needs.

3. Research Method

Based on the comprehensive list of dictionary skills for foreign language at the tertiary level compiled by Nesi (2003), 30 skills for English-Japanese dictionaries were prepared. Using them, three separate surveys were created: two for students and another for teachers. One of the student surveys asked how much they would like to have dictionary instruction offered for each skill, as indicated by a 6-point Likert scale. (In this paper, all the items in Likert scales are reverse coded so that high scores represent high motivation and perceived need.) The general guideline of the survey specified that the instruction might take any form, such as in the classroom, through handout exercises, online practice, and so on. Five learner factors were also added: 1) past dictionary instruction experience, 2) the degree to which they like English, 3) the degree to which they would like to study English, 4) types of dictionaries they currently use, and 5) TOEIC score.

The other student survey was primarily conducted for a different purpose, but contained one item to measure the participants’ general motivation level to receive dictionary instruction, as indicated using a 6-point Likert scale. Only this question is analyzed in the present paper.

The teachers’ survey covered the same 30 skills and asked for the teacher’s assessment of his or her students’ need for dictionary instruction in each skill. Other questions asked about their own experiences with dictionary instruction, classes they have taught, and their students’ proficiency levels.

The first student survey was conducted in late June 2015, and the second was conducted in mid-July of the same year. All students were freshmen and sophomores at two Japanese universities. Their English proficiency ranged from lower intermediate to high intermediate. Valid responses were obtained from 173 students. The teachers’ survey was distributed, and 19 valid responses were collected between July 2015 and March 2016. Twelve instructors were at the same school as the students participating in this survey; the remaining seven instructors teach students of a similar level at other schools.

4. Results and Discussion

4.1 Research Question 1: influence of phrasing of the question on motivation level

The results for the question about students’ general motivation for dictionary instruction had both a mode and median of 3, indicating relatively low motivation level.
The response distribution is shown in Table 1.

**Table 1** Distribution of student responses toward dictionary instruction in general (one question item; n=137).

<table>
<thead>
<tr>
<th>Response #1</th>
<th>Response #2</th>
<th>Response #3</th>
<th>Response #4</th>
<th>Response #5</th>
<th>Response #6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>9</td>
<td>25</td>
<td>44</td>
<td>39</td>
<td>17</td>
</tr>
<tr>
<td>Percentage</td>
<td>6.6%</td>
<td>18.2%</td>
<td>32.1%</td>
<td>28.5%</td>
<td>12.4%</td>
</tr>
</tbody>
</table>

Responses indicating motivation level for dictionary instruction in individual skills had modes ranging from 2 to 6 and medians ranging from 3 to 5, with Response 4 having the highest value for both. The distributions of these modes and medians, compared to the mode and median for general dictionary instruction, are shown in Figures 1 and 2.

![Figure 1](image1)  
**Figure 1** Distribution of modes for individual dictionary skills (n=30)
Mann-Whitney U tests with Holm adjustments were conducted to determine whether any differences are statistically significant. The results show significant differences in 20 out of the 30 dictionary skills, with 19 higher and 1 lower than the results for the general question. Of these 20 skills, 3 are of large effect, 13 are of medium effect, and 4 are of small effect. These results, along with descriptive statistics for each skill, are shown in Table 2. The skills are arranged in the order of combined percentages for Responses 5 and 6, which clearly show high motivation levels for receiving dictionary instruction.

**Table 2** Results from Mann-Whitney U tests between one general question and each individual question, including effect size and descriptive statistics (n=137)

<table>
<thead>
<tr>
<th>Skill Description</th>
<th>U value</th>
<th>Effect size (r)</th>
<th>Median</th>
<th>Mode</th>
<th>Min.</th>
<th>Max.</th>
<th>Combined % of Responses 5 &amp; 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consulting multiple-word items</td>
<td>4645.5**</td>
<td>-0.63</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>59.1</td>
</tr>
<tr>
<td>Using information on collocation</td>
<td>4774.5**</td>
<td>-0.62</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>6</td>
<td>54.7</td>
</tr>
<tr>
<td>Inferring from context</td>
<td>5408.0**</td>
<td>-0.53</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>47.4</td>
</tr>
<tr>
<td>Locating derivative forms</td>
<td>5669.5**</td>
<td>-0.50</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>6</td>
<td>46.7</td>
</tr>
<tr>
<td>Understanding the Japanese definition</td>
<td>6348.5**</td>
<td>-0.41</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>6</td>
<td>45.3</td>
</tr>
<tr>
<td>Distinguishing among homographs</td>
<td>5714.5**</td>
<td>-0.49</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>43.1</td>
</tr>
<tr>
<td></td>
<td>Using sample sentences</td>
<td>6276.5**</td>
<td>-0.42</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>---</td>
<td>-------------------------</td>
<td>----------</td>
<td>-------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>8</td>
<td>Using information on proper inflection</td>
<td>6714.5**</td>
<td>-0.36</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>Distinguishing between relevant and irrelevant information</td>
<td>6723.0**</td>
<td>-0.36</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>Using information on limited usage</td>
<td>6567.5**</td>
<td>-0.38</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>Selecting dictionaries that match consultation needs</td>
<td>7264.0*</td>
<td>-0.28</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>Using general dictionary conventions</td>
<td>6852.5**</td>
<td>-0.34</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>Using information on sentence structure</td>
<td>6785.0**</td>
<td>-0.35</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>Using information on etymology</td>
<td>7287.0*</td>
<td>-0.28</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>Understanding the appropriate forms of the headwords</td>
<td>7339.0*</td>
<td>-0.27</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>16</td>
<td>Using pronunciation symbols</td>
<td>6984.0**</td>
<td>-0.32</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>16</td>
<td>Deciding the target word to be consulted</td>
<td>7289.5*</td>
<td>-0.28</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>16</td>
<td>Using wild cards</td>
<td>8431.0</td>
<td>-0.13</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>19</td>
<td>Finding correct spellings</td>
<td>7628.0</td>
<td>-0.24</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>20</td>
<td>Confirming the definition within the appropriate context</td>
<td>6961.0**</td>
<td>-0.33</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>21</td>
<td>Using conventions for sentence structure</td>
<td>6983.5**</td>
<td>-0.32</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>22</td>
<td>Understanding the microstructure of an entry</td>
<td>7974.0</td>
<td>-0.19</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>23</td>
<td>Using reference information in printed dictionaries</td>
<td>8733.5</td>
<td>-0.09</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>24</td>
<td>Identifying the parts of speech of the target word</td>
<td>8214.5</td>
<td>-0.16</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>24</td>
<td>Using conventions for parts of speech</td>
<td>8476.0</td>
<td>-0.12</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>26</td>
<td>Selecting the target word from unknown words for consultation</td>
<td>8893.5</td>
<td>-0.07</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>27</td>
<td>Using front matter and appendices in printed dictionaries</td>
<td>8490.0</td>
<td>-0.12</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>28</td>
<td>Using the jump function</td>
<td>8136.0</td>
<td>-0.17</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>29</td>
<td>Understanding the</td>
<td>8610.5</td>
<td>-0.10</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
</tbody>
</table>
Based on these results, we cannot say that students’ general motivation level to receive dictionary instruction is equivalent to that in individual skills; this supports the necessity of investigating motivation levels regarding individual dictionary skills for instruction.

These results are comparable to a situation to which many teachers can probably relate. If low-intermediate engineering students are asked how much they would like to memorize vocabulary and take vocabulary quizzes every week, which is a general question, their motivation level is likely to be low. However, if they are specifically asked how much they would like to memorize ten English words that are frequently used in the TOEIC test and have vocabulary quizzes every week, their motivation level is likely to be higher. (In Japan, higher TOEIC scores are advantageous in job hunting.) In a similar way, the differences between the general and specific questions in this study are understandable, although further investigation is necessary to clarify the reasons behind this gap.

4.2 Research Question 2: skills with students’ high motivation and low motivations

4.2.1 High motivation skills

As is seen in Figures 1 and 2, the modes of motivation level for receiving dictionary instruction for all of the individual skills ranged from 2 to 6, and the medians ranged from 3 to 5. In terms of the distribution of modes, one skill scored 6, the highest motivation level, six skills scored 5, 18 skills scored 4, four skills scored 3, 1 skill scored 2 and no skill scored 1. The combined percentage of Responses of 5 and 6, which show a strong motivation for receiving dictionary instruction, ranged from 59.1% (multiple-word consultation) to 12.4% (understanding the alphabetical order and the keyboard layout). Based on these results, it is possible to say that the motivation to receive dictionary instruction for individual skills is not enthusiastically high; however, it is not very low either, as three fourths of skills received positive motivation scores. In particular, approximately one fourth of the skills received clearly high motivation with the modes being 5. The skills with high and low motivation skills will be examined below. For analysis purposes, skills with the combined percentages of 40% and above are treated as “High motivation,” while those with 25% and below are treated as “Low motivation.” The skills with high and low motivation are shown in Tables 3 and 4, respectively.
Table 3 High motivation Skills with combined percentages of Responses 5 and 6 being 40% and above (n=137)

<table>
<thead>
<tr>
<th>Skill</th>
<th>Combined %</th>
<th>median</th>
<th>mode</th>
<th>min.</th>
<th>max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consulting multiple-word items</td>
<td>59.1</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Using collocation</td>
<td>54.7</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Inferring from context</td>
<td>47.4</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Locating derivative forms</td>
<td>46.7</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Understanding the Japanese definition</td>
<td>45.3</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Distinguishing among homographs</td>
<td>43.1</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Using sample sentences</td>
<td>41.6</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Understanding proper inflection</td>
<td>40.1</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
</tbody>
</table>

Consultation of multiple-word items received the highest combined percentage of responses of 5 and 6 (59.1%), indicating that approximately 60% of the students clearly expressed a high level of motivation to improve this skill. In particular, the mode was 6, (“certainly would like instruction to be offered”), accounting for 30.1% of all responses regarding this skill. This skill also showed a high motivation in an earlier study that asked different students about their desire to acquire dictionary skills (Udagawa & Kubota, 2012). Multi-word consultation is regarded as an important dictionary skill, as can be seen in English-Japanese-dictionary-use guidebooks for university students (Murata 2006; Murata 2007), where idioms are often introduced immediately after the basic explanation of word definitions. Many studies have dealt with difficulty of idiom consultation (Tono 2001; Sekiyama 2005; Szczepaniak 2006; Nakayma and Osaki 2009). Given the highly recognized complexity of this skill, its importance, and students’ high motivation to master it, dictionary instruction in this skill should be both popular among and useful for students.

The use of collocation ranked second, with a combined percentage for responses 5 and 6 of 54.7%. These two top skills deal with multiple words, and their importance has been widely acknowledged (Nation 2001; Chen 2016).

Inferences from context ranked third in the survey, showing high popularity among students; however, according to Laufer (1989) and Liu and Nation (1985), correctly inferring a word’s meaning from context requires knowledge of over 95% of the vocabulary in the passage. Therefore, it is necessary to think very carefully about the ideal teaching method for this skill.

Understanding the Japanese definition of an look up item ranked fifth in the survey. This may be an unexpected result for some teachers, as the definitions in Japanese-English dictionaries are written in the students’ first language. However, this high motivation suggests that some students may not be able to understand the definitions, which tend to use succinct expressions. This result also suggests that students are not aware that, in the case of electronic dictionaries, using the “jump”
function allows a user to find an explanation of any unknown word in a Japanese dictionary.

The use of sample sentences, ranked seventh, can facilitate the understanding of the meaning of a word in context and provide a model for its use, including collocation information for writing in English. Considering that collocation instruction itself ranked second, these results suggest that students are keen to learn about appropriate combinations of English words.

Locating word’s derivative forms and distinguishing among homographs ranked fourth and sixth, respectively. These skills are concerned with locating target words; however, they require grammatical knowledge, such as knowledge of sentence structure, the parts of speech, and suffixes and prefixes.

Understanding the proper inflection of a word from dictionaries ranked eighth among the skills. This is a relatively straightforward operation, once users understand the microstructure of an entry. However, students show relatively low motivation to learn the subskills necessary for these 3 skills, as is shown below, and resolving these tensions is important in the classroom.

4.2.2 Low motivation skills

Table 4 Low Motivation Skills with combined percentages of responses 5 and 6 being 25% and below (n=137)

<table>
<thead>
<tr>
<th></th>
<th>median</th>
<th>mode</th>
<th>min.</th>
<th>max.</th>
<th>combined % of 5 &amp; 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>23.4</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>21.9</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>21.9</td>
</tr>
<tr>
<td>26</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>20.4</td>
</tr>
<tr>
<td>27</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>19.7</td>
</tr>
<tr>
<td>28</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>16.8</td>
</tr>
<tr>
<td>29</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>14.6</td>
</tr>
<tr>
<td>30</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>12.4</td>
</tr>
</tbody>
</table>

Among the skills with low motivation (those drawing 25% or less combined percentage for students’ clear motivation), understanding alphabetical order and
English-language keyboard layout had the lowest motivation (12.4%). This is understandable, since this skill is a basic and fundamental one, and university students are likely to have already mastered it. In fact, in the earlier study of university students (Udagawa & Kubota, 2012), this skill scored the highest on self-perceived ability.

Another trend observable in Table 4 is that both of the skills concerned with distinguishing parts of speech show very similar and low motivations. Udagawa and Kubota (2012) found similar results, along with low perceived ability when using these skills. Tono (2001) also reports that non-language major Japanese university students did not use part of speech information in their consultation process.

The results on the selection of a target word (20.4%) are similar those from reported by Prichard (2008): students failed to use this skill. Looking at dictionary-specific skills, both skills for printed dictionaries—the use of reference information and the front matter and appendix—and one of the two skills for electronic dictionaries—using the “jump” function—show low motivation.

Comparing the skills with high motivation to those with low motivation, it can be seen that some of the low motivation skills are necessary to master some of the high motivation skills. In particular, given the structure of most electronic dictionaries, the jump function is extremely useful for understanding the Japanese definition of an English target word; it allows a user to jump straight to the definition of any unknown word. Similarly, understanding the parts of speech and conventions for parts of speech are often useful skills for distinguishing homographs. Further investigation is necessary to examine the circumstances behind these gaps. However, a similar pattern was also observed in our study of students’ wishes to learn (Udagawa and Kubota, 2012).

As this section demonstrates, the levels of motivation to receive dictionary instruction in individual cases covers a wide range. The fact that some low motivation skills are necessary for the mastery of high motivation skills is both a challenge and an opportunity in the classroom. As level of motivation is an important factor in successful learning, research is needed to ensure that enthusiasm for high motivation skills supports the study of low motivation ones. Students in this study are likely to learn the high motivation skills shown in Table 3 successfully, driven by their motivation. However, during the learning process, if the low motivation subskills can be presented as necessary steps toward acquiring high motivation skills, it may encourage students to put in the necessary effort.

Further research is also necessary to identify on which particular subskills university students would like to receive more instruction, as dictionary consultation is a complex process (Tono, 2011) that involves multiple types of knowledge and skills.

4.3 Research Question 3: relationship between learner factors and motivation level
This section presents and discusses the investigation of the relationship between learner factors and their motivation to receive dictionary instruction in individual skills. The learner factors investigated are: 1) whether students received dictionary instruction prior to attending secondary school; 2) how much they like English (evaluated using a
6-point Likert scale); 3) how much they would like to study English (evaluated using a 6-point Likert scale); 4) types of dictionaries they use; and 5) TOEIC scores. For analysis purposes, the two learner factors that use a 6-point Likert scale were categorized into two groups – a high group and a low group – for each. Students’ TOEIC scores, a continuous variable, were also divided into two groups. With regards to dictionary types, more than half of the students reported using many types of dictionaries; thus the comparison was made between students who use a certain type of dictionary and those who do not for each dictionary type. Mann-Whitney U tests with Holm adjustments were applied for each learner characteristic, to examine the differences in motivational level according to each variable between two groups.

4.3.1 Previous experience with dictionary instruction

Of the 137 participants in this study, 45 had received dictionary instruction before, which is one third of all participants. The details of the schools in which they received instruction, and lengths of instruction, are shown in Table 5.

<table>
<thead>
<tr>
<th>Table 5</th>
<th>Length of dictionary instruction students received in junior high school and high school (n=45)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Junior High School</td>
</tr>
<tr>
<td>-----------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Approximately 1 hour or less</td>
<td>23</td>
</tr>
<tr>
<td>Approximately 2 hours or more</td>
<td>10</td>
</tr>
<tr>
<td>Total number of students with dictionary instruction experience</td>
<td>33</td>
</tr>
</tbody>
</table>

The results of the Mann-Whitney U test with Holm adjustments showed no significant differences among the 30 skills between students with or without prior dictionary instruction. Similar results were found regarding students’ wish levels to study English (Udaga & Kubota, 2012), with no significant differences among the 27 skills investigated. This suggests that the relationship between students’ experience with dictionary instruction and their motivation to receive such instruction is weak.

4.3.2 How much students like English (evaluated using a 6-point Likert scale)

The data distribution of descriptive statistics regarding how much the students like English is shown in Figure 3.
For analysis purposes, the data were categorized into two groups: responses of 4, 5, and 6 (n=100) that indicated a high degree of appeal, and those of 1, 2, and 3 (n=37), which indicated low degrees of appeal. The results of the Mann-Whitney U test with Holm adjustments revealed no significant differences between these two groups for any individual skill. As the number of students in each group was not well balanced, the data were retested using the same procedure, but between different two groups: responses of 5 and 6 (n=59), which showed clear a liking of English, and responses of 1, 2, 3, or 4 (n=78), which did not show a clear liking. The results of this re-test also showed no significant differences for any of the skills, confirming the previous results. An earlier study on students’ wish levels (Udagawa & Kubota, 2012) also found similar results.

### 4.3.3 How much students would like to study English

The distribution of data on the descriptive statistics regarding how much students would like to study English is shown in Figure 4.
Figure 4 Distribution of responses for how much students would like to study English (n=137)

We investigated the relationship between the motivation to receive dictionary instruction and the motivation to study English using the same methodology as that used for the students’ liking of English, including the re-test procedure. The results were also the same, revealing no significant results at all. This suggests a weak relationship between students’ motivation to study English and their motivation to receive dictionary instruction in the 30 skills addressed in the present study.

These results are somewhat unexpected, as motivation to receive dictionary instruction can be considered a part of a student’s motivation to study English. One possibility is that there is a real relationship between the two variables. However, the results of current study suggest that the relationship between the motivation to learn English and the motivation to receive dictionary instruction is much weaker than might be expected.

4.4.4 Type of dictionary they use

Four types of dictionaries were investigated in this study: electronic, paper-bound, online, and ones on/in mobile phones. Table 6 shows the number of students who use each type of dictionary.

<table>
<thead>
<tr>
<th>Type of Dictionary</th>
<th>Number of students reporting use</th>
<th>Percentage of users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electronic dictionaries</td>
<td>119</td>
<td>86.9</td>
</tr>
<tr>
<td>Printed dictionaries</td>
<td>19</td>
<td>13.9</td>
</tr>
<tr>
<td>Online dictionaries</td>
<td>39</td>
<td>28.7</td>
</tr>
</tbody>
</table>
As Table 6 shows, many students use multiple types of dictionaries; for analysis, Mann-Whitney U tests with Holm adjustments were conducted on those who use a certain kind of dictionary and those who do not for each type of dictionary. The results indicated no significant differences among the use of any type of dictionary and interest in any of the skills.

These results suggest that the relationship between types of dictionaries and the motivation level for dictionary instruction is weak. In fact, Udagawa and Kubota (2012) found similar results in their study of students’ wish levels.

### 4.4.5 TOEIC scores

The participants’ TOEIC scores ranged from 300 to 705, with a mean of 470.5 and an S.D. of 106.1. The median is 450. The data were divided into two groups for analysis: a high-scoring group, with scores of and above the average of 470 (n= 64), and a low-scoring group, whose results were below the average score (n=73). A Mann-Whitney U test with Holm adjustment revealed significantly higher results (u= 1590.5, p<.05) with medium effect (r=.33) in the high-scoring group than among the low-scoring group for the skills understanding alphabetical order and keyboard layout. As this result was surprising, a retest was conducted with two different groups divided at the median, 450. The results were the same, with a significantly higher motivation found among the high-scoring group for the alphabetical order and keyboard layout skills.

Although it is only speculation, this result may be due to a high level of interest among the high-scoring group to understand the function of all the buttons on an electronic dictionary.

### 4.5 Research Question 4: teachers’ perceived need and students’ motivation for dictionary instruction

Comparing teachers’ perceptions of the necessity of dictionary instruction with their students’ motivation to acquire individual skills can be useful, in that it can offer a clearer understanding of the skills students need. Students are not necessarily aware of what skills will be the most useful to them, while teachers are usually knowledgeable about both students’ needs and dictionary use itself.

The results of the survey on teachers’ perceived necessity, as seen in Appendix, are strongly and positively skewed compared to the students’ results. This is understandable, as many teachers usually hope that their students will learn and master anything educational presented to them. Thus, for analysis purposes, rather than comparing the two sets of raw data directly, we use rankings, which show skills’ relative importance within each data set, as the basis / means of comparison.

The result of Spearman’s rank correlation indicates a significant coefficient of .39 (p <.05), suggesting a weak correlation between teachers’ perceived necessity and students’ motivation.
students’ motivation for dictionary instruction. In order to compare these two pieces of data further, this paper attempts to categorize them into four types of skills: 1) skills both teachers and students regard as being “High necessity”; 2) skills with high teachers; perceived necessity and low student motivation; 3) skills with low teacher-perceived necessity and high student motivation; and 4) skills both teachers and students regard as being “Low necessity.” For analysis purposes, high motivation skills, low motivation skills, high perceived-necessity skills, and low necessity skills are defined as follows:

Combined percentages of responses 5 and 6 in student data
High motivation: 40% and over
Low motivation: 25% and below
Combined percentages of responses 5 and 6 in teacher data
High perceived necessity: 60% and over
Low perceived necessity: 30% and below

4.5.1 Skills with High motivation and High necessity

Table 7 Skills with High motivation and High necessity

<table>
<thead>
<tr>
<th>Skill</th>
<th>Students’ motivation</th>
<th>Teachers’ perceived necessity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rank</td>
<td>%</td>
</tr>
<tr>
<td>Multiple word consultation</td>
<td>1</td>
<td>59.1</td>
</tr>
<tr>
<td>Inferences from context</td>
<td>3</td>
<td>47.4</td>
</tr>
<tr>
<td>Use of sample sentences</td>
<td>7</td>
<td>41.6</td>
</tr>
</tbody>
</table>

With such high necessity perceived by teachers and high motivation by students, the three skills in Table 7 are highly recommended to be dealt with in the classroom, and the outcome of their instruction is likely to be successful.

4.5.2 Skills with Low motivation and High necessity

Table 8 Skills with Low motivation and High necessity

<table>
<thead>
<tr>
<th>Skill</th>
<th>Students’ motivation</th>
<th>Teachers’ perceived necessity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rank</td>
<td>%</td>
</tr>
<tr>
<td>Understanding parts of speech</td>
<td>24</td>
<td>21.9</td>
</tr>
<tr>
<td>Using conventions for parts of speech</td>
<td>24</td>
<td>21.9</td>
</tr>
</tbody>
</table>

Both skills in this study dealing with parts of speech show a large gap between teachers’ recognized needs and students’ motivation to acquire them. It is also assumed
that teachers know that understanding the parts of speech and their conventions can facilitate distinguishing homographs and finding derivative forms; these are both skills with reported high student motivation. Thus, as an educational suggestion, given teachers’ high recognition of necessity and students’ low motivation for instruction, it is important to find a way to engage students in the learning of these skills. One result from this study could become a suggestion for how to motivate students. As the skills in Table 7 are necessary for two high motivation skills—distinguishing homographs and finding derivatives—teachers could present these subskills as a part of other high motivation skills.

4.5.3 Skills with High motivation and Low necessity

No skills were found with to have low teacher-perceived necessity and high student motivation within the frameworks defined in this paper. However, teachers ranked some of the high motivation skills lower than 17th, which may indicate additional attention is needed.

4.5.4 Skills with Low motivation and Low necessity

<table>
<thead>
<tr>
<th>Skill</th>
<th>Students’ motivation</th>
<th>Teachers’ perceived necessity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use of the jump function</td>
<td>28</td>
<td>16.8</td>
</tr>
<tr>
<td>Understanding alphabetical order and keyboard layout</td>
<td>30</td>
<td>12.4</td>
</tr>
</tbody>
</table>

The two skills that students showed a low motivation to pursue and which teachers perceive as being of low necessity are the use of the “jump” function on an electronic dictionary and understanding English alphabetical order and keyboard layout. Alphabetical order is normally acquired very early in the process of learning English, and at the university level, neither students nor teachers think further instruction in this area is necessary. However, students with higher TOEIC scores showed significantly higher motivation to receive dictionary instruction in this skill and that for the keyboard layout than those with lower scores. Further investigations into the reason for this difference would be useful.

While the use of the jump function in an electronic dictionary is simple, it is unclear whether students understand how useful it can be. In particular, this skill is helpful for understanding definitions, which is one of the students’ high motivation skills, as previously mentioned above. This indicates that further research investigating why students and teachers rank this skill so low could be of value.
5. Summary of the Study and Educational Implications

This paper on motivation has several educational implications, as mentioned previously. First, students are more motivated to learn individual dictionary skills than they are to study dictionary skills in general, suggesting that teachers should take care to present skills concretely in the classroom.

Second, this study reveals eight skills for which students show a high motivation, as seen in Table 3. These skills are likely to be good starting points for dictionary instruction. In particular, the three skills in Table 7 show both high student motivation and teachers’ recognition of this need. These skills could be ideal starting point for future lessons. The study also shows ways to maintain students’ motivation to learn other skills. In particular, the low motivation skills depicted in Table 8 can be taught as a component of related high motivation skills, which could lead to better student engagement.

One additional point is worth mentioning. Three high motivation skills, those ranked by students as fourth through sixth, are not highly regarded by teachers. Although these skills do not meet the definitions used in this study, teachers ranked each of them below 17 in their survey. In particular, based on comments by the teachers who participated in the survey, we can see that some teachers do not seem to think that understanding Japanese definitions might be difficult for students. However, not only was this ranked among the students’ high motivation skills, the author has observed difficulty understanding definitions written in students’ first language as a common problem. This suggests that more research is needed to determine what students find difficult about dictionary use, in order to provide them with effective dictionary instruction.

Fourth, in order to provide effective instruction, it is obviously important that students be taught useful skills accurately. However, the teachers’ responses indicate that there is little time in class for dictionary instruction. In fact, the survey results showed that 8 teachers of 19 actually have conducted dictionary instruction at university level, with 6 of them offering less than one hour lessons. This reality along with the generally very high needs perceived by teachers suggests that dictionary instruction needs to be delivered as efficiently as possible at university level. Students’ motivation is very important for efficient instruction as motivated students concentrate better and learn more quickly than less motivated students.

Finally, given the very limited classroom time available for dictionary instruction, it is crucial for students to be able to use dictionaries autonomously. This skill is the ideal outcome of dictionary instruction. As Ushioda (1996 cited in Ushioda, 2011) demonstrated, high motivation is closely related to autonomous learning; thus, this is my hope that this study can contribute to providing effective and efficient dictionary instruction in the time available.
6. Limitations of the Study

The quantitative research conducted for this study is inherently limited, as there were no follow-up interviews conducted (Brown and Harris 2010) and an insufficient number of multi-item scales (Dornyei and Taguchi 2010). Second, the 30 Japanese-English dictionary skills used in this study have not been clearly defined, as other sets of skills are possible and may yield different results. Third, the educational suggestions presented in this paper assume that high motivation skills could raise students’ overall motivation to learn the subskills for which they have shown lower levels of motivation. However, direct research on this would be valuable.

References

Harris and Brown (2010). Mixing interview and questionnaire methods: Practical problems in aligning data.
MEXT. (2008a). 「中学学校学習指導要領」

MEXT. (2008b). 「高等学校学習指導要領」

MEXT. (2009)「高等学校学習指導要領解説」


Appendix: Descriptive statistics of necessity of dictionary instructions as perceived by teachers (n=19)

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>median</th>
<th>mode</th>
<th>min.</th>
<th>max.</th>
<th>combined % of 5 &amp; 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Using general dictionary conventions</td>
<td>6</td>
<td>6&amp;5</td>
<td>6</td>
<td>4</td>
<td>84.2</td>
</tr>
<tr>
<td>2</td>
<td>Understanding conventions for</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>78.9</td>
</tr>
<tr>
<td>3</td>
<td>Consulting multiple-word items</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>73.7</td>
</tr>
<tr>
<td>3</td>
<td>Using sample sentences</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>73.7</td>
</tr>
<tr>
<td>5</td>
<td>Understanding the appropriate forms of the headwords</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>68.4</td>
</tr>
<tr>
<td>5</td>
<td>Identifying the parts of speech of the target word</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>68.4</td>
</tr>
<tr>
<td>5</td>
<td>Inferring from context</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>68.4</td>
</tr>
<tr>
<td>8</td>
<td>Using conventions for sentence structure</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>63.2</td>
</tr>
<tr>
<td>9</td>
<td>Understanding the microstructure of an entry</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>57.9</td>
</tr>
<tr>
<td>9</td>
<td>Using information on proper inflection</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>57.9</td>
</tr>
<tr>
<td>9</td>
<td>Using pronunciation symbols</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>57.9</td>
</tr>
<tr>
<td>12</td>
<td>Using information on collocation</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>52.6</td>
</tr>
<tr>
<td>12</td>
<td>Confirming the definition within the appropriate context</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>52.6</td>
</tr>
<tr>
<td>12</td>
<td>Using information on sentence structure</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>52.6</td>
</tr>
<tr>
<td>15</td>
<td>Deciding the target word to be consulted</td>
<td>5</td>
<td>6&amp;4</td>
<td>6</td>
<td>2</td>
<td>47.4</td>
</tr>
<tr>
<td>15</td>
<td>Selecting dictionaries that match consultation needs</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>47.4</td>
</tr>
<tr>
<td>15</td>
<td>Distinguishing between</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>47.4</td>
</tr>
<tr>
<td>Activity</td>
<td>15</td>
<td>19</td>
<td>19</td>
<td>22</td>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Understanding the Japanese definition</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>47.4</td>
<td></td>
</tr>
<tr>
<td>Finding correct spellings</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>42.1</td>
<td></td>
</tr>
<tr>
<td>Selecting the target word from</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>42.1</td>
<td></td>
</tr>
<tr>
<td>Understanding the macrostructure of</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>42.1</td>
<td></td>
</tr>
<tr>
<td>Locating derivative forms</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>36.8</td>
<td></td>
</tr>
<tr>
<td>Distinguishing among homographs</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>31.6</td>
<td></td>
</tr>
<tr>
<td>Using reference information in printed dictionaries</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>31.6</td>
<td></td>
</tr>
<tr>
<td>Using the jump function</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>26.3</td>
<td></td>
</tr>
<tr>
<td>Using wild cards</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>26.3</td>
<td></td>
</tr>
<tr>
<td>Using information on limited usage</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>26.3</td>
<td></td>
</tr>
<tr>
<td>Using information on limited usage</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>26.3</td>
<td></td>
</tr>
<tr>
<td>Understanding alphabetical order and keyboard layout</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>26.3</td>
<td></td>
</tr>
<tr>
<td>Using information on etymology</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>21.1</td>
<td></td>
</tr>
</tbody>
</table>
Cultural Dictionary: An Instructional Enhancer to Better Reading Comprehension in Literature Classes

Rosario B. Dizon
Mindanao State University-Iligan
rbd_fil@yahoo.com

Alia D. Ramber
Mindanao State University-Iligan
ramber.alia@gmail.com

Manna Hadassah B. Estrada
Dela Salle University Complex, Binan
manna.estrada@dlsu.edu.ph

Melba B. Ijan
Mindanao State University-Iligan
melba.ijan@g.msuiit.edu.ph

Kim Minoza
Mindanao State University-Iligan
kimimoza@g.msuiit.edu.ph

Giselle O. Dangdang
Mindanao State University-Iligan
giselle.dangdang@g.msuiit.edu.ph

Abstract

The transformation of the educational system in the Philippines from 10 years basic curriculum to k+12 program paved the way for teachers to reinvent instructional materials toward more creative and experiential learning using the mother tongue. With this in mind, indigenous symbols, linguistic labels, signals, cultural motifs and expressions from 50 indigenous folktales of 12 Mindanao tribes that underwent validation by the panel of experts were gathered to codify a cultural dictionary. Two sections of high school students were then allowed to use the cultural dictionary before reading the folktales while another two sections of students were required to read the folktales without exposure to cultural dictionary. It was an experiment utilizing the Solomon Four-Group Design. Data generated from the experiment underwent statistical analysis using Statistical Package for Social Sciences (SPSS). Results revealed that there is a significant difference in post-test scores between subject participants who were exposed to Cultural Dictionary orientation and those participants who were not exposed. This means that cognitive-linguistic ability of secondary school students are possibly enhanced by motivational drive through tribal feedback brought by the cultural dictionary so that cultural codes (verbal and non-verbal) embedded in the narratives were understood quickly. Based on the findings, the study concludes that Cultural Dictionary orientation contributes to better reading comprehension of students in Literature classes in secondary education.

Keywords: transformation, curriculum, cultural motifs
1. Culture, Language and Communication

Communication is a very sensitive process. It is always dynamic and adoptive with its varieties of sounds, words, and cultural codes and symbols no matter how brief and simple the message is. It is especially true whenever and wherever students and teachers interact, there is communication.

In the teaching of literature, the best springboard for transmitting and transferring knowledge is a piece of life that is influenced by personal experiences in cultural life. Simple cultural words, anecdotes and stories can unlock literary codes and symbols, thus, perceptions are shaped and behavior are molded. Indeed, through cultural windows, the students develop “constructs”. A student learns and develops his views and concepts that are built which incorporates and reflects the most important attitudes and values woven from one’s culture. The process of communication adjustment is therefore built because undoubtedly, as one, speaks and communicates, he/she bears the imprints of his/her culture.

According to Hall (1983) “culture is communication and communication is culture” to which Birdwhistell (1990) poses a slightly different position saying that “culture focuses on structure as communication on process”. Harris (1998) contends that culture comes down to “customs” of peoples “way of life”. Geertz (1966) views culture as a “symbolic-system-school of thought”. Goodenough (1983) in his argument about culture says, “it consists of the standards for deciding what is... for deciding what can be... for deciding what one feels about... for deciding what to do about it.

In other words, culture in every sense, is not only collection of symbols put together by a person but a “sensitive system of knowledge” (Hall: 1983). It is shaped and constrained by the way human brain comprehends, organizes and processes data of information and creates internal models of reality. (Rico-Weed (1995) says that culture is the filter that reveals how an individual thinks, feels and views other individuals, himself and the world.

Similarly, Whorf makes the point that the linguistic system of each language is not merely reproducing instrument for voicing ideas but rather in itself the shaper of ideas. There is a remarkable correspondence between the thinking of Bhartrhari and the Sapir-Whorf hypothesis. The essential truth embedded in the relationship of language, culture and communication is also found in Levi-Strauss and Piaget (1969).

2. Culture, Language and Literature

In Literature, language and culture are so intertwined, language in literature transmit a collective spirit of a group. De Vito (1971) defines language as “a potentially self-reflective, structured system of symbols which codes objects, events and relations in the world.” Language in Literature serves as “stand-ins” for actual references. In the study of literature, language becomes a primary link to culture. In this argument, the “language becomes the message. It is because words as culture symbols emit cultural meanings that may be similar or may be different to the meanings of other culture.
Literature is a communication code that is closest to culture. In literature, especially oral literature, the reader can find cues and make distinction between the nature of language and the use of it. Words can be analyzed as cultural codes in a given time and place in relation to other system of human behavior. Dialogues are also analyzed in this way. Words represent something. Sometimes, it may also compose plurality of signs by a group of persons living in a particular place and have developed their own culture.

Through literature, the students learn how a particular people lived their lives in a certain time and place, tried to institutionalize their codes and symbols and succeeded in defining their priorities and life patterns. The author jot down what people say and how they saw it in actual setting. Literature communicates the experiences of folks, their mind set and prejudices including their faith as well. The culture is discovered through language signals and “words are language signals”, (Hymes: 1973).

Literature not merely modifies but also interprets. It is indeed, an interpretative art using language, creative cultural language. The author with his beliefs shapes his material into a lifelike situation. Details of characters, plot and setting are so chosen to make the story credible. The writer too becomes an interpreter of his story using his own culture.

3. On Enhancing Reading Comprehension

For instruction purposes in multi-cultural classes, maintaining a pleasant discussion requires the “ability to comprehend the relationship of the medium to the first language” (Gudykanst: 1993). Before a teacher, on the other hand, can manipulate and control events and behavior, he has to tailor understanding, cater teaching styles to suit the gender, status, age, mental ability and cultural background that might be similar or very different to the literary genre that is being studied. If only to be more clear and transparent, a teacher orients his/her class to visual instructional material like that of a cultural dictionary that carries the culture blueprint of an individual.

According to Lanigan, (1986) “Printed word should be in the student’s speaking, meaning and vocabulary.” The printed words are but merely symbols that serve as cues to the reader who must organize an understanding of what they meant. The ease with which the reader can associate meanings depends largely upon his background of experiences. The student reads using his experiences and language skill, interrelating them to device meaning from the printed symbols.

To a large degree, the most important thing to consider is that, “reading as a tool for learning will be ineffective unless it is accompanied by thinking.” The ability to develop, expand and refine comprehension ability involves gradual acquisition of skills through reading. In this foundation, reading defined as the ‘recognition of printed or written symbols that serve as stimuli to the recall of meaning built up through the reader’s past experience (Harris: 1990). New meanings are derived through the manipulation of concepts already in the readers’ possession. Organization and clarification are guided by purpose on the reader’s mind, which is influenced basically
by his culture and experience. The mind cannot relate, interpret, evaluate nor reflect correct meanings unless it is lighted by the said two interrelated elements.

According to Gillet (1990), the rate of development in the reader’s auditory and visual discrimination is also grounded in his/her experiences. This means that if the reader’s experience has been extensive and related, reading comprehension is quite easier but if the learner’s experiences are narrow and quite dissimilar, comprehension becomes difficult.

Under favorable circumstances, when the reader’s experiences are rich, he/she will not only be ready to read but eager to read and learn because reading is a complex learning. The acquisition of an adequate background of experiences and normal physical development with satisfactory classroom adjustment could be the positive factors for reading readiness. (Jewel et al., 1986).

An effective reading instruction/curriculum is geared more than to develop the abilities and skills of visual recognition and identifying words, grouping words into thought units, noting details and following directions. According to Temple (1990), the overall goal of reading instruction is to help each learner become as able and diversified a reader as his capabilities, the available facilities and instructional program permit. To achieve this, certain factors must be considered and satisfied. These are 1) Basic understanding of words; 2) Progressing maturity in reading habits and attitudes; 3) Independence in reading; 4) Ability in the use of basic study skills; 5) Increasing maturity in essential comprehension skills like the ability to evaluate, interpret and appreciate.

4. Cultural Dictionary as Instructional Support Material

When we consider standard learning enhancement, we expect that our students will show more than satisfactory learning outcomes. Although this is not easy, a teacher by heart can use his/her intuition to augment audio/visual instructional aides to bring in ideal learning. Student have different learning styles. A passionate teacher always welcome challenges to produce greater learning impacts and outcomes despite greater diversity in student populations. Common materials in the classroom tend to be limited in scope. Textbooks, manuals and even workbooks/study kits have limiting skills. Specific multimedia, maps or the globe, videos, the Internet can be better used for reinforcing learning. These tools however, though they augment instruction, cannot sustain the delivering knowledge that can be carried by a dictionary, particularly, a dictionary that is cultural. The need of a dictionary is largely been ignored nowadays when so many gadgets fluctuate in the home and school. This is very true even studies conducted from the 1980s onwards revealed the gap between the typical dictionary structure and the reference skills possessed by the average dictionary user.

Dictionaries are essential tools for effective learning. Taking advantage of the information compilers put into dictionaries requires (a) an understanding of their value as well as limitations (dictionary awareness); (b) specific skills enabling the user to find the information searched for and, once found, extensive knowledge of the
lexicographic conventions used in order to interpret it properly and integrate it into the context which prompted the search (reference skills). That is why a classroom orientation is necessary so that students can develop interest in consulting a dictionary not only in searching for deep meaning but even in cultural equivalents as well. Later on, the natural interest is developed as a habit and a lifestyle.

Learners grow at different rates. He/she learns to read with his/her eyes, ears, energy, as well as cultural experiences. Any differences found within readers in any of these elements affects their reading skills.

Concerned teachers can quickly recognize that individual differences in reading abilities exist among students. Just because a particular child excels in reading and understanding science, it doesn’t necessarily follow that the same child will excel at reading poetry or folktale. A certain student may grasp the overall meaning of a sentence, and able to recognize unfamiliar words and yet his knowledge of figures of speech could remain very limited. Another child may be able to recognize parts of speech but unable to group them into thought units or what and how figures of speech came about. The extent of these variations, according to Lanigan (19), the ways of diagnosing them directly and adjusting to individual differences are frequently not fully understood.

5. Theoretical Framework

The present study is grounded on the theoretical foundation on the relationship of language, culture and communication:

Ferdinand de Saussure (1925), one of the formulators of STRUCTURALISM viewed language as a system of signs that define one another through their relations rather than through meanings. He distinguished sharply between the system of language (langue and the actual use (parole).

The Sapir-Whorf hypothesis states that each language shapes the reality of its speakers and that no two cultures share the same understanding of the world.

According to Cecilia Lopez, (1987) “Philippine languages, without doubt, must also have borrowed internally from one another”. In Mcfarlan’s linguistic grouping of Philippines languages, Tagalog, which is the basis of Filipino, belongs to the Meso-Philippine languages together with Tausug. Maranao, Maguindanao, Manobo, B’laan and Tiruray belong to the Southern Philippine Languages.

In teaching Literature whether it be in English and/or Filipino as stipulated in Department of Educ and CHED order no. 59 (s.1996), cultural barriers are needed to be removed by clearing ideas. With this approach, Filipino students of different cultures could have feeling of belongingness.

Codes are words and signs of the culture of an individual. The culture, which is the environment, controls the individual’s reactions and behavior. Verbal and non-verbal reactions are also molded by the culture. These elements are present in any literary piece. The author puts in much representations of culture for the teacher and the students to unlock and understand objectivity. When they are identified and
understood, literary devices used for aesthetic beauty will also readily be discovered.

In orienting the students about the use of a cultural dictionary, cultural awareness is introduced. Students are led unconsciously to adjust themselves and speak out their thoughts, in the process, they are thinking aloud their culture. Unfamiliar words that may be present in story a tale or narrative will be mispronounced but after detection, pronunciation is corrected and the obstacle in comprehension is removed. The search for meaning is innate but learning provides both stability and familiarity as well as satisfying curiosity. The brain is designed to perceive and generate patterns. The ideal teaching process presents information allowing brain to extract patterns and create meanings. Although emotions are crucial to learning, instruction support should be marked by mutual respect and acceptance. Learning always involves rich environmental stimuli that accommodates presentation of cultural recipe like cultural dictionary that can enhance comprehension.

![Figure 1 Conceptual Framework](image)

The experimental phase of this study determines better the experience of cultural dictionary orientation and instruction to really make the literature class worthwhile between the student and the teacher.

With cultural dictionary orientation focusing on the cultural conventions present in a literary piece folktale for that matter, the student experiences cultural awareness. The cultural awareness through the use of cultural dictionary will serve as motivation drive; an unlocking device to prepare the mind processes for better appreciation and understanding of everything in the story.

When culture influences reading, the brain works to identify printed symbols and associate them with correct meanings and messages in order to get the ideas being conveyed by the author. Comprehension comes either fast or very slow depending on the stock of information in the brain.

### 6. Methodology

This study utilized the Two Non- Equivalent Groups Design (One Experimental- One Control Group Design) to carry out pilot experiment and another involving four
collegiate groups from Iligan Institute of Technology of Mindanao State University using the Solomon Four-Group Design. The pilot experiment was intended to discover flaws in the actual experiment.

One Experimental-One Control Group Design

<table>
<thead>
<tr>
<th></th>
<th>Pre-Test</th>
<th>Treatment</th>
<th>Post Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental Group</td>
<td>O₁</td>
<td>X</td>
<td>O₂</td>
</tr>
<tr>
<td>Control Group</td>
<td>O₃</td>
<td>(X)</td>
<td>O₄</td>
</tr>
</tbody>
</table>

Legend:
X = Experimental Stimulus/Treatment (Cultural Dictionary Orientation)
(X) = Folktale reading activity only
O₁, O₂, O₃, O₄ = Measures of Reading Comprehension

SOLOMON FOUR-GROUP DESIGN (Creswell: 1994:134)

<table>
<thead>
<tr>
<th></th>
<th>Pre - Test</th>
<th>Treatment</th>
<th>Post Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1 R</td>
<td>O₁</td>
<td>X</td>
<td>O₂</td>
</tr>
<tr>
<td>Group 2 R</td>
<td>O₃</td>
<td>C</td>
<td>O₄</td>
</tr>
<tr>
<td>Group 3 R</td>
<td>X</td>
<td></td>
<td>O₅</td>
</tr>
<tr>
<td>Group 4 R</td>
<td>C</td>
<td></td>
<td>O₆</td>
</tr>
</tbody>
</table>

Legend:
R = random assignment of subject groups
X = exposure of a group to an experimental (treatment). (Cultural Dictionary Orientation preceding reading Literary piece)
C = exposure of a group to Control Condition (reading Literary piece only minus Cultural Dictionary Orientation)

a) Research Samples

There were two sets of samples chosen to participate in the study. The first set (who participated in the pilot study) consisted of 2 intact classes of Filipino 5 Literature, Department of Filipino and Other Languages, MSU-IIT, Iligan City during the first semester SY 2011-2012. Twenty-eight (28) students constituted the Experimental group and twenty seven (27) students comprised the control group.

This was actually a preliminary try-out (rehearsal) or a dry run of the experiment whose objective was to discover/anticipate problems, possible weaknesses in the design and detects contamination effects in the administration of the actual/real experiment. A second objective was to generate data for the validity and reliability testing of the criterion post test.

The second sample (which participated in the actual or real experiment) was randomly selected from the poll of two (252) hundred fifty-two students enrolled in Filipino 5 – Lecture classes during the second semester, SY 2011-2012 at CASS.
MSU-IIT, Iligan City. They were randomly assigned to 2 experimental groups and 2 control groups with 30 students each. Only twenty (20) were males and one hundred (100) were females whose ages ranged from sixteen to seventeen years old. All scored 85 PR or higher in the MSU Systems Admission and Scholarship Exam SASE and with a final grade of 2.0 or higher in Filipino I.

6.2 The Pre-Post Comprehension Test (Criterion Test): Content Validity and Reliability Analysis

The Pre-Post Test referred here was the General Comprehension Test (the criterion test which was first content validated/evaluated by three professors from Filipino Department of the College of Arts and Social Sciences, MSU-IIT and again, re-evaluated by another professor from MSU; a Ph.D. Language Studies major and another professor with a Ph. D. in Linguistics.

It was subjected also to students’ evaluation/validation using twelve (12) College of Science and Mathematics (CSM) students for objective comprehensibility and clarity of each item. The second student tryout of this test instrument was done in one major class of Filipino 22 (Sanaysay). There was no revision made because they all understood what each item was asking for.

This was the instrument used as Pre-Test and consequently Post Test for the pilot study preceding actual experiment. As pre-test, it was administered to 55 students (2 intact classes) of Filipino 5 offered last first semester of SY 2011-2012 (EG = 28, CG = 27). As post-test it was also administered to both the Experimental Group (EG) and the Control Group (CG) after the folktale reading activity.

After the raw scores of the students have been obtained and tallied, this test instrument was subjected to reliability and validity analysis using Statistical Package for Social Sciences (SPSS) program.

The result of these statistical computations (see Appendix B) showed that numbers 2, 5, 13, 20, 29, 39, 44 and 45 must be modified or revised because it was discovered that the students were almost certain of the correct answers to be chosen from the choices given. On the other hand, items 5 and 20 appear to be rather too difficult (based on the Pre-test data in the Pilot Experiment).

<table>
<thead>
<tr>
<th>Item</th>
<th>R</th>
<th>sig.</th>
<th>t-value</th>
<th>df</th>
<th>sig</th>
<th>Wrong Response</th>
<th>Correct Response</th>
<th>Total Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.288</td>
<td>.033</td>
<td>2.191</td>
<td>53</td>
<td>.033</td>
<td>19</td>
<td>36</td>
<td>55</td>
</tr>
<tr>
<td>5</td>
<td>.362</td>
<td>.007</td>
<td>2.830</td>
<td>53</td>
<td>.007</td>
<td>47</td>
<td>8</td>
<td>55</td>
</tr>
<tr>
<td>13</td>
<td>.499</td>
<td>.000</td>
<td>4.191</td>
<td>53</td>
<td>.000</td>
<td>19</td>
<td>36</td>
<td>55</td>
</tr>
<tr>
<td>20</td>
<td>.305</td>
<td>.023</td>
<td>2.333</td>
<td>53</td>
<td>.023</td>
<td>50</td>
<td>5</td>
<td>55</td>
</tr>
<tr>
<td>29</td>
<td>.306</td>
<td>.023</td>
<td>2.336</td>
<td>53</td>
<td>.023</td>
<td>7</td>
<td>48</td>
<td>55</td>
</tr>
<tr>
<td>39</td>
<td>.321</td>
<td>.017</td>
<td>2.467</td>
<td>53</td>
<td>.017</td>
<td>42</td>
<td>13</td>
<td>55</td>
</tr>
</tbody>
</table>
Table 1 shows the partial result of the 50-item comprehension test used in the Pilot Experiment in that it was subjected to validity analysis using Pearson correlation technique and t-test. The result of the SPSS computation on validity and reliability analysis shows significant correlation on items 2, 5, 13, 29, 39, 44 and 45 with total scores. The insignificant correlation indices imply that each item contributed to a high or low total score meaning a non-correct response is correlated with a low total score and correct response is correlated with high scores.

Further, the t-test was used to find out if those who got them wrongly has significant total scores lower that those who answered the item correctly. The statistical results were consistent.

Note that since this was a pre-test and students are not supposed to be familiar yet with the correct answers, therefore the above items have to be modified. Thus, in relation to the above findings, these numbers were modified/revised and the modified version of the pre-test was used with the Experimental Group in Actual Experiment.

Based on the second set of data generated form the pre-test of Actual Experiment, reliability and validity indices were again computed.

Table 2 Paired Samples T-Test Based on Pre-Post Test Scores of EG1 and CG1

<table>
<thead>
<tr>
<th>Group</th>
<th>Pre-Test Mean</th>
<th>PostTest Mean</th>
<th>Mean Difference</th>
<th>T - Value</th>
<th>Significance Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>EG1</td>
<td>20.63</td>
<td>30.67</td>
<td>10.03</td>
<td>13.754</td>
<td>.000</td>
</tr>
<tr>
<td>CG1</td>
<td>19.53</td>
<td>19.57</td>
<td>.04</td>
<td>.052</td>
<td>.959</td>
</tr>
</tbody>
</table>

Considering the mean difference of 10.03 between pre-test and post-test performance of EG1, the T-value of 13.754 and probability value of .000, Table 2 shows how post-test scores are significantly different from the pre-test scores of EG1. In other words, it can be said that the post-test scores are significantly greater than the pre-test scores of this group.

CG1 data, on the other hand reveals different interpretation. With its mean difference of only .04, T-value of .054 and probability value of .959 which is >.05 or greater than the observed level of significance, it is therefore very hard to conclude that there is a significant difference between pre-test scores and post-test of this group. The results proved otherwise, meaning the very small difference may just be due to chance factor.

7. Results and Discussion

Learning is a process integrating experience and concepts, observations and actions, “(Ora: 1995:2) the impulse of experience serves as moving force and direction to ideas. Delaying or suspending quick action/reaction is an instinct essential for observation.
and decision building. The mind has to weigh meanings, associate and integrate sounds, words and concepts to experience before decision building takes place. In the absence of association and integration, “blind impulse” is resorted to. The integration process allows a learner to mature because “… the learner is directly in touch with the realities being studied rather than thinking about possibilities facilitates effective learning. This flow of developmental learning based on Dewey’s model of experiential learning was one of the conceptual pillars of this experiment.

On Actual Experimental Group, (EG)

As the subject participants of EG read the folktales, they could spontaneously associate to their common experiences the sounds, lexicons and semantics of the cultural terms they were encountering along. The cultural dictionary orientation builds ideas and concepts towards better understanding of each folktale. Although there maybe multiple cultural experiences, the fact that the teachers deal with Philippine stock of languages and therefore from Malayo-Polynesian family, understanding of meanings of cultural codes may have more similarities than differences.

Sounds and Lexical symbols may seem different but the cultural orientation cleared doubts for comparisons and integration to their personal encounters. Meanings then became clear and precise so the learning process was not hampered. Observations were assimilated into the mind. Theories from implications were deduced. These then became guideposts and directions to faster comprehension.

Kurt Lewin’s (1985) dialectic tension and conflict between immediate and concrete experiences and abstracts are ironed out by Cultural Dictionary orientation making the students experience the “here-and-now phenomena” validating abstract evaluation of situations they found in the story. The questions on basic figures of speech were a test of students’ awareness on their capability to associate and link general literary devices embedded in the plot of the stories. The premise was that if the students understood the cultural context of the literary codes, they could print out the beauty and objective comparisons to details from internal to external literary elements of each text. To come up with this, the researchers considered the different dimensions of questions: literal, association, inferential, evaluative and implications. In doing this, she had in mind the accommodation of concepts (schemas) to experience in the world and assimilation of events and experiences from the world into the existing concepts of the Cognitive Developmental Tradition by Jean Piaget (1969).

As he postulated, learning occurs in the first level through seeing and feeling and handling through action. In the second level, called the representation stage, there is reflective orientation. At this point the learner begins to internalize actions and images. In the third stage, is the beginning of symbolic developmental stage. Learning in this stage is governed by the logic of classes and relations- the learner learns to select concepts based on her experiences. And then she/he, he comes up with the fourth stage, called the formal operations. This is a mental movement in learning and symbolic processes to concrete operations based on the representational logic he/she previously constructs. The learner develops possible implications and
confirms this through awareness. This learning cycle from accommodation to divergence, from divergence to assimilation ends up to integration and convergence.

In the Pre-Test, the EG1 group already saw, smell, feel and imagine the cultural worlds through the cultural dictionary conventions. But they could not associate yet the images and emotions. In reading Literature, the author’s selective use of native words found in a codified cultural dictionary allowed the students to enter the “gates” of the cultural world so that they can be brought inside and breathe the air of the seven tribal worlds.

Cultural Dictionary orientation became the cultural guide; the “awareness instrument” for them to develop reflective orientation. The subject participants were able to associate therefore the images and actions of the characters in the tales with the codes that they encountered in the pre-test. In the third level, they were able to select from their own experiences in their culture the correct association (similarities and differences) from what they read in the piece of folk literature. Then they found it easier to draw implications and justify relationships from their own sounds and cultural realities. Confirmations of word meanings became natural because iconic symbols were associated with real experiences.

Awareness was really the key to effect optimum comprehension. Without this, no matter how many times they may read any piece of literature including folktales, the unlocking key to comprehension would never be present.

The statistical result shows significant increase of reading comprehension test scores from pre-test to post-test of the Experimental group. They were not paralleled by the Control subjects. By the reinforcing findings that even with or without pre-test, the stagnant low post-test scores of the Control group 1 and 2 only suggest the idea that the Cultural Dictionary Orientation helped a lot to situate students’ senses and mind to familiar ground as a bridge for new association and concept, inferences and integration of knowledge such as literary devices, value judgements and others. This is a simple semblance of “cognitive constructs”. Lastly, the findings suggests that cultural dictionary orientation directly affects comprehension in reading literary pieces.

8. Conclusion

This means that cognitive-linguistic ability of secondary school students are possibly enhanced by motivational drive through tribal feedback brought by the cultural dictionary orientation so that cultural codes (verbal and non-verbal) embedded in the narratives were understood quickly. Based on the findings, the study concludes that orientation to Cultural Dictionary contributes to better reading comprehension of students in Literature classes in secondary education. Further, values that can be developed to students through exposure to the cultural dictionary are being non-judgmental, emphatic and trustful.
References

Abstract

From a cross-linguistic perspective, idioms may have various equivalents, which differ in interlingual similarity, starting from full equivalents occurring when source language idioms and target language idioms have identical lexical, formal and stylistic features, followed by partial and parallel counterparts, finishing with the zero equivalents – the source language fixed idiomatic expressions with no idiomatic equivalents in the target language. In terms of bilingual lexicographic description, it can be assumed that the most serious problems will be caused by zero equivalents, in particular by the semantically and connotatively complex idioms of culture-bound character. The paper aims to give an insight into the bilingual lexicographic description of idiomatic expressions which serve as labels for rich meanings and multiple associations in native language users. The culture-boundness of the idioms, their semantics and connotative potential demand special treatment from the lexicographer. The bilingual lexicographic description should present the meaning and connotative potential of the idiom in such a way that irrespective of non-equivalent character of a given idiomatic expression, a prospective dictionary user would be able to decode the idiom and to use it in various contexts. In the paper, one develops a model of bilingual description of such idioms. The proposed model is an extended one, in which an emphasis is placed on the inclusion of additional linguo-cultural information allowing for providing a true picture of the unit and its status in the source language. In the model, indirect defining, e.g. stylistic labels, is also of great importance.

Keywords: culture-boundness, idiom, semantics, cultural connotations, bilingual lexicography

1. Culture-Boundness and Idioms

In studies on phraseology, various approaches are adopted by scholars doing research within different methodological frameworks. In fact, many terms are used in different meanings by respective scholars, even the key term – idiom. It should be admitted that there is no general agreement on the definition of this term, yet certain features of idiomatic expressions are repeated in many linguists’ works, for instance, polylexicality and fixedness. Generally speaking, the term idiom can be used in a narrow sense, which is limited to word-like expressions of figurative character, and in
a broad one, which comprises reproducible language units of various kind, both word-like and sentence-like multiword expressions, either of figurative or non-figurative character (Fiedler 2007: 28). In the broader approach, the word *idiom* is used as an umbrella term to name collocations, idioms proper, proverbial expressions, etc.\(^1\). It can be added that the term *phraseeme* is also used in a similar way.

Although the interest in cultural aspects of phraseology varied in earlier research studies, over the last two decades the cultural component has been included in many works, which to a significant extent is related to the adoption of cognitive approaches, focusing on the “relevant knowledge structures which underlie phrases” (cf. Piirainen 2008: 207). The relation between fixed expressions and culture has been analyzed from various angles, for instance, from an ethnolinguistic perspective\(^2\). As for the culture-oriented linguistic terminology, one observes inconsistency in the use of the terms in phraseology (Sabban 2008). Another problem is the applications of terms like *culture* or *culture-specific* without defining their meanings, which is needed for their operationalization (cf. Sabban 2004, 2007, 2008; Dobrovol’skij 1999; Piirainen 2008).

Among the terms used most often in phraseological studies, the following ones should be mentioned: *culturally marked, culture-specific* and *culture-bound* (cf. Sabban 2008: 230). As stated by Sabban (ibidem: 231), the adjective *culture-bound* – although not so widespread – is preferable over the other two. First of all, its use implies that one has to determine in which way a given language item is *culturally bound*. Secondly, the term indicates that cultural and linguistic aspects can be analyzed on two planes, which are undeniably related. Lastly, the application of the term in question creates the possibility of including the study on both similarities and differences in phraseological stocks of two or more languages, whereas the adjectives *culture-specific* and *culturally-marked* put an emphasis on the discrepancies.

Furthermore, the term *cultureme* – the usage of which does not have to automatically involve ethnospecificity – is used in phraseological studies (cf. Pamies 2007). The cultureme is defined as “extralinguistic cultural symbols, which behave like metaphorical models, motivating figurative expressions in language (lexical or phraseological)” (Pamies 2017: 101). It is worth emphasizing that this concept is useful from the perspective of bilingual lexicography (cf. Szerszunowicz 2014a, Pamies 2017).

Culture-boundness of idioms may be analyzed from various perspectives, for instance, the focal issue can be the semantics of the idioms and their connotations. It can be assumed that there are certain idiomatic expressions which convey culture-bound meanings and evoke cultural connotations. From a cross-linguistic perspective, there will be various relations between the source language units and their target language counterparts: they can be very similar, partly identical or not shared by two languages. The last problem, i.e. zero equivalence in the in semantics and connotations of the source language unit and its target language counterpart will be

---

\(^1\) Generally speaking, phraseology can be understood in a broad sense and in a narrow one (cf. Colson 2008).

\(^2\) For an account of the state of the art see Sabban (2007) and Piirainen (2008).
discussed on the example of selected Polish units and their English equivalents.

2. Cross-Linguistic Equivalents of Idioms

While discussing cross-linguistic equivalents of idioms, the distinction should be made between translational equivalents and systematic ones (Dobrovol’skij 2011). The former ones are the representations of idioms in the translation, chosen from several candidates, both phraseological and non-phraseological, after a multiaspectual analysis of the use of a given unit in the original.

Systematic equivalents are the target language counterparts of the source language idioms at the system level. Establishing systematic equivalents of idioms is of great importance from the point of view of bilingual lexicography (Vidović Bolt and Szerszunowicz 2017; Szerszunowicz 2014a, 2017). Basing on various division of systematic equivalents, Dobrovol’skij (2011) proposes a typology comprising four kinds of cross-linguistic equivalents labeled as full, partial, parallel and zero.

The first term is used to refer to the source language idioms and their equivalents sharing lexical and formal properties, having the same meaning and stylistic value. In the case of partial equivalents, subtle changes may occur, for instance, the word order can be slightly different. Parallel counterparts employ different imagery, still expressing the same meaning. When source language idioms do not have phraseological equivalents, then they have zero equivalents.

The last kind of equivalents mentioned comprises two kinds of units: one of them is the idiom whose meaning is not verbalized in the target language unit, but whose meaning is known to the target language users; the other occurs when the idiom verbalizes the meaning unknown to the target language users. Such units are of lacunary character and they create gaps, also called *lacunae* in the target language culture (Szerszunowicz 2017). From a cross-linguistic perspective, the most problematic ones are those which belong to the latter kind and are deeply rooted in the target culture (Szerszunowicz 2014a). Among them, it is possible to determine the expressions which express complex notions, i.e. the units conveying meanings which are too complicated to be verbalized in a simple definitions typical of an entry.

3. Culture-Bound Idioms Expressing Complex Notions in Dictionaries

The study is based on the analysis of the lexicographic entries for the unit *polskie piekło / polskie piekiełko* (lit. Polish hell / Polish little hell). The idiom was selected after a thorough linguo-cultural analysis of general dictionaries, phraseological dictionaries and other lexicographic sources including specialized dictionaries, for

---

3 In several recent works, due to the fact that full equivalence of all parameters of the target language and source language idiomatic units is rather improbable, the term quasi-equivalent is considered as more appropriate (Fiedler 2007; Szerszunowicz 2017). For instance, one can expect differences in frequency, stylistic value and/or register of the corresponding expressions.
instance, the dictionary of political jargon (Szerszunowicz 2014a). The focal issue of the search was to find Polish multiword fixed expressions which are non-equivalent from a Polish-English perspective (Szerszunowicz 2017), i.e. they do not have corresponding English idioms.

All the excerpted units fulfilling the above criteria, like *Matka Polka* [lit. Mother Pole] or *Polak potrafi* [lit. A Pole can (do)], express complex notions which are culture-bound, thus deserve special attention in bilingual lexicography. Then, one of them, *polskie piekło / polskie piekiełko*, was selected and a corpus analysis was conducted. Its results and those of a questionnaire study, in which the participants were asked to explain the meaning of the expression⁴, were used for creating a model of a bilingual entry for a semantically and connotatively complex multiword unit.

### 3.1 Monolingual Dictionaries

As for general monolingual dictionaries of the Polish language, only one of the consulted works contains the idiom *polskie piekło*, mainly *Praktyczny słownik współczesnej polszczyzny* which offers the following explanation: ‘polityczna diagnoza polskich kłótni, sporów’ [lit. a political diagnosis of Polish quarrels and disputes] (PSWP, 28: 272). Similarly, only one phraseological dictionary, *Wielki słownik frazeologiczny języka polskiego*, includes this unit, which is defined as follows: ‘zawiść, niedocenianie czyjeś wartości, obłuda cechująca życie publiczne w Polsce’ [lit. envy, underestimating somebody’s merit, falseness characteristic of the public life in Poland] (WSFJP: 199).

The unit *polskie piekło* can be found in other lexicographic sources, for instance, *Słownik wydarzeń, pojęć i legend XX wieku* (SWPiL)⁵, in which the entry is more comprehensive and provides the information on the cultural background, as presented below:

**POLSKIE PIEKIEŁKO** wyrażenie spopularyzowane przez Tadeusza Mazowieckiego (ur. 1927), polskiego publicysty i polityka, premiera rządu III Rzplitej w 1989-90, na posiedzeniu Komitetu Obywatelskiego 31 marca 1990 zdaniem: „Abyśmy tej zaczynającej się polskiej demokracji nie zamienili w polskie piekło (...) swarów, podgrzyzań i walk [lit. an expression popularized by Tadeusz Mazowiecki (b. 1927), Polish publicist and politician, the Prime Minister of the Third Republic of Poland in the period 1989-1990, at the meeting of Citizens Committee on 31 March 1990 in the following sentence: „So as not to turn the budding democracy into the Polish hell (...) of quarrels, dishonest dealings and struggles] (SWPiL: 324)

The expression is also used in a quotation included in the dictionary of Polish and foreign winged words (WSCPiO: 269). In this work, the unit occurs, but it is not

---

⁴ The study was conducted twice, in 2013 – 137 participants, in 2017 – 128, with no significant differences in the results in terms of semantics of the unit *polskie piekło / polskie piekiełko*. For more information on the first research study see Szerszunowicz (2014a: 333-336; 2017: 235-237).

⁵ All translations own.
explained in any way, which results from the micostructure adopted by the authors of the dictionary in question.

Finally, a more comprehensive description is offered by an on-line source — Wikipedia (WIKI). The unit is explained in a more detailed way by listing the components of the general meaning of the phrase based on the previous research study (Szerszunowicz 2017: 236-237), which is illustrated below.

POLSKIE PIEKIEŁKO
- synonim zazdrości, zawiści połączonej z ciągłym narzekaniem, apatią, brakiem jakiejkolwiek inicjatywy [lit. a synonym of jealousy, envy combined with constant complaining, apathy, lack of any initiative, etc.];
- prywata, działania poza prawem [lit. private interest, being beyong law],
- pojmowanie „patriotyzmu” pod własną korzyść [lit. understanding “patriotism” for one’s own benefit],
- brak zdolności działania „ponad podziałami”, wypracowania ogólnego konsensusu, kompromisu [lit. the lack of the ability to work “beyond divisions”, to work out a consensus or a compromise],
- nieumiejętność konstruktywnej dyskusji, warcholstwo, kłopotliwość [lit. the inability to have a constructive discussion, lawlessness, being quarrelsome] (WIKI).

The last of the descriptions (WIKI) comprises the biggest number of semantic components of the complex meaning expressed by the phrase polskie piekło / polskie piekiełko. From the point of view of a non-native speaker, it can be assumed that this presentation is most useful, since it enables the dictionary user to construct a mental image of what a “Polish hell” is. The entry is informative and its scope facilitates the creation of a mental basis by the prospective users. The basis is necessary for their decoding the actual meaning of the unit polskie piekło / polskie piekiełko in various contexts.

3.2 Bilingual Dictionaries

The bilingual dictionaries consulted for this study (e.g. WSAP; NSFK), which included the most important Polish-English dictionaries published within the three last decades, did not contain the idiom polskie piekiełko. In fact, the unit is question is not included in the vast majority of monolingual dictionaries, either those general, or phraseological, which – to some extent – may account for its absence in bilingual dictionaries. The second factor causing the lack of the unit is the fact that there are relatively few Polish-English phraseological dictionaries. As for the Wikipedia entry (WIKI), it does not offer an English translation of the Polish entry for polskie piekiełko.

The questionnaire study shows clearly that the vast majority of participants know the unit and are able to explain it (cf. Szerszunowicz 2014b). The position of the idiom is also corroborated by its frequency in the corpus data (cf. Szerszunowicz 2017).

---

6 The number of English-Polish phraseological dictionaries available is much bigger.
Therefore, it can be concluded that the idiom *polskie piekło / polskie piekielko* should be included in bilingual general and phraseological dictionaries, since it is an important expression from the linguo-cultural perspective. A prospective dictionary user, who is not a native speaker of Polish, should be provided with a comprehensive explanation of the meaning of the phrase.

4. A Model of Description of Culture-Bound Idioms Expressing Complex Notions in a Bilingual Dictionary

The proposed model of lexicographic description of culture-bound idioms expressing complex notions includes several elements, among which the most important is the explanation of the meaning expressed by the described unit. The definition should be comprehensive and comprise not only the meaning or meanings but also the connotations of a given phrase and additional information regarding the cultural background, contextual use of the described phrase, etc. Determining the meaning of the unit should begin with the analysis of the previous lexicographic description in the dictionaries available, then the corpus data analysis should be conducted so that the actual meaning could be described.

Furthermore, questionnaires and in-depth semi-structured interviews with native speakers of the source language can give an insight into the understanding of a particular idiom (Szerszunowicz 2014b). Such an integrated approach enables the lexicographer to take into considerations changes in the meaning of the expression as well as to analyze its stylistic value in modern texts.

In the case of culture-bound idioms expressing a complex notion, the meaning can be discussed so that the description of the complex meaning is divided into categories in which particular semantic and associative components could be placed (cf. Szerszunowicz 2014b). What is proposed here is to present the explanation not in a form of definition, but rather as an enumeration of components of the notion which is hard to comprehensively explain in a traditional form.

As for labels, they are also very important from the point of view of describing the unit in a bilingual dictionary. The labels support the explanation and provide the information on the idiom which is very useful for the dictionary users, especially if the specific character of the idiom is taken into consideration.

The examples of the idiom use constitute a source of information on the actual usage. If the meaning of the unit is described in several categories, as suggested above, the examples given should correspond with the above mentioned elements. The contexts of usage should be signaled, so that the dictionary user could develop an awareness of when and how to use the unit.

The cultural component is also of great importance, since the unit expresses a notion which is culture-bound (Cf. Szerszunowicz 2011). The additional information can comprise the reasons for the creation of the image of the Polish hell against its socio-historical background, references to the use of various texts of culture, starting with literary works and finishing with modern ones, for example, memes in which the
unit is employed.

The entry for the unit *polskie piekło / polskie piekielko* presented below illustrates how this approach allows for providing the prospective dictionary user with the information which can be transformed by the user into a conceptual foundation for creating the image of the cultural phenomenon described by the idiom in question.

**polskie piekło / polskie piekielko**

*pejorative, informal*

used as a label referring to all Polish negative characteristics, those which are elements of Poles’ autostereotype

POLISH NATIONAL CHARACTER [Polish bad qualities, Polish national character, Polish mentality]

JEALOUSY, ENVY AND INABILITY TO COOPERATE [being jealous of other people’s success, wealth etc., being envious of other people’s success, wealth, etc., being selfish, using other people, creating problems]

NEGATIVE ATTITUDE TO LIFE [being too critical of everything, hatred, bad attitude to life, aggression, complaining about everything, being quarrelsome, irritability, creating conflicts, underestimating other people’s achievements, merits, etc.]

FALSENESS [being suspicious, intolerance, false piety, pseudo-Catholicism, two-sidedness]

SELFISHNESS [aiming to have profits for oneself at all expense, using public property, etc. for one’s own benefit, nepotism]

BEING PASSIVE AND OVERAMBITIOUS [being passive, lack of will, lack of action, waiting for being given instead of doing things oneself]

PROBLEMS RELATED TO THE ACCEPTANCE OF THE LEADERSHIP AND LAWLESSNESS [unfulfilled ambitions combined with envy result in the questioning or rejecting of the leadership, inability to decide on who should be appointed the leader]

**Typical use**

everyday speech, press texts, spoken public discourse

**Examples**

Dodam jeszcze, że z drugiej strony, Polska nie daje wiele szans. Kiedy próbuje się tutaj zrobić coś pożytecznego, można natrafić na polskie piekielko. […] Tutaj dobre podejście to siedzenie z założonymi rękami i użalanie się nad sobą. Jeżeli jesteś osobą zdolną, pracowitą i chętną, aby coś zmienić, zaraz znajdzie się banda ludzi których to wku...rzy i spróbują Ci zatruć życie. (NKJP)

O dziwo, po przyznaniu nagrody nie rozplomieniło się zwykle w takich sytuacjach polskie piekielko. W licznych prasowych artykułach nikt nie ośmielił się podważać weryfikacji Akademii Noblowskiej. Przeważała narodowa duma i uczucia patriotyczne. (NKJP)
A Unia? Cóż, pewien jestem, że wstąpić do niej musimy, tylko najbardziej obawiam się samych Polaków. Czy my będziemy potrafieli wykorzystać możliwości, czy nie zwycięży polskie piekiełko. (NKJP)

W Kulturzentrum mało tego, że dostali bzika na punkcie zmian, to jeszcze polskie piekiełko, wstręt człowieka brał, gdy na to wszystko patrzył. Kłótnie, podchody, sytuacja jak na tonącym okręcie, każdy tylko rozgląda się, jak tu drugiemu podłożyć świnię, a siebie uratować. (NKJP)

To najistotniejsza część polskiego piekiełka, gdzie każdy uważa, że ma prawo i kompetencje gadać o wszystkim, a jak ktoś próbuje wyjść poza gadanie, wtedy pojawia się armia oceniaczy i ulepszaczy. Jak zauważył Witkacy – za tym wszystkim kryje się zwykle kompleks niższości, tyle tylko, że terapia, którą obrałeś, prowadzi donikąd, również pacjent jakby nie ten. (NKJP)

Obiecywałem sobie, że z pokorą przyjmę decyzję jury debatującego nad projektami siedziby Muzeum Sztuki Współczesnej w Warszawie. Że nie wstąpię do polskiego piekielka konfesjonalnego, a gdy zauważyłem, że Warszawa zasługuje na coś dużo ciekawszego. (NKJP)

Cultural information

The expression was popularized by Tadeusz Mazowiecki who was a Polish publicist and politician. He was appointed the Prime Minister of the Third Republic of Poland and was in the office from 1989 to 1990. In his speech delivered at the meeting of Citizens Committee on 31 March 1990, he included the phrase, saying: “So as not to turn the budding democracy into the Polish hell (...) of quarrels, dishonest dealings and struggles”. The phrase was used as a label for the Polish bad characteristics which tend to affect the situation in the country. The unit several elements constituting the autostereotype of Poles, i.e. it summarizes how the Polish view themselves, with a focus of envy and unfulfilled ambitions, socially accepted complaining and problems with accepting the leadership.

Additional information on the contextual use of the phrase in texts of culture

[https://ksiegarnia.bellona.pl/?e=ksiazka&bid=5731](https://ksiegarnia.bellona.pl/?e=ksiazka&bid=5731)

Articles: [https://koduj24.pl/polskie-piekielko](https://koduj24.pl/polskie-piekielko)

Blog Polskie piekielko: [http://polskiepiekielko.pl](http://polskiepiekielko.pl)

Cartoons: [http://www.kwajk.pl/wylegarnia,110.html](http://www.kwajk.pl/wylegarnia,110.html)

Set in the framework of hybrid lexicography and combining linguistic aspects with the cultural ones, the inclusion of the above information allows for a better presentation of
the specifics of the described notion. The usage of the expression illustrates its broad scope and shows how a short phrase can be used in order to refer to many aspects of the autostereotype of Poles. It can be assumed that by consulting the dictionary, the language user will develop the learning autonomy of the person using the dictionary (cf. Yamada 2014; Szerszunowicz 2015). From the point of view of a non-native speaker of Polish, it is of great importance to learn about the linguo-cultural character of the phrase.

5. Conclusion

To sum up, it can be concluded that the idiom *polskie piekło / polskie piekielko* is a language label for a complex meaning embedded in the Polish culture. From the point of view of language users, it is very economical to use a short phrase to express a complex notion: whenever it is used in a given context, some of its semantic components dominate. Therefore, a non-native speaker of Polish should be presented with a comprehensive description of the meaning of the phrase.

The expression is culture-specific, thus it creates a cross-linguistic and cross-cultural gap. Its lacunary character and complex semantics make the lexicographic description difficult. The approach proposed in the paper aims at filling in the gap so that the prospective dictionary user could construct the mental image of what the idiom *polskie piekło / polskie piekielko* names.

Opting for the description of the meaning in semantic categories rather than offering the traditional definition results from the specific character of the unit. The notion of *polskie piekło / polskie piekielko* evades simple description: the broad semantic categories are open, which stimulates the ability of the dictionary users to interpret the phrase in a particular context, showing them the right direction of interpretation when they encounter the unit in new contexts. Thanks to such an approach, they will be able to understand the usage of the unit even though the actual meaning will not be precisely described in the entry. An additional benefit for the users is that fact that such organization of information within the entry fosters developing their autonomy in the process of learning a given foreign language.

In conclusion, it should be emphasized that the lexicographic presentation of such units as the one described in the present paper involves providing the prospective user not only linguistic but also cultural information. It results from the fact that understanding its complex meaning and cultural significance demands that the person who is acquiring the language to build a cognitive framework in which the phrase can be set. The proposed model allows for including more information which facilitates framework building potential of the user, thus increasing the chances of doing it successfully.
References

Dictionaries, lexicons, corpora and other reference sources


Other Literature

Pamies, A. 2017. The Concept of Cultureme from a Lexicographic Point of View.


Creating Dictionary Entries in a Bilingual Dictionary of Proverbs: The Case of Tshivenə

Munzhedzi James Mafela
University of South Africa
mafelmj@unisa.ac.za

Abstract

When reference is made to dictionaries, what comes to mind for many native language speakers is a lexicography document which provides the headword and its definition; for non-native language speakers, it is a lexicography document which provides the headword and its equivalent. People do not think much about the other types of dictionaries such as a dictionary of expressions. A dictionary is made up of dictionary entries, the components of which consist of some of the following: headword, spelling, pronunciation, word classes, morphology, syntax, semantics, usage, derived forms and cross-reference. It is not difficult to create dictionary entries for monolingual and bilingual dictionaries; but it is a challenge for lexicographers, especially in the developing languages such as Tshivenə, to create dictionary entries for a dictionary of expressions such as the dictionary of proverbs. Tshivenə, one of the minority languages in South Africa, has produced one dictionary of proverbs titled ṱhalusamaipfiyaMireroTshivenə – English: English - Tshivenə Dictionary of Proverbs. This paper seeks to highlight how lexicographers handled the problem of creating dictionary entries in ṱhalusamaipfiyaMireroTshivenə – English: English - Tshivenə Dictionary of Proverbs.

Keywords: Content word, dictionary entry, dictionary of expression, equivalent, function word, headword, keyword, lexicography, proverb, Tshivenə

1. Introduction

A dictionary is a lexicography document which provides the headword and its definition. Looking at aspects such as content and language, dictionaries can be classified into different types. When looking at the aspect of language, we can classify dictionaries into either monolingual, bilingual or trilingual dictionaries, to name a few. However, we can also talk about a concise or comprehensive dictionary when we look at the size of a dictionary; dictionary of idioms, dictionary of proverbs, dictionary of animals, etcetera. when looking at the content.

The first type of dictionaries produced in the African languages of South Africa, including Tshivenə, were bilingual dictionaries which were meant for learners of the language. When missionaries arrived in the country, they learnt African languages so that they could communicate with the native speakers. One way they used to achieve this was to compile bilingual dictionaries. Providing headwords and their equivalents in the foreign language was not difficult for them because they could easily enquire about the information from the native
speakers. It was difficult for them to compile dictionaries of expressions such as dictionary of idioms and dictionary proverbs. As a result, it took time before African languages could produce dictionaries of proverbs and idioms.

Proverbs and idioms are important in the life of an African. Grauberg (1989:97) says: “…isolated communities still exist in Asia and Africa where the social and economic structure has not changed for generations, and where proverbs, stories and songs constitute a repository of wisdom of continuing relevance.” Proverbs are metaphorical statements that reflect a general truth by reference to specific phenomenon or experience (Okpewho 1992). Okpewho (1992) further states that they are used when making speeches in public or for presenting cases in traditional legal disputes, spicing up the talk and making the points more firmly. This is testimony that proverbs and idioms are used in everyday conversations among Africans. Therefore, there is a need to record them in dictionaries for the purposes of learning by the youth in the native languages and non-native speakers. Whereas some literary texts have been animated with proverbs and idioms, some have been provided with proverbs as titles as mentioned by Grauberg (1989:96): “In the wake of Taverner’s work, and other collections, proverbs became widely popular in the sixteenth century. They provided titles of plays, were illustrated in tapestries and engraved on cutlery.” If a reader does not understand the meaning of proverbial expressions in a literary text, he or she misses the message. In the past the youth used to learn the meaning of proverbs and idioms through listening to oral prose narratives. With the introduction of modern technology, the narration of African folktales during the evening has ceased. The establishment of lexicography units by the Pan South African Language Board (PanSALB) played a role in producing the first dictionaries of proverbs in the indigenous African languages.

Seeing that proverbs are taken seriously in society as a storehouse of deeper philosophical truth and are used for such serious purposes of resolving conflict among citizens (Okpewho 1992), African languages started to produce dictionaries of expressions. In Tshivenda, Nêluvhalani (1987) produced a dictionary of idioms, IfaLashulaMaambele and the Tshivenda National Lexicography Unit produced a dictionary of proverbs, ThalusamaipfiyaMireroTshivenda – English: English - Tshivenda Dictionary of Proverbs, in 2012. Both lexicography works mentioned above are bilingual, with Tshivenda as the source language and English the target language. This does not imply that reference was not made to proverbs and idioms in the Tshivenda general bilingual dictionaries before the publication of the two dictionaries mentioned above. A dictionary user is confronted with proverbs, idioms and riddles in Van Wamelo’s Tshivenda – English Dictionary which was published in 1982. Van Warmelo (1982) made use of proverbs, idioms and riddles as illustrative examples in this dictionary. Mafela (2008:32) states that, in addition to the provision of parts of speech, translation equivalents, and phrases or sentences, Van Warmelosometimes adds proverbs to further enhance the user’s understanding of the meaning. However, it is difficult to find a proverb one is searching for in this dictionary because they are not listed as headwords, but as components of dictionary entry. In this regard, Kosch (2016:152) states: “Given their multilexical composition, there are a number of possible entry points of proverbs, which complicates the look-up process, because users may not be able to predict these entry points.” Moreover, the number of proverbs included is limited. Unlike Van Warmelo’s (1982) dictionary, Nêluvhalani’s (1987) dictionary of
idioms is user-friendly because the user can easily find the idiom as they can be searched through keywords.

It is not difficult for lexicographers to create dictionary entries for monolingual and bilingual dictionaries. The components that make a dictionary entry are common knowledge as Atkins (1985) writes that a dictionary entry will consist of some or all of the following components: headword, spelling, pronunciation, word classes, morphology, syntax, semantics, usage, derived forms and cross-reference. Nevertheless, it is a challenge for lexicographers, especially in the developing languages such as Tshivenda, to create dictionary entries for a dictionary of expressions such as the dictionary of proverbs. This can be explained through the view of Landau (1984) who sees dictionaries of proverbs as consisting of items that contain one or more sentences and are based on other criteria: traditional use in the first instance; especially aptness or wit, or the fame of the person quoted, in the second. As indicated above, Tshivenda, one of the minority languages in South Africa, has produced one dictionary of proverbs titled Thalusa maipfiya Mirero Tshivenda – English: English - Tshivenda Dictionary of Proverbs. The paper seeks to highlight how lexicographers handled the problem of creating dictionary entries in Thalusa maipfiya Mirero Tshivenda – English: English - Tshivenda Dictionary of Proverbs.

2. Definition of terms

The terms dictionary, dictionary entry, proverb and dictionary of proverbs seem to be key concepts which need a recap for the purposes of understanding the discussion that follows.

Many people are aware of the existence of the document called ‘dictionary’ because they refer to it when they want to check for the meanings of words. Van Sterkenburg (2003:4) cites Zgusta (1971) who defines a dictionary as follows:

A dictionary is a systematically arranged list of socialized linguistic forms compiled from the speech-habits of a given speech community and commented on by the author in such a way that the qualified reader understands the meaning ….

Landau (1984:5) on the other hand defines a dictionary as:

… a book that lists words in alphabetical order and describes their meanings. Modern dictionary often includes information about spelling, syllabification, pronunciation, etymology (word derivation), usage, synonyms, and grammar, and sometimes illustrations as well.

Both scholars emphasise the function of a dictionary as that of informing readers about the different meanings of words. Hacken (2009:400) adds to the above when he says: “… dictionaries pretend to be authorities on the existence and meaning of words.” Hacken’s (2009) assertion above means that if a user cannot find the word he or she looks-up in a dictionary, it is concluded that the word does not exist. Indeed, a dictionary pretends to be an authority of the existence of words. In his definition of ‘dictionary’ above, Landau (1984) further adds other information, i.e. spelling, syllabification, pronunciation, etymology, usage, synonyms, grammar
and illustrations, which enhance the understanding of the meaning. On the function of a
dictionary Fromkin and Rodman (1998:66) say:

All dictionaries, from The Oxford English Dictionary (often referred to as the OED and called
the greatest lexicographic work ever produced), to the more commonly used collegiate
dictionaries, provide the following information about each word: (1) spelling, (2) the 'standard'
pronunciation, (3) definitions to represent the word’s one or more meanings, and (4) parts of
speech, e.g. noun, verb, preposition. Other information may be included such as the etymology
or history of the word, whether the word is nonstandard (such as ain’t) or slang, vulgar, or
obsolete. Many dictionaries provide quotations from published literature to illustrate the given
definitions, as was first done by Johnson.

These elements, together with the headword are referred to as dictionary entry. Van Sterkenburg
(2003:4) in this regard states:

A dictionary distinguishes itself from other sources of information in that it does not offer
information in a coherent order, but divided into thousands of short chapters or sections. In
lexicography these are usually referred to as articles or dictionary entries, meaning the headwords
and everything that is said about them.

We can therefore briefly say that a dictionary entry is made up of the headword and all
information about the headword, i.e. information such as spelling, syllabification, pronunciation,
etymology, usage, synonyms, grammar and illustrations. Lemnitzer and Kunze (2005) point out
that most dictionary entries bear structural characteristics such as information items, part of
speech and equivalent(s), which serve the function of providing information about the headword

Many people are aware of the existence of dictionaries, whereas a few of them are
accustomed to a dictionary of proverbs. They are not aware that one can refer to a dictionary in
order to understand the meaning of a proverb. They think that a proverb is part of a community’s
speech which does not need to be checked in a dictionary, while others think that proverbs are
used by elderly people and do not bother themselves by checking their meanings.

Proverbs are very important in a community. Among other functions of proverbs, they
serve as carriers of the culture of a community. Okpewho (1992) defines a proverb as a piece of
folk wisdom expressed with terseness and charm. Guma (1977) sees a proverb as a bare
statement of fact, which is usually a short and simple sentence consisting of a subject and
predicate, which may either be positive or negative. Proverbs contain a truth about life accepted
by people and appeals to their imagination by the neatness and beauty in which it has been
framed (Okpewho, 1992:226). Seeing that proverbs are valuable in a language as they are
carriers of the truth about the life of a community, it is important to have a dictionary of proverbs
which serves as a reference for the understanding of these expressions by non-speakers of the
language and by the native youth. A dictionary of proverbs mainly focuses on the meanings of
the proverbs.
3. Dictionary of proverbs and dictionary entries

As stated by Landau (1984), elements that make up a dictionary entry are the headword, alphabetisation, pronunciation, grammatical information, equivalent, meaning, etymology, and illustrative examples. It is not difficult to create dictionary entries for common dictionaries such as monolingual and bilingual dictionaries because they are made up of the elements mentioned above. However, it is difficult for lexicographers, especially in the developing languages such as African languages, to create dictionary entries for dictionaries of expressions. Dictionaries of expressions include dictionaries of proverbs and idioms, just to name few. As proverbs are made up of sentences, it becomes a difficult task for a lexicographer to adopt a criterion which makes it easy for users to find proverbs for which they look-up. There should be some form of formal organisation of proverbs in order to make the dictionary user-friendly.

Commenting on the English proverb dictionaries, Predota (2003:97) says: “The oldest English proverb dictionaries completely lack any logical or formal organization of proverbs and idioms, making them not easy to consult.” Kosch (2016) states that lexicographers are faced with a challenge when it comes to the lemmatisation of proverbs. There are different ways of creating dictionary entries for a dictionary of proverbs. According to Predota (2003) a dictionary of proverbs can follow either of the following construction: alphabetical order of the first words, alphabetical order according to groups of themes, and alphabetical order of keywords.

Swanson (1975) notes that in creating a dictionary of proverbs, the lexicographer should first consider the number of entries. According to Swanson (1975:64), “Entries will be, as conventionally, primarily words.” In a more realistic selection of entries, Swanson (1975) mentions the following criteria: relative frequency, i.e. most frequent words, function words which constitute the syntactic cement of discourse, and word-formations.


ṰhalusamaipfiyaMireroTshivenđa – English: English -Tshivenđa Dictionary of Proverbs, is a bilingual dictionary of proverbs, published in 2012 by the Tshivenđa National Lexicography Unit under the editorship of S.L. Tshikota. As the title of the dictionary is too long because it caters for the two languages, that is, Tshivenđa and English, it will hence forth be referred to as Tshivenđa – English: English - Tshivenđa Dictionary of Proverbs. This lexicography work is a bi-directional bilingual dictionary, with the first part consisting of Tshivenđa as the source language and English the target language; and the second part consisting of English as the source language and Tshivenđa the target language. Lexicographers did not separate the two parts as it would be expected because the last headword (zwivhuya) of the first part is followed by the first headword (abdomen) of the second part without any form of break.

Generally, dictionary users expect to find the front matter, main matter and the back matter in a dictionary, even if they rarely read the front matter and the back matter. Tshivenđa – English: English -Tshivenđa Dictionary of Proverbs satisfies the expectation of an ordinary dictionary user because it consists of the three parts - the front matter, main matter and the back matter. The main function of the front matter is to guide users on how to use the dictionary. Some dictionaries
focus on the orthography and grammar of the language, whereas Tshivena – English: English - Tshivena Dictionary of Proverbs focuses on how headwords are presented in its front matter. In this regard, Tshikota (2012:vi) states:

Lemmas are written in bold letters. The first letter of proper names is to be capitalized. The lemma will be given appropriate translation equivalents. Proverbs and translation equivalents will be written in italics. Parts of speech are written in italics. The entries of the lemma will include the lemma, lemmatic address, speech category, translation equivalent of the lemma, the proverb and its literally translation equivalent and the proverb’s free translation and its translation equivalent, and the usage examples, …”

The information in the extract above helps dictionary users know what to expect regarding the presentation of dictionary entries in the main matter of Tshivena – English: English - Tshivena Dictionary of Proverbs. The information in the back matter of this dictionary includes a list of Tshivena archaic words whose sources are the Tshivena literary works. These are the words which are mainly used as headwords, even though some of them are not included as headwords, for example, mapenzhe, fhininaandlundani. The discussion of dictionary entries below concentrates on headwords, part of speech, equivalent, proverbs and meaning.

4.1 Headword

Tshivena – English: English - Tshivena Dictionary of Proverbs is a dictionary of proverbs as stated in the title. The arrangement of the proverbs is through keywords identified in the proverbs concerned. The keywords are arranged alphabetically to make the dictionary user-friendly. The importance of alphabetical order of keywords, as Sterkenburg (2003:98) argues is that: “Organisation by means of keywords in alphabetical order is currently unanimous choice of the paroemiologists, who regard it as the most appropriate and user-friendly organizational system.” The keywords are referred to as headwords which need to be defined. A keyword is the headword by which the word or expression being defined is identified. The headword should be in the standard form of the language, or from a preferred dialect (Landau 1984:76 – 77). Lexicographers in the dictionary under investigation use words which are in the standard form including archaic Tshivena words as headwords.

Since a proverb consists of many words, a dictionary user is bound to identify more than one keyword and to try to look-up the proverb. In this dictionary lexicographers make use of noun, verb, adverb, conjunction, and pronoun as keywords to look-up proverbs. A dictionary user will therefore use the trial method to find the proverb for which he or she is looking up. Many of the headwords in the dictionary are nouns, followed by verbs; therefore, one should identify the nouns first, followed by verbs before one can consider other parts of speech.

To make it easy to find the proverbs, keywords are written in bold and arranged alphabetically. This is in line with many dictionaries as Landau (1984) writes that dictionaries usually alphabetise letter by letter rather than word by word. One keyword may be found in a number of proverbs. As a result, one headword in Tshivena – English: English - Tshivena Dictionary of Proverbs may provide a number of proverbs in its definition as shown below.
lurangadzin calabash ndilurangaluyangamulila it is the calabash plant, it follows the path, ndimusiiriwana a tshitevhezela zwavhabebivhawe is said when the child is imitating parents; wakokodzaluranga, mafhuri a atevhela if you pull the calabash plant the pumpkins follow, aralimufumakadzi a malwa e navhana, navhanavha a mu tevhela a woman is married with her children (Tshikota, 2012:40)

In the example above, the keyword luranga (calabash) appears in the two proverbs, Ndilurangaluyangamulila (It is the calabash plant, it follows the path) and Wakokodzaluranga, mafhuri a atevhela (If you pull the calabash plant the pumpkins follow). Lexicographers have used the semi-colon to separate the two proverbs. To make it easier for dictionary users to identify the two proverbs under the headword luranga, lexicographers should have numbered them, instead of using the semi-colon. This has been adopted throughout the dictionary.

4.2 Part of speech

As indicated above, the function of a keyword is to guide the dictionary user towards finding the proverb which is being looked up. It is therefore not necessary to reflect on the grammatical information of the keyword or headword because the focus is on the proverb. However, it is important for a lexicographer to identify the part of speech of the headword, even though the dictionary of proverbs is about the meaning. Dictionary users would like to know the parts of speech of the content words and function words. Tshivenda – English: English - Tshivenda Dictionary of Proverbs make use of headwords which are nouns, verbs, adverbs, pronouns, and conjunctions for the identification of proverbs. These are categorised into content words (nouns and verbs) and function words (adverbs, pronouns and conjunctions). Few function words are reflected in the dictionary as there are not more than five pronouns, five conjunctions and ten adverbs used as headwords in the whole dictionary to identify the proverbs. Lexicographers relied much on content words because the majority of headwords are nouns, followed by verbs. All the parts of speech in the entries are written in Tshivenda without providing the English translation in the first part of the dictionary. The English translation of Tshivenda parts of speech is provided in the front matter together with the full form of the words. A dictionary user will have to refer to the front matter in order to know the English translation of the part of speech of a headword. Furthermore, only the abbreviation form is shown for a part of speech, for example, dzin for dzina (noun), and jiit for jiiti (verb), nfor noun and vfor verb.

Some parts of speech are incorrectly stated, for example, headword okhola is indicated as a noun (dzin) whereas it is a verb stem. Parts of speech of headwords all and although are written in Tshivenda as iša and išang respectively, instead of writing them in English (pronoun and conjunction) because English is the source language in this regard. This reveals that the dictionary was not properly edited before publication.

4.3 Equivalent

Before the proverb can be provided, the English equivalent of the headword is given. Al-Kasimi (1983) states that the major task of a bilingual lexicographer is to find appropriate equivalents in
the target language to the units of the source language. Since Tshivenđa – English: English - Tshivenđa Dictionary of Proverbs is a bilingual dictionary, it is important for lexicographers to provide equivalents of headwords, although the focus of the dictionary is on the proverb. The English translation of the headword assists non-speakers of Tshivenđa to understand the meaning of the keyword, and comprehend the proverb better. For example:

bokoṭodzin he-goat, billy-goat bokoṭolipfumbulįtshidzhenalįvulamapango when a he-goat enters, it opens the poles, muthuwaζiitsozvimisi e fhethu u vhonalangaziiitsozwawe a bad person shows himself wherever he is (Tshikota, 2012:3)
boynmutukananmutukana ha endi, huendamukalaha , a boy does not travel, an old man does, vhaaluwahanatshenzhemoyakutshimbidzelekwawithu u fhiraζhaswa old people have more experience than young people (Tshikota, 2012:170)

In the first example, the English equivalent of the headword bokoṭo is he-goat or billy-goat, the Tshivenđa equivalent of the English headword boy is mutukana in the second example. Although boy is the headword in the second example in which English is the source language, the Tshivenđa equivalent mutukana is used as a keyword to identify the proverb because only Tshivenđa proverbs are given. This is due to the fact that lexicographers of this dictionary do not provide the English proverbs, but English literal translations of Tshivenđa proverbs. Both equivalents in the examples above come after the part of speech, dzin (noun) and n (noun) respectively.

4.4 Proverb

Tshivenđa – English: English - Tshivenđa Dictionary of Proverbs is a bilingual dictionary of proverbs which is bi-directional. The focus of the dictionary is to provide a proverb in the source language and its equivalent in the target language, accompanied by the meaning in both languages. The element of dictionary entry that follows the equivalent of the headword is the proverb in which it (headword) is found and its accompanying target language equivalent. In the two examples given above, bokoṭolipfumbulįtshidzhenalįvulamapango and mutukana ha endi, huendamukalahaare proverbs in the two dictionary entries. Both proverbs are followed by literal English translations, when a he-goat enters, it opens the poles and a boy does not travel, an old man does respectively. Dictionary users expect to find proverbs both in the source and target languages; however, the proverb is provided in one language, that is, Tshivenđa. In the first part of the dictionary, where Tshivenđa is the source language, an English translation of the Tshivenđa proverb is provided instead of the English equivalent proverb. Lexicographers can give the English translation of the Tshivenđa proverb only if there is no English equivalent proverb. In the second part, with English as the source language, a Tshivenđa proverb is given with an English translation. Since English is the source language, an English proverb was supposed to have been given, accompanied by its Tshivenđa equivalent or translation. As this is a bilingual dictionary of proverbs, it should make it possible for two people who speak different languages to know and understand proverbs in both languages. Equivalent proverbs do exist between Tshivenđa and English, for example:
In the first example above, lexicographers define the headword maði as follows:

maðidzin water  maði a tevhuwa o tevhuwa, u kumbelandivhudenga if water poured out, is poured out, zwithuzwatshinyalazwotshinyala a zwitshavhuedzedzea  is said when things are difficult to repair; … (Tshikota 2012:43)

The English equivalent of the Tshivenđa proverb Maði a tevhuwa o tevhuwa, u kumbelandivhudenga is given as If water poured out, is poured out, which is a literal translation of the Tshivenđa proverb. The English equivalent of the Tshivenđa proverb is Do not cry over spilt milk.

### 4.5 Meaning

The last element of dictionary entry in Tshivenđa – English: English - Tshivenđa Dictionary of Proverbs is the meaning of a proverb. The meaning of a proverb is the most important element of dictionary entry in a dictionary of proverbs because in many instances this is what dictionary users search for. Proverbs are metaphorical in nature; they do not mean what individual words mean. It is the duty of the lexicographer to interpret the meaning of the proverb in line with the society’s interpretation. Lexicographers of this dictionary managed to provide meanings of proverbs as interpreted by the Vhavenđa. In many cases the pragmatic meaning of proverbs is achieved by using the verb as communicative action because proverbs communicate a message to community. Looking at the proverb Wako kodzaluranga, mafhuri a atehlela (If you pull the calabash plant the pumpkins follow) above, we will notice that it is sending a message to the community, that is, if a man gets married to a woman who has children by another man, he must be ready to take the children as his own. This action is to prevent a situation where the children will be left without a person who looks after them. The pattern of providing the meaning is the same in the whole dictionary. Lexicographers provide the meaning in Tshivenđa, which is followed by its English translation.

### 5. Conclusion

Tshivenđa – English: English - Tshivenđa Dictionary of Proverbs is the first dictionary of proverbs in Tshivenđa and it is a real contribution in the language because users have a reference work regarding finding the meaning of proverbs, especially school children, non-speakers of Tshivenđa, researchers and literature books readers. The Tshivenđa National Lexicography Unit should be commended on producing a work of this nature. The dictionary introduces dictionary users into the organisation of dictionary entries in the front matter.
The discussion above reveals that Tshivena – English: English - Tshivena Dictionary of Proverbs is made up of two parts; the first part consisting of Tshivena as the source language and the second part consisting of English as the source language. However, the two parts have not been separated as the tabulation of dictionary entries continue without any form of break. Lexicographers could have skipped one page and start with the second part to make the dictionary user-friendly. The front matter of the dictionary states that the entries of the lemma will include the lemma, lemmatic address, speech category, translation equivalent of the lemma, the proverb and its literal translation equivalent and the proverb’s meaning and its translation equivalent, and the usage examples. All the elements of the dictionary entry as mentioned above have been included in the definition except usage examples. The exclusion of usage examples denies dictionary users a better understanding of the meaning of the proverbs because they are not used in context. It is also important to note that not all archaic words listed on the back matter were used as headwords. Lexicographers relied much on nouns and verbs as headwords. Few pronouns, adverbs and conjunctions were used in this regard. A bilingual dictionary is about providing equivalents in the target language. In the two parts of the dictionary, with English as both target and source language, English proverbs are not provided. English proverbs should be provided if they exist. As it is, dictionary users are denied the opportunity to know proverbs in English in this bilingual dictionary of proverbs. Given the above-mentioned, focus on the issues can add value to Tshivena – English: English Dictionary of Proverbs, when the dictionary is revised.

References


How to Build a Specialized Chinese-English/English-Chinese Electronic Dictionary for Chinese Business English Learners

Zhiyi Zhou
Guangdong University of Foreign Studies
20160310047@gdufs.edu.cn

Abstract

During the learning process, in which the focus is shifted from acquiring knowledge of English to studying concepts of business in English, Business English (BE) learners need a specialized Chinese-English/English-Chinese electronic dictionary so that their business related knowledge can be structured efficiently and effectively. From the perspective of BE learners, this paper mainly studies the disadvantages of existing mainstream C-E/E-C electronic dictionaries and discusses how to build a specialized C-E/E-C electronic dictionary for Chinese BE learners. It cannot be denied that those electronic dictionaries are well-developed in China. However, they do have some disadvantages when being employed during the learning process by BE learners. (1) Existing electronic dictionaries are developed to satisfy the needs of all users rather than to merely emphasize on the demands of BE learners; (2) The knowledge that BE learners is provided by existing electronic dictionaries is fragmented, overlapped and even unprofessional information; (3) There is no explanation and comparison among synonymous business terms. Therefore, it is needed to take the needs of BE learners and the features of business terms into account so as to build a customized electronic dictionary of C-E/E-C business terms that (1) boasts a professional and specialized database; (2) updates timely so as to catch up with the development trend of economy and technology; (3) groups business terms that will be applied in the same register and provides learners with specialized equivalents, detailed definitions and practical examples; and (4) makes comparison among synonymous business terms for Chinese BE learners.

Keywords: electronic dictionary, Business English term, Business English learner

1. Introduction

“In 2006, Business English was approved by the Ministry of Education as an undergraduate major. Since then, Business English major has been successively introduced into 216 colleges and universities in China.” (Zhong et al. 2015) From then on, Business English (BE) related researches, accordingly, have always been hot topics, of which the study of BE dictionary is a significant part. The BE learners mentioned in
this paper mainly refer to full-time undergraduate BE major students in China. “The first two years of BE major students are filled with intensive English teaching and training courses with business background teaching materials so that the BE program can equip students with knowledge and skills of English, and meanwhile, pave the way for the next stage of business-course teaching in English.” (Zhang 2011) After the English learning and training courses of the first two years, majority of BE learners may have laid a solid foundation of general English knowledge and opt to further attach greater importance to the accumulation of business related knowledge in English. Furthermore, electronic dictionary has become a heated trend for both English learner and teacher. Therefore, during this learning process, in which BE learners are going through the middle phase from learning language to acquiring business related knowledge in English, a specialized C-E/E-C electronic dictionary that can serve as an encyclopedia for them to acquire business related knowledge is in great need.

There are few scholars at home and abroad who study on specialized C-E/E-C electronic dictionary. Fuertes-Olivera studies on the usability of free internet dictionaries for teaching and learning Business English (2012) and problems encountered in Spanish universities when applying internet dictionaries for teaching and learning Business English (2014). While in China, scholars pay more attention to study BE dictionaries from the perspectives of teacher and lexicographer. They carried out investigation into English dictionary use by BE majors (Zhang et al. 2012), corpus-based study of BE lexicography for English learners (Hu et al. 2013), studies on translation standards, principles and methods of terms in BE dictionary (Li et al. 2013), etc.

From the perspective of BE learners, this paper studies the disadvantages of existing mainstream C-E/E-C electronic dictionaries and discusses how to build a specialized electronic dictionary of Chinese-English for Chinese BE learners.

2. Disadvantages of Existing Electronic Dictionaries

The electronic dictionary mentioned in this paper mainly refers to online dictionary and offline dictionary (like apps). “The term electronic dictionary (or ED) can be used to refer to any reference material stored in electronic form that gives information about the spelling, meaning, or use of words.”(Hilary et al. 2000) Nowadays, in order to acquire the knowledge of business terms, BE learners in China are able to choose such mainstream electronic dictionaries as “youdao”, “Dict.cn” and “iCIBA”. “Adopts the users’ perspective and investigates the effectiveness of electronic dictionaries for different categories of users, in particular students and translators.” (Granger 2012) It cannot be denied that those electronic dictionaries are well-developed for English learners in China. However, they do have some disadvantages when being employed during the learning process by BE learners. The following tables are searching results that were taken from “youdao”.

(1) Existing electronic dictionaries are developed to satisfy the needs of all users rather than to merely emphasize on the demands of BE learners. They manage to
gather a huge number of vocabulary and example sentences through all kinds of electronic resources to ensure that each user can, at least, find the equivalents, definitions and examples of his or her target words. Nonetheless, they try to satisfy the needs of learners from all walks of life, but it turns out that they cannot perfectly meet the demands of everyone, which vary from person to person. As a consequence, for BE learners, these mainstream electronic dictionaries can only serve the basic needs of looking up simple equivalents of words up, since the general and specialized definitions of a certain word are mixed together, some miss the specialized meanings that this word contains and some even do not mark which subject this word belongs to.

The Chinese equivalents of “square” that is provided in *Oxford Business English Dictionary for Learners of English* is “平仓” as a noun. While there are a lot of general meanings but no business meanings of the word “square” that are demonstrated in “youdao” (see table 1), let alone a business tag.

<table>
<thead>
<tr>
<th>Headword</th>
<th>square</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chinese Equivalents</strong></td>
<td></td>
</tr>
<tr>
<td>n.</td>
<td>平方；广场；正方形</td>
</tr>
<tr>
<td>adj.</td>
<td>平方的；正方形的；直角的；正直的</td>
</tr>
<tr>
<td>adv.</td>
<td>成直角地</td>
</tr>
<tr>
<td>vt.</td>
<td>使成方形；与…一致</td>
</tr>
<tr>
<td>vi.</td>
<td>一致；成方形</td>
</tr>
</tbody>
</table>

(2) The knowledge that BE learners is provided by existing electronic dictionaries is fragmented, overlapped and even unprofessional information. Some BE terms merely have their Chinese or English equivalents and lack the further definitions and examples. In addition, although developers tend to share interesting and valuable Business English related news and articles with BE learners, those contents are fragmented rather than being systematically composed. And what BE learners is able to obtain from those news and articles is business related information rather than knowledge. Besides, equivalents, definitions and examples are gathered and displayed separately without a unified and systematic arrangement. In order to provide all-round and professional information of a word, these electronic dictionaries combine their own equivalents and definitions of a certain word with that from different kinds of dictionaries and websites together rather than integrate these information and delete the overlapped parts. In order to look a business term up in these existing electronic dictionaries, it is needed for BE learners to click link to link and sometimes there is even no approving results displayed on the screen. Hence, this may lead to a time-consuming searching process and make it harder for BE learners to find out the
knowledge they long for efficiently and effectively.

“Growth-Share Matrix” is explained in *Principles of Marketing* as “a portfolio-planning method that evaluates a company’s strategic business units in terms of its market growth rate and relative market share. SBUs are classified as stars, cash cows, question marks, or dogs.” (Kotler et al. 2011) with further elucidation of the “SUBs”. While, first of all, the information that is provided in “youdao” (see table 2) is limited and fragmented. There are merely short Chinese and English phrases serving as definitions and a sentence, which cannot be of significant help for BE learners to understand this term, as a sentence example. In addition, the equivalents of “占有率矩阵” and definitions of “BCG 矩阵” are showed for 2 times. Furthermore, there are 5 Chinese equivalents as “增长份额矩阵”, “占有率矩阵”, “BCG 矩阵”, “成长占有率矩阵” and “成长—份额矩阵” that are offered, which may confuse BE learners and make it hard to distinguish which equivalent is of great frequency and specialty.

<table>
<thead>
<tr>
<th>Headword</th>
<th>Growth-Share Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equivalents</td>
<td>增长份额矩阵;占有率矩阵</td>
</tr>
<tr>
<td>Online Definitions</td>
<td>BCG 矩阵;BCG 矩阵;占有率矩阵;成长占有率矩阵;</td>
</tr>
<tr>
<td>E-E Definitions</td>
<td>Growth-share matrix</td>
</tr>
<tr>
<td>Examples</td>
<td>This paper mainly used the analysis tools of the BCG's Growth-Share Matrix and the McKinsey &amp;Company's Three Horizons of Growth theories to study the adjustment of strategy of HGMP.</td>
</tr>
<tr>
<td></td>
<td>文中主要运用了波士顿咨询公司的成长—份额矩阵和麦肯锡管理咨询公司的三层面增长理论来研究汉光机械的战略调整问题。</td>
</tr>
</tbody>
</table>

(3) There is no explanation and comparison among synonymous business terms. The information of synonymous business terms is displayed in existing electronic dictionaries in an isolated way. In addition, even if BE learners look target terms up in these electronic dictionaries respectively, it is hard to spot the differences between these words. However, it is the differences of different entries, whether Chinese one or English one, that matters significantly in economic activities. In business situation, there are not only specific words but also general words with business meanings that are employed as Business English terms in different business situations. What’s more,
in certain cases, different business terms may have the same Chinese expression but they are required to be applied in different situations. These concepts are clear for experienced businessman, while ambiguous for callow BE learners. Therefore, this confusion of concepts may become a barrier for BE learners to integrate their knowledge and practice. And it is needed for those specialized electronic dictionaries to offer explanation and comparison among synonymous business terms for BE learners.

There are 8 equivalents of “授权” are offered in “youdao” (see table 3). These words can be applied to different objects, within different period and in different way. However, there is no further explanation and comparison among those synonymous equivalents. What’s more, “franchise”, which is not collected in this group of equivalents, can also serve as the English equivalent of “授权”.

<table>
<thead>
<tr>
<th>Headword</th>
<th>授权</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equivalents</td>
<td>n. authorization ; warranty; empowerment; delegation</td>
</tr>
<tr>
<td></td>
<td>v. accredit ; empower; license; authorize</td>
</tr>
</tbody>
</table>

3. How to Build a Specialized Electronic Dictionary of C-E/E-C for BE Learners

As the paper has analyzed above that existing electronic dictionaries have their disadvantages and cannot fully satisfy the needs of BE learners. Then it is significant to build a specialized electronic dictionary of C-E/E-C according to the needs of BE learners and the features of BE terms.

The focus is still on user-friendly, while it is supposed to be narrowed down to the perspective of BE learners. For BE learners, their needs is far beyond more words and more interesting functions. “Effective Second Language Acquisition (SLA) needs the assistance of a new generation pedagogical dictionaries. And being encyclopedic and specific are significant direction for future pedagogical dictionaries development.” (Zhang 2009) Therefore, during this learning process, the BE learners need a specialized C-E/E-C electronic dictionary that can serve as an encyclopedia to help them to form their business related knowledge structure so as to digest those intricate knowledge efficiently and effectively in future study.

In order to build a customized electronic dictionary of C-E/E-C business terms for BE learners, it is needed (1) to build a professional and specialized database. In order to build a professional and specific database, it requires the focus on the dictionary itself and the coordination among developers, lexicographers and professionals from general English teaching and Business English teaching. “In mainland China, electronic dictionaries were developed with electronic technology as basis. Early period products were developed independently by information technology enterprises,
and their commercial object was evident.” (Zhang 2007)

Nowadays, though developers have invited professionals to work on compiling dictionary together, they pay more attention to develop extra attractive functions rather than to the dictionary itself, and their emphasis is still on attracting visitors, drawing investments and increasing profits. Moreover, it is required to ensure that BE learners can, at least, find the Chinese definitions of his or her target words. “The invention of wiki technology has brought about one of the most spectacular changes in lexicography, i.e. the integration of collaborative or community-based input.” (Granger 2012) The developers of these dictionaries can also invite learned users to offer definitions for words without authorized definitions. In this way, so long as BE learners look their target words up in the dictionary, there would demonstrate relevant definitions and sentence examples.

(2) Besides, updating timely is significant for a specialized C-E/E-C electronic dictionary in that business terms need to be updated every now and then so that it can catch up with the development trend of economy and technology in this information age. What’s more, since business covers a wide range of fields, there generated a great diversity of terms derived from those fields. “(International) Business English, an inter-discipline program, takes English as the carrier and linguistics, management, economics, science of law and psychology as major courses.” (Zhang et al. 2008)

According to Oxford Business English Dictionary for Learners of English, which was published by Huaxia Publishing House in 2011, business terms can be further divided into following subjects as accounting, commerce, economics, HR, insurance, IT, law, marketing, stock exchange, technical, transport, etc. And Fuertes-Olivera (2012) reckons that “business/economics is a very broad subject field that can be broken down into around forty different sub-fields.” And it is due to this feature, the collection of BE terms are required to vacuum up the newest information and snap the representative terms so as to enable BE learners to keep pace with the changing times.

(3) It is also supposed to group terms that will be applied in the same register and provide learners with specialized equivalents, detailed definitions and practical examples. In order to provide learners with specialized equivalents, detailed definitions and practical examples, it is needed to combine resources from existing printed dictionaries, electronic dictionaries and learning websites with prudent selection and systematic arrangement. “Usually, the study of lexicography does not include the practical needs and cognitive ability of dictionary users, which, to a great extent, would influence the using effect of dictionaries.” (Huang et al. 2008)

From the perspective of BE learners, this dictionary, firstly, is developed to directly provide BE learners with a specialized and professional equivalents for those classic business terms and general words with business meanings. And then it offers detailed definitions and practical examples so as to serve as an encyclopedia for BE learners to obtain the information of target terms in an effective and efficient way. It may be able to, accordingly, be of great help on the knowledge construction phase of BE learners. “During the translation process of law related terms, the mere reference of general bilingual dictionaries is not enough. It is necessary to look them up in specialized Chinese and foreign language dictionaries.” (Zhang 1997)
Moreover, making comparison among synonymous business terms for BE learners is of great importance. In business situation, there are not only specific words but also general words with business meanings are employed as Business English terms. In addition, both specific and general content of a certain word would be involved. “Business English is usually presented as a hybrid that mixes specific and general content and is concerned with the teaching and learning of the strategic communication system in the business domain.” (Fuertes-Olivera 2012) And because business activities are carried out in different and specific situation, it can be sure that the definition of a business term is certain so long as it is properly used in that situation.

In order to make comparison among synonymous business terms for BE learners, the professional guidance of professionals who study on BE teaching and BE practicing is required. “Most ESP teachers have a language teaching background and do not have first-hand experience of the content and context of other discipline or business.”(Dudley-Evans et al. 1998) It cannot be denied that merely professionals of general English cannot distinguish clearly words from words applied in business situation in that Business English covers a wide range of fields and business terms are of great specialty, like jargon. Therefore, this requires the assistance of teachers who are familiar with business concepts and are expert at BE teaching as well.

4. Conclusion

Chinese BE learners aim to study the special use of English in business situation. After the English learning and training courses of the first two years, majority of BE learners may have laid a solid foundation of general English knowledge and opt to further attach greater importance to the accumulation of business related knowledge in English. “As the teaching and learning environment of Business English consists of an array of communicative-oriented (i.e. reading, writing, and translating) and cognitive-oriented (i.e. acquiring knowledge on business concepts) use situations.” (Fuertes-Olivera 2012) During the learning process, in which the focus is shifted from acquiring knowledge of English to studying concepts of business in English, the Business English learners need a customized specialized electronic dictionary of C-E/E-C business terms so that their business related knowledge can be structured efficiently and effectively. Therefore, from the perspective of Business English learners, this paper mainly studies the disadvantages of existing C-E/E-C electronic dictionaries and discusses how to build a specialized C-E/E-C electronic dictionary for Chinese BE learners.

It cannot be denied that those electronic dictionaries are well-developed in China. However, they do have some disadvantages when being employed during the learning process by BE learners. (1) Existing electronic dictionaries are developed to satisfy the needs of all users rather than to merely emphasize on the demands of BE learners; (2) The knowledge that BE learners is provided by existing electronic dictionaries is fragmented, overlapped and even unprofessional information; (3) There is no explanation and comparison among synonymous
business terms. Therefore, it is needed to take the needs of BE learners and the features of business terms into account so as to build an electronic dictionary of C-E/E-C business terms that (1) boasts a professional and specialized database; (2) updates timely so as to catch up with the development trend of economy and technology; (3) groups business terms that will be applied in the same register and provides learners with specialized equivalents, detailed definitions and practical examples; and (4) makes comparison among synonymous business terms for Chinese BE learners.

References


胡春雨，何家宁. 基于语料库的商务英语学习词典编纂研究——兼评《牛津商务英语学习词典》. 广东外语外贸大学学报，2013(06): 第38-41页.

李镔，何家宁. 商务英语学习词典术语词目翻译的标准、原则与方法——以《牛津英汉双解商务英语词典》(新版)为例. 长沙铁道学院学报(社会科学版)，2013(02): 第160-162页.


张相明，谢衍君，何家宁. 商务英语专业学生英语词典使用调查研究. 重庆邮电大学学报(社会科学版)，2012(03): 第116-121页.
张武保，商务英语专业本源论及动态模式．华北水利水电学院学报（社科版），2011（03）：第7-10页．

张武保，严新生．学科创新与商务英语的复合型特点研究．广东外语外贸大学学报，2008（02）：第101-104页．

章宜华．对我国电子词典发展策略的几点思考．辞书研究，2007（02）：第1-11页．

章宜华．关于双语教学与双语专科学习词典的几点设想——以英汉词典为例．外语界，2009（04）：第30-37页．

仲伟合，张武保，何家宁．高等学校商务英语本科专业的定位．中国外语，2015（01）：第4-10页．
Microstructure in Dictionaries of Turkish Grammar and Linguistics

Terms: Examples of Mood and Modality Terms

Duygu Kamaci
Eskişehir Osmangazi Üniversitesi
dkamaci@ogu.edu.tr

Abstract

Dictionaries of terms, which take their focus points from researchers, learners, and translators, are reference guides that provide information about a specific area for dictionary users. Within these, dictionaries of linguistics and dictionaries of grammatical terms have a place in special importance for who do scientific work in these fields. In the beginning Turkish dictionaries of terms, which are the primary aims to find the Turkish. Turkish equivalents of the foreign words, showed development and diversity in terms of lexicography in progress of time. However, this diversity has brought the problem of standardization on the other hand. In this paper five major printed, which published in the last 20 years, dictionaries of Turkish grammar and linguistics terms was examined comparatively in point of microstructure of mood and modality terms for drawing attention to definition and standardization problems. At the same time headwords (in other words records), information of etymology, usage labels and definitions, which are the elements of microstructure, were analyzed and discussed this question: how should be an ideal example of microstructure in a dictionary of grammar and linguistics terms? Finally some suggestions about the topic of the paper are offered. The purpose of this study, as a result of the analysis of the data obtained from the corpus, is to make a contribution to Turkish lexicography literature about dictionaries of terms and to express some expectations of dictionary users from dictionaries of term.

Keywords: dictionary of terms, mood and modality, microstructure, standardization

1. Mood and modality (concept, term and definition)

The concept of modality is defined in various forms by most researchers until today. The dynamics of these definitions are morphology, semantics, logic and philosophy. At the same time some researchers claim that modality can not be defined clearly (Van der Auwera and Plungian 1998; Salkie 2009; Declerck 2011).

Salkie (2009:79) identified four different approaches for definitions of modality. The first of these approaches is “a broad interpretation”. Palmer's definition (2001:1) “modality is concerned with the status of the proposition that describes the event” is an
example for this approach. Salkie criticizes this definition because of being broad. So in this definition the scope of the modality is not clear.

Another approach is “a narrow definition”. For instance “propose to use term ‘modality’ for those semantic domains that involve possibility and necessity as paradigmatic variants” (van der Auwera and Plungian 1998: 80), this format is in sufficient for the concept of modality because it includes another semantic domains like volition and evidentiality.

In the third approach which is determined by Salkie “a list of sub-types of modality”. A modality definition in this way “to be understood as a semantic category which covers such notions as possibility, probability, necessity, volition, obligation and permission” (Downing and Locke 2002:382) is helpful in terms of providing information about the field but for the definitions in this manner there is a problem of delimitation.

Finally Salkie expresses that “avoidance of the problem by focusing on a set of expressions”. Coates (1983) has searched use of English modals in her study but she does not even define the term of modal.

In brief Salkie emphasizes that the impossibility of the certain and unique description of the modality and offers four category for description of modality: possibility and necessity, epistemic and deontic, subjectivity and extremess of the modality scale (2009: 79-87).

In the context of our research four categories were identified in Turkish and foreign literature for modality definitions: morphological, semantical, logical and philosophical. In morphological definitions researchers focus on the relation between tense, aspect and modality. Bybee et al. (1994) and Nuyts (2005) stand out in this definition tendency. Bybee et al. say that it is not as easy to define the concept of modality as time and aspect. According to them modality need a comprehensive definition. A definition such as ‘grammaticalization of the attitudes and opinions of the speaker’ inadequate because modality is a comprehensive concept (1994: 176). Nuyts places modality at a higher level from time and aspect. Even though modality at the same time contain these categories (2005: 15-16).

The second type of modality definition is semantical definition. In this type instead of modal, modal verb and modalizervarious semantic categories as possibility, necessity, probability, permission, ability used while defining modality. Barbiers(2002:1-3), with this approach, describes modality as a semantic concept. According to him it expressed by modal verb, verb inflection, modal adverbs and modal particles syntactically. Moreover there are quite a lot of researchers who define modality semantically (von Fintel 2006, ErkEmeksiz 2008; Larreya 2009).

Different from morphological and semantical, in logical definitions some concepts are used as state, entity, event, entire statements and speaker’s attitudes. The most known researcher of the logical ones, Palmer (2001: 1) defines modality “concerned with the status of the proposition that describes the event”. Bybee(1986), also like Palmer, associating modality with the entities and the events adds entire statements to his definitions. Also Erguvanlı Taylan 1997, Corcu 2005, Güven 2015, Frawley 2009 preferred to describemodality in this way.
Another definition tendency philosophical definition of modality, from the point of view content close to logical definition, the terms of reference world, non-factual world and expressed world are used in these definitions. Beside these terms the notions of truth, possibility and necessity, which philosophical ones, are also used in variety definitions.

According to Chung and Timberlake (1985) modality encodes the comparison of expressed world and reference world. When this worlds overlap actuality modality is obtained, if not non-actuality modality (or irrealis modality) is obtained (Frawley 2009: 397). While defining the modality Çapak (2004) and Cevizci (2010) choose the terms of apodeictic and problematic in their definition approaches. According to Çapak (2004: 1444-145) modality is “the judgement on a judgement” and “the necessity, possibility, contingency and impossibility of something to be true” accordingly Cevizci(2010: 943).

As it is seen there are four different approaches in the literature for modality and all of which are related to each other. Through them when considered in terms of linguistics in our opinion logical definition is optimal for modality because of essential definition1. In conclusion modality is an semantic phenomenon that express the subjective approach of the speaker in the face of event or situation.

On the other hand for mood researchers preferred semantic, morphologic and pragmatic approaches for definitions. As it is known in philosophy, with the terms of mode and modus, mood is described as “the way of existence, appearance and present of something” (Cevizci 2010: 942-943). In semantics and morphology when defined mood the researchers consider based on the relation between mood and modality. Based on this relation for some researchers mood grammaticalized form of modality (Bybee et. all 1994; Bybee and Fleischman, 1995; Lyons, 1995, Palmer 2001, Frawley 2009).

According to Frawley (2009: 386) mood is a grammatical phenomenon and it is a morpho-syntactic device. In Turkish literature the majority of researchers accept Frawley’s opinion in accordance with general acceptance (Deniz Yılmaz 2004; Aslan Demir 2007, ErkEmeksiz 2008, Çürük 2010, Sargın 2014). In some studies (Karademir 2012; Kadiu 2012) mood is associated with verb and person categories at the same time.

Kocaman (1981: 85) describes mood in semantic framework, morphologic approach is insufficient for Kocaman because of possibility of disrupt linguistic integrity. Apart from these two approaches the term of mood is associated with utterance. In regard to van der Auwera and Plungian (1998) moods are main utterance types like declaration, interrogation, imperative and wish.

According to our opinion while modality is evaluated on the semantic platform, the mood must evaluate on the morphologic one. In that their difference comes proceed from these platforms. In this way mood is an morphologic phenomenon express the subjective approach of the speaker in the face of event or situation. This distinction between mood and modality will bring the question of “how these terms will take place

---

1 A kind of definition clearly specifying the concept completely different from others (Öner 1986: 36).
in dictionaries of terms?".

2. Microstructure of mood and modality in dictionaries of terms

Microstructure is an internal design of a reference unit which produces particular information, like spelling, pronunciation, grammar, definition, usage and etymology, about the headword (Hartmann and James 1998: 94):

In this part of the study we will discuss the microstructure of the terms mood and modality in five different dictionaries. The reviewed dictionaries are not include the part of comment on form (spelling, pronunciation and grammar), all of them have etymology information but only one of them has usage labels. Thus we will consider only comment on meaning (definition, etymology and usage label).

Dictionary of Grammar Terms (Dilbilgisi Terimleri Sözlüğü) which the earliest study in the corpus only has the term of mood. The headword is in this dictionary as it follows (Topaloğlu 1989: 101):

Topaloğlu has chosen the term of mood for modality, so that in the definition part of microstructure of headword mood has defined the notion of modality. In etymology information the term is given in French and Ottoman Turkish. Usage labels are not used in this dictionary.

Vardar (2002) gives both terms mood and modality in Expositional Dictionary of Linguistics Terms (Açıklamalı Dilbilim Terimleri Sözlüğü). Usage labels are not take place as in the previous one. He prefers to give in addition French, English and German etymology information. Both the concept of mood and modality have been defined in two different headwords at the same time. So it is impossible to precisely draw the boundaries between the definitions of mood and modality in the dictionary.
There is conceptual confusion in both definitions. Whereas modality is related to semantic and mood-related to morphology. We can say that this dictionary is not sufficient for definition but it is satisfactory in terms of etymology (2002: 135):

An important name in Turkish grammar with Dictionary of Grammar Terms (Gramer Terimleri Sözlüğü) Korkmaz (1992) just described the term of mood in her study as a grammatical item. This dictionary which can satisfy the dictionary user in terms of definition but it is deficient in regard to not having the term of modality (Korkmaz 1992: 103):

According to the criteria we have set in the previous section for mood and modality Dictionary of Linguistics (Dilbilim Sözlüğü) (İmer et all. 2011) is the closest dictionary to our idea among the others by definitions. This dictionary only gives English as etymology information. But unlike the others usage labels are used in the dictionary. According to these labels modality is a semantic term and mood both semantic and morphologic one. While usage labels is an innovation for such dictionaries the distinction between these labels is not as clear as the definitions (İmer
As Dictionary of Linguistics (İmreț et al. 2011) in Dictionary of Basic Grammar Terms (Temel Dilbilgisi Terimleri Sözlüğü) (Yılmaz 2014) mood and modality terms are described in detail. According to Yılmaz, mood is a grammatical category related with verb. On the other hand modality is associated with the truth-value. He does not mention that modality is a semantic category. In addition to this he does not use usage labels. The information of etymology only given in English (2014: 64):

**kip (Ing. mood)** Konuşmanın niyetine bağlı olarak bağlam içinde; -bil-dirmeye zannı ve çerçevesine zannının ayırdığını çerçevesinde kendisinin veya diğer iletişimsel katımcıların da eylemsel/durumsal bir hazırlanmış olmadığını, böyle bir hazırlık içinde olduğunu, böyle bir hazırlanmış planlı veya plansız, yakın veya uzak zaman diliminde gerçekleştirileceğini, ya da eylemsel/durumsal bir hazırlanıla ilgili olarak yine kendisinin veya diğer iletişimsel katımcıların duaryu, düşüncesi, tasarısı, istek ve tespitleri aktardığı fillerle ilgili dil bilgisel kategoriyi denir.

**kiplik (Ing. modality)** Kiplik; fiili karışıldığı anlamanın doğruluk ve geçerlilik ölçüdüği açısından kaynağını bilgi ve anıncıardan alan "bilgi (olasılık)" veya eylemin gerçekleştirilmesini isteyen/sağlanan kişitenin özel bakış açısını yansıtan ve bunu gerçekleştirmesi beklenen kişi/kısıların tutumu gosteren "yükümlülük (gerekliilik)" işlemlerile ifade edilmesidir. Sözdiziminde genellikle "bilgi (olasılık) kipliği" (Ing. epistemic modality) yargısal önerme yapısında karşılanırken, “yükümlülük (gerekliilik) kipliği” (Ing. deontic modality) olay odaklı olduğu için eylemler aracılığıyla karşılırlar. Olasılık kipliği, olası (Ing. possible) ya da olasılık (Ing. possibility) bildiren sözlüksel birimlerle de karşılırlar: “It’s possible they’ll phone this evening” ya da “Maybe/perhaps they’ll phone this evening”.

et al. 2011: 175-176):
3. Conclusion

The first thing that stands out in the examined dictionaries there is no certain distinction between mood and modality. Moreover two of these dictionaries do not contain headword of modality (Topaloğlu 1989 and Korkmaz 1992). Thus we can say that the term of modality was subsequently entered into the dictionaries (with Vardar 2012). Once modality entered the dictionaries there has been a conceptual confusion between mood and modality. The only dictionary reflecting the semantic and morphological separation between these two terms is Dictionary of Linguistics (İmer et al. 2011) according to our point of view. These two different concepts must be accurately reflected in the dictionaries. This also shows that headword’s entry into dictionaries does not make any sense in itself, so microstructural parts such as definition, usage label play an significant role in themselves. The distinction between the terms of mood and modality must be taken into consideration while defining.

Another thing that attracts attention is about etymology information. In three of the dictionaries (Topaloğlu 1989, Korkmaz 1992 and Vardar 2012) this information was given in languages such as German, French and English. In two of them the information was given in only English. This tendency is seen in recent years (İmer et al. 2011, Yılmaz 2014). It may be related to the volume of dictionaries or to the English literature. This alteration can be evaluated positively in regard to volume of dictionary but may not be satisfactory from a dictionary user perspective. It brings to mind the following question: What is the scope of the etymology information in dictionary of terms?

Usage label which a kind of lexicographic indicator (Burkhanov 1998: 256) has importance as the others. The conceptual distinction between mood and modality is consolidated by usage label as morphology or semantics. The only dictionary that uses a usage label while defining terms is Dictionary of Linguistics (İmer et al. 2011). Usage labels are not available in others. It is a question whether usage labels will be used for dictionaries to be written or to be printed after that.

All these differences bring the problem of standardization among dictionaries of terms. In particular the scope and typology of grammar and linguistics dictionaries should open to discussion. Lexicographers should think on the question of is there a difference between these two types, or should it a combination of grammatical and linguistic terms?

As a result we can say that the closest microstructure to our opinion is available in Dictionary of Linguistic (İmer et al. 2011). Because usage label is very important for dictionary users. It is also very significant that the conceptual distinction between mood and modality is made clearly in the dictionary.
References


Some Aspects of Language in Samuel Johnson’s *The Rambler*

Masaaki Ogura
Osaka Ohtani University
oguramasa@osaka-ohtani.ac.jp

Abstract

The 18th century saw a surge of interest in the purity of language, which was advocated for by a well-known group of writers, who mourned the ‘deterioration’ of language and insisted on English being fixed. Samuel Johnson, the author of *The Dictionary of The English Language*, is said to have been among them initially. In previous discussions of Johnson’s view of language such as in Lynn (1992), scholars have analyzed the thoughts he presents in his works, but his language usage itself has not yet been systematically studied, although it is likely to tell us what kind of language he preferred and could deepen our understanding of his language philosophy. I hold that analyzing Johnson’s use of language can shed light on his philosophy, and thus this study examines aspects of Johnson’s usage in *The Rambler*, a collection that conveys his thoughts on various topics. Through examining frequently used expressions in *The Rambler*, I attempt to identify some aspects of usage that differentiate *The Rambler* from other works of Johnson’s age. In the discussion of his language, a corpus-stylistic approach is used, making it possible to draw numerical comparisons between literary works. I contend that knowing Johnson’s language preferences in this work will be significant because this work was in circulation between 1750 and 1752, a period that overlaps with the time when Johnson was creating his dictionary, which raises the possibility of knowing his language use during the compilation process reflected in his own writing.

**Keywords:** Samuel Johnson, *The Rambler*, corpus stylistics, word-clusters, writing styles

1. Introduction

Samuel Johnson, who is known for his works such as *A Dictionary of the English Language*, *The Vanity of Human Wishes*, and *The Rambler*, has been said to have preferred to employ a unique writing style that includes the use of abstract and academic words, complicated sentences and rhetoric and among studies on these aspects is Wimsatt (1948) who focused on how Johnson’s language developed in *The Rambler*. Johnson is also known for his unique perspective and attitude on language and some potential change in his attitude towards language during the process of compiling the dictionary has been discussed as in Barnbrook (2005). Johnson is said to
have wanted to fix the language in vain and resorted to writings by canonical authors for usage examples in his dictionary. However, there can be no adamant conclusion that he regarded those authors’ writings as an ideal for inclusion in his dictionary. Rather, there might be a possibility that those writings were the only recourses that were available to assess. Examining which interpretation is right, though there might be no clear-cut conclusion, requires an investigation of his writing preferences. To be more specific, comparison between his own writings and sentences included in the Dictionary can reveal his attitudes towards language; when writing styles included in the Dictionary are similar to Johnson’s own writing, this suggests that Johnson resorted to literary works for usage examples in accordance with his ideals. However, a comprehensive analysis of Johnson’s writings has not been attempted, except for some studies regarding his vocabulary use as conducted by Brocklebank (2013). When discussing a writing style, examining what kind of phrases Johnson used should be the center of focus, which can bring part of his language patterns under the spotlight. As part of this attempt, the present study intends to explore some frequently used phrases in The Rambler and tries to uncover some language aspects that are recurrent and unique to this work.

2. Method

The Rambler is a periodical collection of essays on various topics in circulation during the period between 1750 and 1752, mostly contributed by Samuel Johnson, though some essays were written by other authors. In the present study, which seeks to specifically investigate aspects of Johnson’s writing styles found in The Rambler, essays written by other writers were excluded from the data. The text data was obtained from Project Gutenberg and spelling was adjusted to modern English versions, in accordance with Baron et al. (2009), who investigated the influence of spelling variations on the research results obtained and reported that the date with corrected spellings provides more robust results. Dates found before the start of each essay were also eliminated. To analyze the text, WordSmith Tools 7.0 (Scott 2017) was used and the number of tokens used for the word list was 295,076. In this study, a phraseological approach was employed, without starting from a particular hypothesis. With this bottom-up approach, I expect some features of Johnson’s writing to emerge. The unit of word-clusters, also referred to as n-gram, is adopted for this study, where word-clusters of three to five words are examined. I did not confine the length of the clusters to five words so that shorter clusters thought to be found in other literary works of his age could be included in the list.

3. Analysis

The following is a list of top 25 3-5 grams clusters extracted from The Rambler.
Table 1 25 3-5 grams clusters extracted from *The Rambler*

<table>
<thead>
<tr>
<th>N</th>
<th>Word Cluster</th>
<th>Raw Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>OF THE WORLD</td>
<td>105</td>
</tr>
<tr>
<td>2</td>
<td>IT IS NOT</td>
<td>91</td>
</tr>
<tr>
<td>3</td>
<td>OF THOSE WHO</td>
<td>78</td>
</tr>
<tr>
<td>4</td>
<td>IT MAY BE</td>
<td>69</td>
</tr>
<tr>
<td>5</td>
<td>THE POWER OF</td>
<td>68</td>
</tr>
<tr>
<td>6</td>
<td>IS TO BE</td>
<td>65</td>
</tr>
<tr>
<td>7</td>
<td>ONE OF THE</td>
<td>65</td>
</tr>
<tr>
<td>8</td>
<td>OF HIS OWN</td>
<td>63</td>
</tr>
<tr>
<td>9</td>
<td>PART OF THE</td>
<td>60</td>
</tr>
<tr>
<td>10</td>
<td>THE REST OF</td>
<td>60</td>
</tr>
<tr>
<td>11</td>
<td>THOSE WHO HAVE</td>
<td>57</td>
</tr>
<tr>
<td>12</td>
<td>IT HAS BEEN</td>
<td>54</td>
</tr>
<tr>
<td>13</td>
<td>IN THE SAME</td>
<td>53</td>
</tr>
<tr>
<td>14</td>
<td>THOSE WHO ARE</td>
<td>53</td>
</tr>
<tr>
<td>15</td>
<td>FOR WANT OF</td>
<td>51</td>
</tr>
<tr>
<td>16</td>
<td>NOT TO BE</td>
<td>48</td>
</tr>
<tr>
<td>17</td>
<td>THE HAPPINESS OF</td>
<td>48</td>
</tr>
<tr>
<td>18</td>
<td>THE NECESSITY OF</td>
<td>48</td>
</tr>
<tr>
<td>19</td>
<td>OF THE SAME</td>
<td>46</td>
</tr>
<tr>
<td>20</td>
<td>WITH ALL THE</td>
<td>46</td>
</tr>
<tr>
<td>21</td>
<td>AND THAT THE</td>
<td>45</td>
</tr>
<tr>
<td>22</td>
<td>THAT I WAS</td>
<td>44</td>
</tr>
<tr>
<td>23</td>
<td>OF THE MIND</td>
<td>42</td>
</tr>
<tr>
<td>24</td>
<td>OF THEIR OWN</td>
<td>42</td>
</tr>
<tr>
<td>25</td>
<td>THAT IT IS</td>
<td>42</td>
</tr>
</tbody>
</table>

As can be seen, these clusters are relatively frequently used in this text. However, this does not necessarily mean that these phrases are unique to *The Rambler*. When a particular writing style is uniquely seen in a text, it is a hasty conclusion to attribute this to the individuality of the author because there is no denying that particular style is shared among other groups of works that are contemporary to a particular text in which that style is employed. In that case, it is possible that *The Rambler* just shared a tendency to use these phrases with other authors during his age. Therefore, it is necessary to examine the frequency of these phrases in other writings of his period. Here, *The Corpus of Late Modern English Text*, 3.0 (*CLMET*, 3.0) is used to this end. This is a large historical English corpus, which comprises approximately 34 million words, made up of three sub-period sections as shown in the following table.
### Table 2 The number of words in CLMET 3.0

<table>
<thead>
<tr>
<th>Sub-period</th>
<th>Number of authors</th>
<th>Number of texts</th>
<th>Number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1710-1780</td>
<td>51</td>
<td>88</td>
<td>10,480,431</td>
</tr>
<tr>
<td>1780-1850</td>
<td>70</td>
<td>99</td>
<td>11,285,587</td>
</tr>
<tr>
<td>1850-1920</td>
<td>91</td>
<td>146</td>
<td>12,620,207</td>
</tr>
<tr>
<td>TOTAL</td>
<td>212</td>
<td>333</td>
<td>34,386,225</td>
</tr>
</tbody>
</table>

(https://perswww.kuleuven.be/~u0044428/clmet3_0.htm)

This large corpus has genres of drama, letter, narrative fiction, narrative nonfiction, others and treatises. Since *The Rambler* was published during the period between 1750 and 1752, the sub-period section of 1710-1780 was used to examine the frequency of clusters presented in Table 1. Since the present study attempts to locate some aspects of writing styles that are unique to *The Rambler*, three texts by Samuel Johnson included in the subcorpus were eliminated. The following table shows the frequency per million words in each genre and *The Rambler*.

### Table 3 The adjusted frequency in each subcorpus

<table>
<thead>
<tr>
<th>N</th>
<th>n-gram</th>
<th>rambler</th>
<th>drama</th>
<th>letter</th>
<th>narrative fiction</th>
<th>narrative non fiction</th>
<th>other</th>
<th>treatise</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>OF THE WORLD</td>
<td>356</td>
<td>76.850</td>
<td>126.00</td>
<td>94.793</td>
<td>113.98009</td>
<td>118.04</td>
<td>136.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>29083</td>
<td>69393</td>
<td>27383</td>
<td>27656</td>
<td>27656</td>
<td>17131</td>
</tr>
<tr>
<td>2</td>
<td>IT IS NOT</td>
<td>308</td>
<td>100.86</td>
<td>205.38</td>
<td>122.49</td>
<td>53.190709</td>
<td>268.08</td>
<td>351.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>60067</td>
<td>13893</td>
<td>54178</td>
<td>50761</td>
<td>50761</td>
<td>21653</td>
</tr>
<tr>
<td>3</td>
<td>OF THOSE WHO</td>
<td>264</td>
<td>7.2047</td>
<td>25.796</td>
<td>44.583</td>
<td>46.677561</td>
<td>117.33</td>
<td>130.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>14765</td>
<td>69624</td>
<td>13792</td>
<td>16646</td>
<td>16646</td>
<td>09067</td>
</tr>
<tr>
<td>4</td>
<td>IT MAY BE</td>
<td>234</td>
<td>55.236</td>
<td>49.609</td>
<td>59.949</td>
<td>55.90421</td>
<td>191.99</td>
<td>214.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>14653</td>
<td>03122</td>
<td>17089</td>
<td>72694</td>
<td>72694</td>
<td>50339</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>28747</td>
<td>87686</td>
<td>35988</td>
<td>47</td>
<td>11646</td>
<td>0433</td>
</tr>
<tr>
<td>6</td>
<td>IS TO BE</td>
<td>220</td>
<td>67.244</td>
<td>236.13</td>
<td>86.136</td>
<td>20.082206</td>
<td>191.99</td>
<td>140.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>00448</td>
<td>89886</td>
<td>35385</td>
<td>78</td>
<td>72694</td>
<td>35454</td>
</tr>
<tr>
<td>7</td>
<td>ONE OF THE</td>
<td>220</td>
<td>91.259</td>
<td>569.51</td>
<td>272.26</td>
<td>499.34135</td>
<td>452.97</td>
<td>153.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>72036</td>
<td>16785</td>
<td>01335</td>
<td>79</td>
<td>13355</td>
<td>59947</td>
</tr>
<tr>
<td>8</td>
<td>OF HIS OWN</td>
<td>214</td>
<td>33.622</td>
<td>87.311</td>
<td>114.05</td>
<td>119.95047</td>
<td>121.59</td>
<td>99.726</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>00224</td>
<td>89496</td>
<td>49208</td>
<td>84</td>
<td>82706</td>
<td>38706</td>
</tr>
<tr>
<td>9</td>
<td>PART OF THE</td>
<td>203</td>
<td>81.653</td>
<td>128.98</td>
<td>135.91</td>
<td>278.97984</td>
<td>339.19</td>
<td>596.54</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>43401</td>
<td>34812</td>
<td>36438</td>
<td>56</td>
<td>51759</td>
<td>51153</td>
</tr>
<tr>
<td>0</td>
<td>THE REST OF</td>
<td>203</td>
<td>60.039</td>
<td>105.17</td>
<td>110.59</td>
<td>104.75313</td>
<td>64.710</td>
<td>68.901</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>28971</td>
<td>11462</td>
<td>21528</td>
<td>27</td>
<td>19079</td>
<td>86742</td>
</tr>
</tbody>
</table>
Based on this cross-tabulation, correspondence analysis was conducted to see how the phrases in Table 3 differentiate The Rambler from other literary works.

4. Results

Correspondence analysis produced the following result in Figure 1, which shows a bi-plot of the result, on which the relation between the phrases and the texts are...
observed. The Eigen value of dimension 1 and dimension 2 is approximately 69%.

Figure 1 The result obtained from correspondence analysis

Among the clusters in Figure 1, for want of, those who have, the happiness of, those who are, of those who, the necessity of, the power of, it has been and it may be are in the same quadrant, indicating that these clusters differentiate The Rambler from others. I then examined what kind of surroundings these phrases occurred within. Let me start with for want of, the happiness of, the necessity of, and the power of as these phrases seem to have a noun+of frame. The phrase for want of is followed by considering, knowledge, employment, and assurance more than twice. The following excerpts would be a typical example of the use of phrases meaning ‘necessity’. In this line, Johnson associates the notion of knowledge with dignity, suggesting that without sufficient knowledge, one cannot understand the dignity of a person’s business.

*But let him not too readily imagine that another is ill employed, because, for want of fuller knowledge of his business, he is not able to comprehend its dignity. (The Rambler, No.9)*

The similar phrase the necessity of occurs 48 times and interestingly 28 instances are with the gerund form after of and the similar thing applies to the power of. This phrase is used 68 times in The Rambler and there are 39 instances where the gerund form follows this chunk, suggesting that Jonson insisted on the importance of actually doing something. As shown in this excerpt, Johnson sometimes argues for the importance of
some quality of human beings by connecting that to something else important in life and this can be seen in contexts where other words related to human quality others, mankind, life, are often collocated by the happiness of.

There is indeed some tenderness due to living writers, when they attack none of truths which are of importance to the happiness of mankind, and have committed no other offence than that of betraying their own ignorance or dullness. (The Rambler, No.93)

Such phrases as those who are, those who have and of those who are also worth attention. Their relatively frequent use indicates the possibility that Johnson wrote some articles to discuss some general characteristics of people as in;

It has been always the practice of those who are desirous to believe themselves made venerable by length of time, to censure the new comers into life, for want of respect to grey hair and sage experience, ... (The Rambler, No.50)

But surely nothing can more evidently show the value of this quality, than that it recommends those who are destitute of all other excellencies, and procures regard to the trifling, friendship to the worthless, and affection to the dull. (The Rambler, No.72)

It is interesting to see that another key phrase, it has been, is used to describe a human quality and that for want of, which was already discussed, is used to put emphasis on important human qualities. This interpretation is potentially drawn by the frequent use of it has been with observed. Interestingly, among 36 instances where it has been is used, there are 11 instances where observe follows and 4 instances where remarked follows, which indicates that it has been is often used to describe Johnson’s observation of human nature. In addition, it may be is followed by such words as observed, doubted, remarked, considered, and thought, each of which is used in a context where Johnson discusses general issues. From these points, it can be said about Johnson’s writing style in The Rambler through the clusters used in The Rambler that his work is different from other writing genres as he uses phrases that convey his life philosophy. On a surface, essays in The Rambler discuss various issues but the results obtained lead us to observe that his philosophy lies at the center of his arguments.

5. Conclusion

The present study conducted a phraseological examination of The Rambler to assess some features in Johnson’s writing style. Based on the concept of word-cluster, or n-gram, frequent phrases consisting of 3 to 5 words were extracted and, based on the top 25 clusters, the author attempted to locate what differentiates The Rambler from other literary works produced in the same time period. The analysis revealed that some phrases do play a role in making The Rambler unique in its writing style, specifically in conveyance of Johnson’s message regarding human nature. However, the analysis utilizing word-clusters only tells some aspects of his language use and the result itself
is a tentative conclusion that triggers further investigation. To be specific, since word-clusters used in the present research contained content words such as happiness, want, or necessity, it is easy to note the analysis is potentially influenced by a genre of writing. Whether these phrases set The Rambler apart from other writings when this work is compared with some other philosophical writing where these words are expected to appear more often is worth examination. In addition, in terms of syntax, it would be necessary to examine the text using p-frame, by which it might be possible to observe Johnson’s writing at a more abstract level.

References


Two Lexicographers of English Collocation Dictionaries in Japan: Katsumata and Irie

Kazuo DOHI
Tokyo City University
dohik@tcu.ac.jp

Abstract

This paper considers two Japanese lexicographers and the collocation dictionaries they compiled in the first half of the last century in Japan. The compilers are Senkichiro Katsumata (1872-1959) and Iwae Irie (1866-1929). Th main focus of this paper is on Katsumata’s Kenkyusha’s Dictionary of English Collocations (1939) and Irie’s A Grammatical Dictionary of the English Language (1915). The former is better-known than the latter partly because of its revision in 1958 and the publication of its handy version in 1960 after Katsumata’s death. My study shows that the two compilers, who never had an opportunity to study abroad, had a positive influence on each other for the development of English collocation dictionaries in Japan.

This paper will show how the two Japanese compiled their dictionaries, what they included in their collocation dictionaries, and how they influenced on each other. Irie had a strong interest in grammatical collocation, while Katsumata was keen to work with lexical collocation after he developed the framework for an idiomatic English dictionary printed in 1909.

The prototype of Katsumata’s dictionary can be traced back to his 1909 work; the framework of the former work strongly influenced the classification of his English collocation dictionary. His 1911 work motivated Irie’s first collocation dictionary. Although Irie’s dictionary was not as systematically organized as Katsumata’s, Irie’s dictionary clearly had a partial influence on Katsumata’s.

This paper will show how the two lexicographers contributed to the development of collocation dictionaries for Japanese learners of English.

Keywords: Senkichiro Katsumata, Iwae Irie, English collocations, dictionary
1. Introduction

This paper, based on my previous studies (Dohi 2012, 2016), aims to show the development of what is called English collocation dictionaries in the first half of the twentieth century in Japan. Here, primary focus is on the dictionaries compiled by two authors: A Grammatical Dictionary of the English Language printed in 1915 by Iwae Irie, and Kenkyusha’s Dictionary of English Collocations printed in 1939 by Senkichiro Katsumata. These Japanese lexicographers deserve special mention for two reasons: Firstly, they are considered to have been the first teachers of English in the world who were interested in collecting English idiomatic combinations or collocations, and, secondly, they greatly influenced each other in the compilation of their dictionaries of English collocations. It should be added that neither of them had any opportunity to study abroad and so they were eager to provide tools to Japanese people for mastering English expressions. This resulted in them both compiling English collocation dictionaries.

2. Iwae Irie

Irie, who had studied and taught German, got involved in proofreading Shinshiki Nichi-Ei Jiten (A Japanese–English Dictionary) printed in 1905, although his name does not appear anywhere in its title page or preface.1 The dictionary contains an appendix, which consists of ‘Table of the irregular verbs’ and ‘Words used with various prepositions.’ Of particular interest is the 14-page list of expressions in the latter: Examples include ‘Admission to (a person), into (a place)’ or ‘Break into (a house), (oneself) of (a habit), through (restraint), (ill news) to (a person), with (a person).’ I believe that Irie introduced the above-mentioned list into the dictionary. The appendix mainly includes example phrases of what may be called ‘grammatical collocation’ (Benson, Benson and Ilson 2009). The main reason for showing the detailed grammatical collocation to Japanese learners at the time is that they were not good at distinguishing the subtle differences in the use of prepositions or adverbs. Indeed, the turn of the twentieth century saw publications of many reference books related to the use of particles (prepositions and adverbs), such as Student’s Mistakes in English: How to Avoid Them (1905) by Eastlake, or Eigo Zenchishi Yoho fu

1 Irie’s describes his painstaking contribution to the dictionary compilation in his essay Jisho Hensan Kushidan, written around 1929.
Setsuzokushi [How to Use English Prepositions and Conjunctions] by Nakahara (1898). Therefore, in the development of collocation descriptions in dictionaries, the fact that Irie introduced the information on the usage of selected words with their related prepositions or adverbs in the appendix of a dictionary at the beginning of the last century is of great importance.

After compiling work on the above Japanese-English dictionary, Irie produced a few other dictionaries: Chukai Waei Shin Jiten (A New Critical Japanese-English Dictionary) (1907) and the smaller-sized Shuchin Waei Jiten (Yuhodo’s Pocket Japanese-Dictionary) (1914). In addition to these two dictionaries, he was also actively involved in a lexical-oriented dictionary: Shokai Eiwa Jiten (A New English-Japanese Dictionary) in 1912, which gives detailed information on grammar and usage of entry words. However, this dictionary did not sell well due to fierce competition with the other English-Japanese dictionaries at that time. The appendix of the 1912 dictionary also contains ‘Zenchishi Yoho [How to Use Prepositions],’ showing which particles to use with the entry words. The 1905 dictionary enumerates 1334 entry words with their particles shown, and the 1912 dictionary shows 2157 entry words, which amount to 60 percent more than the previous dictionary. Although the entry words and their description are quite different, Irie clearly focused on particle description in both dictionaries.

Irie’s great contribution to collocation is found in the dictionary of Eibumpo Jiten (A Grammatical Dictionary of the English Language)(1915). He says in its preface that many Japanese learners of English feel difficulty in using even fairly easy words because they are used in different grammatical constructions. He also writes that he compiled the dictionary so that learners can easily and conveniently understand each word’s usage. What he means by ‘grammatical constructions’ is not entirely clear because he does not distinguish between ‘grammatical collocations’ and ‘lexical collocations.’ A brief study reveals that Irie’s work includes both types of collocations, as is shown below in the entry word pain. (Note that the numbers, which are not given in the original, are shown here to classify the examples.)

---

2 The dictionary was reprinted under the different title of Eisakubun Jiten (A Dictionary of English Composition) in 1918. The reason for the title change is, he wrote, that the use of ‘grammatical’ in the first dictionary title misleads users into thinking that it focuses on explaining grammatical terms although the compiler’s intention was that it should be used as a reference for making clear English grammar and usage of words in writing.
an acute (or a sharp), a severe, a shooting, a slight, a sudden, a violent pain

This ointment will allay (or soothe, relieve) the pain.

She suffers greatly from a pain in the back.

The pain is still almost more than she can bear.

I feel a pain in my chest.

The pain leaves him for a time.

He is in great pain.

He was mad with pain.

This pain is enough to drive one mad.

The examples are divided into five groups: (1) indicates an adjective and noun combination, (2) indicates a verb and noun combination, (3) indicates a noun and preposition combination, (4) indicates a noun (as a subject) and verb combination, and (5) indicates a preposition and noun combination. The examples given show grammatical as well as lexical collocations, though Irie does not make the distinction clear. Irie selected these from approximately three hundred thousand examples he collected. He apparently thought alphabetical arrangement with some collocate words shown in bold is convenient for users to understand, but the above entry shows that this is not always appropriate for collocation examples.

Irie’s dictionary not only deals with nouns, verbs and adjectives as entries but also pronouns, adverbs, conjunctions, prepositions, definite and indefinite articles, and interjections. Irie collected as many authentic examples from as many sources as he could. But partly because of his short editorial work for the dictionary, it is not systematically organized. Still, the dictionary can be regarded as the first dictionary of English collocations in Japan.

3. Senkichiro Katsumata

Katsumata, who began to study English after he got a job for foreign exchange at a post office, was extremely energetic in studying the language and printed several reference books of English. His first experience in dictionary compilation was with
an idiom dictionary where he helped develop the framework in the classification of English idioms. The work was later printed as *Eiwa Sokai Jukugo dai Jiten (A Dictionary of English Phrases with English and Japanese Explanations)* (1909), but his name was only mentioned in its preface, and he was not listed as one of the editors.\(^3\)

The turn of the twentieth century in Japan saw a lot of reference books or dictionaries of English phrases such as the one mentioned above. Katsumata classified English idioms into at least eighteen types in the idiom dictionary. Examples of five patterns which he considers to deserve special attention are shown for reference (examples are all taken unchanged from the original): (1) transitive verb + object (take place), (2) verb + preposition + noun (go to the hammer), (3) verb + adverb (swear off), (4) adjective + noun (heavy rain), and (5) preposition + noun (on paper). Katsumata’s detailed classification shows that not only was he interested in idiomatic combinations in English but also he tried to make their patterns explicit so that Japanese learners could understand more accurately. His development of the framework for this trailblazing dictionary undoubtedly encouraged him to study idiomatic expressions in more detail, which contributed to the compilation of his collocation dictionary (1939).

I briefly mention a second important work by Katsumata for reasons given below. It is titled *Eiwa Reikai Yogo dai Jiten (A Dictionary of English Particles and Other Grammatically Important Words)* (1911), but only the first volume was printed. That volume deals with more than 130 selected words (between A and L), and includes many adverbs and prepositions as well as auxiliary verbs, conjunctions, pronouns, and adjectives. It is not considered to have contributed greatly to description of collocation. The book’s description, however, reveals that Katsumata was more interested in the senses, the usage and the patterns of individual English words, or the idiom of each word, than in the systematic classification of English.

---

3 The preface says, “It goes without saying that a work on a scale like this could hardly have been brought through by single-handed labor. The plan was originally laid down by Mr. Katsumata of Waseda University, collection and arrangement of materials having gone on under his guidance for more than three years, when circumstances obliged him to hand over the work to the present editors who...have at last been enabled to bring it to a successful conclusion after nearly four years’ incessant labor....” This suggests that Katsumata was engaged in the preparation of the dictionary around 1903 through 1905, roughly corresponding with time when he began to collect what is considered “idiomatic combinations” in English.
The third work by Katsumata considered here is *Eiwa Katsuyo Gosen-ku (Exercises in Phrase-Making)* (1920). It focuses on the description of verb and noun combinations, such as pay [attract, draw, turn] attention. Its preface claims that the aim is to focus on how to use English verbs, specifically which nouns can be used with which verbs. In this way, it aims to show idiomatic English combinations. Katsumata says that every language usually has conventional noun and verb combinations. To violate them leads to unidiomatic English. According to the book, the transitive verbs shown below can be used in combination with the noun attention, although not all of them actually used.  

*attract, bestow*, call, claim*, devote, distract, draw, enage*, engross*, fix*, focus, give, have, invite*, occupy*, pay, receive, require, rouse*, show*, stimulate*, take*, turn*

The first and most comprehensive collocation dictionary Katsumata compiled is *Eiwa Katsuyo dai Jiten (Kenkyush’a Dictionary of English Collocations)* (1939). Its preface says that the dictionary was compiled based on the original idea that idiomatic combinations or collocations can be classified into some types. For example, the noun ntry pain in Katsumata (1939) has far more collocation phrases or sentence examples than Irie’s. Especially worthy of attention is his focus on a few major types: V. (=verbs), Q. (=adjectives) and P. (=prepositions). The first category V. shows verb collocates with the noun, the second Q. shows adjectives or adjectival collocates, and P. shows prepositions or prepositional collocates. The dictionary also includes two minor types: M. (=modifiers) and O. (=others). M. includes adverbs or adverbial collocates, while O. includes other types of expressions including set phrases and grammatical collocations. Part of the descriptions of the noun and the verb entries for pain given below show the level of detail that went into Katsumata’s dictionary (Japanese translations are deleted here).

**pain.** n. V.

*allay* the pain of toothache/ *alleviate* pain/ *bear* the pain with fortitude/ This *eased* his pain slightly/ *experience* sore pains/ I *feel* a pain in my side/ it gives me pain

---

4. The words shown with an asterisk are not included in the present-day English collocation dictionaries for foreign learners: BBI (2009), Oxford (2009), and Macmillan (2010).

5. Katsumata used the term ‘idiomatic combinations’ but Masuda (1959: 29) writes that later in the 1930s the term collocation was used after H.E. Palmer came to Japan.
to.../ mitigate pain/ mollify pain/ palliate pain/ reduce the pain to the minimum
Q. The pain was so acute that it seemed like having cut with a knife/ assiduous pains have been taken to.../ I have a bad pain here/ a burning pain/ I have a good pain in my tooth
P. He is at great pains to do this work/ I have been at some pains to investigate the matter

pain. v. M.

It pains me deeply.
P. The father is pained by his boy’s conduct

(all examples are taken unchanged from the original; V. indicates verb collocates with the noun, Q. indicates adjectival collocates, M. indicates adverbial collocates, P. indicates prepositional collocates)

Compared with his 1920 work, *Eiwa Katsuyo Gosen-ku*, Katsumata’s collocation dictionary gives Japanese learners four times more examples to show what he considers to be collocations in English. However, not all of the examples he gives are actually strong or restricted collocations. In order to make the difference clear, compare the transitive verbs given below, which were added in the 1939 work and excludes phrasal verbs here, as collocates with the noun attention. Even considering that English collocations in the reading materials in those days are not exactly the same as those in the present-day English, not all of them may be regarded as strong collocations. This indicate that Katsumata endeavored to collect as many collocations as he could for his reference work, even if they were not actually strong collocations. His determined effort deserves special attention in the historical development of collocation dictionaries in the first half of the last century. (See footnote 4 about the use of asterisks.)


6 Badger & Rodale (1937) also shows some of the following verbs: absorb, arouse, arrest, excite, fasten, provoke, relax, secure, sustain, and win.
4. Their influence on each other

Irie and Katsumata exerted a positive influence on each other in the development of collocation dictionaries. Irie’s dictionary is not systematically organized in its description. For example, in addition to nouns, verbs and adjectives, its entries include proper nouns, grammatical terms, and foreign expressions. Secondly, the parts of speech do not exactly correspond in some examples. A few style labels, such as poetic, in the front matter are nowhere to be found in the body of the work itself. No distinction is drawn between grammatical and lexical collocations. The core of Irie’s dictionary focuses on its examples. However, as the example sentences or phrases in the above-mentioned entry for pain show, they are merely alphabetically ordered, and the use of boldface is not necessarily appropriate. The number of examples is not always proportional to the importance of entries. Although Irie devoted himself to collecting a huge number of examples for his work, his superhuman effort did not lead to the finished work he had expected, because he spent a limited time compiling it.

To examine the influence they had on each other, I will focus on the entries and their examples in Katsumata (1911), Irie (1915) and Katsumata (1939). A brief comparison of some of the entries in Katsumata (1911) and Irie (1915) reveals that the former undoubtedly has influenced the latter: below are shown their examples for the adverb entry already. It is clear that Irie partially plagiarized or imitated Katsumata. The entry for against, for example, also includes similar or identical examples from Katsumata.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Comparison of the examples for already</th>
<th>(Dohi 2016 : 56)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Katsumata (1911)</td>
<td>Irie (1915)</td>
</tr>
<tr>
<td>1</td>
<td>Light the lamp – It is lighted already.</td>
<td>It is lighted already.</td>
</tr>
<tr>
<td>2</td>
<td>Before this letter reaches you, you will have already reached home.</td>
<td>Before this letter reaches you, you will have already reached home.</td>
</tr>
<tr>
<td>3</td>
<td>Does he seem to be recovering ? – He has almost recovered already.</td>
<td>He himself was there already.</td>
</tr>
</tbody>
</table>

7 The preface in Irie’s 1918 edition says that he came up with the idea of compiling this type of dictionary, which would later be considered a ‘collocation dictionary’, around the time he was working on the 1905 Japanese-English dictionary. He always kept it in mind while he collected materials for two Japanese-English dictionaries and an English-Japanese dictionary printed in 1907, 1914 and 1912, respectively.
To clarify whether Irie (1915) had a positive influence on Katsumata (1939), a brief comparison is made of the entries, which shows that Katsumata (1939) partially follows Irie (1915) in its entries. Table 2 below shows the number of entries found in both or either of them. Although Katsumata added many of his own entries, it is clear that most of the entries are found in both. This supports the claim that Katsumata referred to Irie in the selection of entries for his collocation dictionary.

Table 2  Comparison of the number of entries in four parts of Irie (1915) and Katsumata (1939)  
(Dohi 2016 : 59)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>E</th>
<th>O</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irie &amp; Katsumata</td>
<td>36</td>
<td>44</td>
<td>46</td>
<td>45</td>
</tr>
<tr>
<td>Irie</td>
<td>14</td>
<td>6</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Katsumata</td>
<td>17</td>
<td>54</td>
<td>21</td>
<td>23</td>
</tr>
</tbody>
</table>

Table 3 shows that Katsumata (1939) also contains some example sentences or phrases that are identical or similar to those in Irie (1915). The entry accede is another case where examples are taken from Irie.

Table 3  Comparison of examples in abstain  
(Dohi 2016 : 60)

<table>
<thead>
<tr>
<th>Irie (1915)</th>
<th>Katsumata (1939)</th>
</tr>
</thead>
<tbody>
<tr>
<td>to abstain from luxuries</td>
<td>abstain from food</td>
</tr>
<tr>
<td>to abstain from quarreling</td>
<td>abstain from flesh and fish</td>
</tr>
<tr>
<td>Abstain from speaking ill of others</td>
<td>abstain from luxuries</td>
</tr>
<tr>
<td>must entirely abstain from wine</td>
<td>abstain from speaking ill of others</td>
</tr>
<tr>
<td>Abstain from all wrong</td>
<td>abstain from quarreling (wine)</td>
</tr>
</tbody>
</table>

Irie printed his ‘grammatical’ dictionary nearly a quarter century before Katsumata’s collocation dictionary came out, which makes a fair comparison between them difficult. But both Irie and Katsumata were interested in grammar and usage of each individual word in English, and they endeavored to compile a comprehensive dictionary of English collocations. Their partial plagiarizing of each other’s work may have been a result of their not always being able to collect enough materials to

---

8 Four parts, in each of which fifty entries are included in Irie’s work, are examined: a through absurdity, each through emanate, O through of, and sack through saying. It should be remembered that Katsumata’s work was printed about a quarter century later and has fifty percent more pages. The printed letter size and the book sizes are also quite different.
use in their dictionaries together with their awareness of the detailed descriptions in their competitor’s work.

5. Conclusion

In the first half of the last century, the idea of copyright protection did not exist, and dictionaries often imitated the descriptions in other dictionaries or reference works. This study shows that Katsumata and Irie were both interested in idiomatic combinations or collocations at the beginning of the last century, and they each compiled a reference book for Japanese learners. In the pre-electronic age, all they could do for collecting materials was to read as much as possible in their own way, taking a careful note of related expressions. This, however, often leads to the compilers’ subjective bias in their choice of materials and presents serious difficulties in accurately selecting idiomatic combinations or collocations, whether they are grammatical or lexical. But their genuine and unbounded enthusiasm for a new type of dictionary, a dictionary of English collocations, enabled them to compile their own reference works.

The following reasons are given for Katsumata’s and Irie’s contribution to compilation of English collocation dictionaries: The first is that both were eager to study the English language, although they had quite different educational backgrounds and had no experience of studying abroad. The second is that both were engaged in other reference works: Irie was involved in compiling an English-Japanese dictionary, and Katsumata worked on a book of English conversation and a book of everyday English words and phrases before he began to study idiomatic English combinations. The third is that both were interested in each individual word and tried to show their senses, grammar and usage as well as appropriate examples. The fourth is that both had experiences in teaching the English language, however long or short it may have been. In addition, Katsumata was actively involved in translating from Japanese into English, which compelled him to collect idiomatic combinations for improving his writing. They came up with the idea of compiling a new type dictionary as their mission in life.

Katsumata and Irie worked hard and did all they could to compile their collocation dictionaries, whether or not they were conscious of the differences between grammatical and lexical collocations, or whether or not their dictionaries include strong or restricted, weak, or free lexical collocations. Their hard labor and their unending perseverance in the task of collecting factual data were instrumental in the gradual development of English collocation dictionaries as well as the interest
during the twentieth century in the idea of collocations for more accurate English expression.

6. Further research

In terms of the historical development of English collocation dictionaries in the first half of the last century in Japan, although Katsumata is most widely considered to have contributed to the compilation of collocation dictionaries, Irie also made important contributions. In those days they collected authentic English from reading materials, such as newspapers, magazines, books, encyclopedias, brochures, leaflets, and letters. Katsumata is said to have used letters he took out of wastebaskets. Katsumata was a professor at a university, submitted many articles on English collocations to The Rising Generation, and also was deeply engaged in using practical English in his part-time jobs, while Irie was only regarded as a dictionary compiler. For material or data collection, Katsumata is famous for his ‘note-book habit,’ by which he wrote down anything he found useful in his notebook while he was reading, giving a lecture, or taking a walk. He never smoked nor drank, and he spent his money on notebooks. That was because he wanted to collect useful and instructive materials for the purpose of compiling a dictionary so that Japanese people may write in English accurately. His activities in a wide range of areas made his dictionary well-known. In contrast to this, it has not been made clear how well Irie was valued for his arduous work. His death prevented him from revising what is regarded as the first collocation dictionary in Japan. Although it was reprinted many times, his dictionary has rarely been the center of scholarly attention.

A number of other important considerations are also still unclear, such as how many copies of their dictionaries were printed or sold in the first half of the twentieth century, how influential the dictionaries were in teaching English at school, and whether or not teachers of English and students used either or both of them. This has created insufficiencies in my study, and it will take time to investigate these matters because the publishers do not always want the facts known. But the fact that Katsumata’s original dictionary was revised before his death, and that a handy version was also printed just after his death makes it clear that his dictionary was favorably received and sold well. Irie’s dictionary was reprinted often until the 1990s, which suggests that Irie’s dictionary sold well in previous years as well. Although Japanese are regarded to have been the first people in the world to be interested in idiomatic combinations or collocations, still much research is needed to clarify the development of collocation dictionaries in the last century.
References


Abstract

The relationship between two or more words which are identical in form but not in meaning calls “homonymy”. The identity may be in pronunciation (‘homophones’, e.g. *fair* and *fare*), in spelling (‘homographs’, e.g. *minute* ‘division of time’ and *minute* ‘tiny’) or both (‘homologues’, e.g. *band* ‘ribbon’ and *band* ‘group of musicians’) (Hartmann and James 1998: 69). Ottoman Turkish which written with Arabic script has many homographs and interlingual homographs (or false friends) because characteristics of Arabic script sometimes has more than one way of reading and also Ottoman Turkish has many borrows from Arabic and Persian. Because of this reason Ottoman Turkish and Ottoman Literature (or Divan Literature) is very rich in respect to homographs and false friends. This is also very important for Turkish poems and prose and literary arts.

Turkish Lexicography has a long history which starts at 10th century and a great number of dictionaries have been written in a large area until today. The most important Turkish dictionaries were written in 19th century and in this century there were many different type of dictionaries. One of them is Eser-i Şevket which written in 1851 by Mehmed Şevket Efendi. This dictionary is a pun dictionary and has many homographs and false friends. Mehmet Şevket Efendi aimed to prepare a dictionary which Ottoman poets utilize. This dictionary has 745 page and 17,473 headwords and has some different macro and microstructures and also has some different headword types.

The aim of this study is to make an analysis about structure of Eser-i Şevket and to show how to present homographs and interlingual homographs in this dictionary.

**Keywords:** homonymy, homophone, homograph, Eser-i Şevket, dictionary of homonymy, Turkish Lexicography, false friends.

1. Basic Concepts

Before talking about the dictionary, we need to talk about some basic concepts about lexical semantic. According to Günay, *polysemy* indicates an economic rule and a functional state at the top, but we cannot speak for the *homonymy* on the same
In general, homonymies have different origins and these words have been pronounced in the same way depending on their use over time.

*Homonymy* is the relationship between two or more words which are identical in form but not in meaning and lexemes are to be regarded as two “different” words (Hartmann and James 1998: 69; Svensén 2009: 95) and if the homonymy is about only spelling, it is named as *homograph* and if the homonymy is about only pronunciation it is a *homophone*.

According to Zgusta homonymy begins at the point when the speakers of a language are unable to conceive different senses are connected and he gives an example from Russian mir “world” and mir “peace” as connected then they are not single words with polysemous meaning and he says that we have to regard both pairs of words as pairs of homonymous words as pairs of homonyms (1971: 74) and according to Hartmann and James the decision as to whether two words are homonymous or different senses of the same period may not be easy to make criteria include, in addition to spelling, pronunciation and meaning, the etymology of the word(s), their grammatical status and the scope and user convenience intended (1998: 69).

Svensén says that polysemy is manifested in the microstructure of the dictionary, the lexemes being presented as different senses within one and the same entry, homonymy is manifested in the macrostructure of the dictionary, the lexemes being presented in different entries (2009: 96).

Another concept that need to be mentioned “false friend”. “One of two or more words or phrases from different languages which are similar in form but not in meaning are false friends (Hartmann and James 1998: 56). They can be compared to homonymies in a single natural language. False friend is also has two type: *interlingual homographs* and *interlingual homophones*. These concepts can be shown as in the Figure 1.:
Table 1 Distinction about homonymy

<table>
<thead>
<tr>
<th></th>
<th>SPELLING</th>
<th>PRONUNCIATION</th>
<th>MEANING</th>
<th>SAME LANGUAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>homonymy</td>
<td>+/-</td>
<td>+/-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>homophone</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>homograph</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>interlingual</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>homograph</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>interlingual</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

2. The Distinctive Features of Ottoman Turkish and Contemporary Turkish about Homonymy

Until the adoption of the new Latin alphabet, Turkish people used Arabic script from 10th century to 1928. Contemporary Turkish has not too much homonymy except inflected words but we cannot say same thing for Ottoman Turkish because of Arabic script. Ottoman Turkish which written with Arabic script has many homographs and interlingual homographs (or false friends). There are some reasons about that.

First of all Turkish is a phonetic language and has nine vowel. On the other hand Arabic script has only three vowel letter and because of this reason some words has more than one reading e.g. صن “son” (final) and صن “sun” (present) or اوت “ot” (grass) and اوت “öt” (warble) etc. The comparison can also be seen from the Table 2.:  

Table 2 Comparison of Arabic vowels and Latin vowels for Turkish

<table>
<thead>
<tr>
<th>Arabic letters</th>
<th>Latin letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a, e</td>
</tr>
<tr>
<td>و</td>
<td>o, ö, u, ü, v</td>
</tr>
<tr>
<td>ى</td>
<td>i, i, y</td>
</tr>
</tbody>
</table>
Secondly, consonants of Arabic letters also cause different readings. For example اوک "ön" (front) and اوک "avun" (console) has same spelling but different pronunciation (Table 3):

**Table 3** Comparison of Arabic consonants and Latin consonants for Turkish

<table>
<thead>
<tr>
<th>Arabic letters</th>
<th>Latin letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>اوک</td>
<td>k, g, ğ, тж</td>
</tr>
</tbody>
</table>

Another reason is that Ottoman Turkish has many borrowings from Arabic and Persian and if the reader does not know the word whether it is Arabic or Persian or Turkish, (s)he cannot know the right pronunciation. Only if the reader see all context, maybe (s)he can guess the right pronunciation but if (s)he does not know Arabic or Persian morphology some, will not find the correct pronunciation again.

According to Yavuzarslan, in historical and contemporary dictionaries, lexicographers come up with homograph and homonymous words at the beginning of the problems that are dragged into the error and in the dictionary of Ottoman Turkish and Chagatay Turkish which were written with Arabic script about 13.-19. century, homographic and homonymous words were given in the same entry because compilers do not know their etymological information (2010: 66). Most compilers gave them in the same entry because they think that the word is polysemous.

But these features are not always bad. Because of them, Ottoman Turkish and Ottoman Literature (also known Divan Literature) is very rich in respect to homographs and false friends. This is also very important for Turkish poems and proses and literary arts. Pun (in Turkish “cinas”) is a very important literary art and Ottoman poets use homographs and false friends at their poem and proses for harmony. By going this way Mehmed Şevket Efendi, who was a poet, wrote a dictionary at 1851 and named it “Eser-i Şevket”.

### 3. Eser-i Şevket

As mentioned above Eser-i Şevket is a pun dictionary and has a lot of homographs and false friends. It was written 1851 by Mehmet Şevket Efendi who was actually a poet. Mehmet Şevket stated the dictionary type and audience and purpose of dictionary clearly on the preface and we learn from the preface the dictionary’s type and also we learn that the user is poets who writing poems with Ottoman Turkish.

The dictionary consists of 745 pages and one volume. It includes 17,473 lemmata and has alphabetical order. But the alphabetical order has some different features. Not only initial but also final alphabetical order. For example, as you can see from Figure 2, the title shows that this chapter (fasıl el elit bi el hâ) includes elif (ı) initial letter and hâ (ح) final letter. The dictionary includes 29 chapters (bab) and every chapter includes 29 chapters inside (fasıl). Babs show initial letter, fasilş show final letter.
Figure 2 The macrostructure of Eser-i Şevket

In Eser-i Şevket each entry continues in a prose and follows one another. There is not any headword or entry column. This creates difficulty for the user when searching for headwords and definitions. Headwords are given in parentheses (see Figure 2.) and every homographs and false friends were given consequitive.

In Eser-i Şevket, every microstructure, information is given in a special order (see Figure 3.).
These information categories is not be found every headword. Especially grammatical information is given only these headwords which the inflected verb or plural name is a headword.

4. The Analyse of Eser-i Şevket and Its Structure about Homonymy

Because Eser-i Şevket is a pun dictionary, there are a large number of homonymies in this dictionary. In this chapter some examples are shown from dictionary.

First example is homographic examples (see Table 4.). These headwords are same spelling but different pronunciation and compiler shows pronunciation information clearly. Meaning is totally different. Meaning of first word is “cup, plate”, meaning of second word is “to delay”. Compiler did not give part of speech information but first one is a name, second is a verb.

<table>
<thead>
<tr>
<th>Headword</th>
<th>Arabic script</th>
<th>Pronunciation</th>
<th>Etymology</th>
<th>Definition</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>inâ’</td>
<td>أناء</td>
<td>bi-kesi’l-elîf ve meddi’in-nûnû’l-meftûha ve sükkûnû’l-hemze</td>
<td>Arabî</td>
<td>Kâse ve kap manasına cemi âniye ve evâni gelir.</td>
<td>6</td>
</tr>
<tr>
<td>enâ</td>
<td>أناء</td>
<td>bi-fethî’l-elîf ve meddi’in-nûnû’l-meftûha ve sükkûnû’l-hemze</td>
<td>Arabî</td>
<td>Te’hîr eylemek ve eğlenmek manasına.</td>
<td>6</td>
</tr>
</tbody>
</table>

Second example is interlingual homograph (see Table 5.). These headwords are same.
spelling but different pronunciation and different languages. Compiler gave three interlingual homographs. Languages are Arabic, Persian and Turkish. Meaning is totally different. Meaning of first word is “smoked eyes and eye disease”, meaning of second word is “yesterday night and shoulder” and meaning of third word is “dream and bosom”.

Table 5 Examples about interlingual homograph in Eser-i Şevket

<table>
<thead>
<tr>
<th>Headword</th>
<th>Arabic script</th>
<th>Pronunciation</th>
<th>Etymology</th>
<th>Definition</th>
<th>P.</th>
</tr>
</thead>
<tbody>
<tr>
<td>deveş</td>
<td>دوش</td>
<td>bi-fethateyn</td>
<td>Arabî</td>
<td>Göz dumanlı olmak manasına ve bir adamın gözü bir maraz sebebiyle fâsid olmak manasına.</td>
<td>274</td>
</tr>
<tr>
<td>düş</td>
<td>دوش</td>
<td>huş vezinde</td>
<td>Farsî</td>
<td>Dün gece manasına ve omuz manasınadır.</td>
<td>274</td>
</tr>
<tr>
<td>düş/düş</td>
<td>دوش</td>
<td>bi-zammi'd-dâl</td>
<td>Türkî</td>
<td>Rüya manasına ve ağnâmın göğüs mahalidir.</td>
<td>274</td>
</tr>
</tbody>
</table>

Third example is inflected homograph (see Table 6.). These headwords are same spelling but different pronunciation and different inflection. Compiler gave three inflected homographs. Languages are all Arabic but compiler did not give this information. Meaning of first word is “(s)he spilled”, meaning of second word is “Something is spilled” and meaning of third word is “spill”.

Table 6 Examples about inflected homograph in Eser-i Şevket

<table>
<thead>
<tr>
<th>Headword</th>
<th>Arabic script</th>
<th>Pronunciation</th>
<th>Grammar</th>
<th>Definition</th>
<th>P.</th>
</tr>
</thead>
<tbody>
<tr>
<td>deffaqé</td>
<td>دفق</td>
<td>bi-fethi'd-dâl ve teşdidü'l-fâü'l-meftûha</td>
<td>Tef'il bábından fî'i mâzî-i ma'lûmdur.</td>
<td>Tedfik etti yani döktü manasına.</td>
<td>278</td>
</tr>
<tr>
<td>duffiqé</td>
<td>دفق</td>
<td>bi-zammi'd-dâl ve teşdidü'l-fâü'l-meksûra</td>
<td>Bâb-1 mezkürdan fî'i mâzî-i mechûldür</td>
<td>Tedfik olundu.</td>
<td>278</td>
</tr>
<tr>
<td>deffiq</td>
<td>دفق</td>
<td>bi-fethi'd-dâl</td>
<td>Bâb-1 mezkürdan emîrdir.</td>
<td>Sen tedfik eyle, dök demek olur.</td>
<td>278</td>
</tr>
</tbody>
</table>

Forth one is inflected interlingual homograph (see Table 7.). These headwords are same spelling but different pronunciation and different languages. Compiler gave three interlingual homographs. Languages are Arabic, Persian and Turkish. Meaning is totally different. Meaning of first word is “to memorialize”, meaning of second word is “other” and meaning of third word is “(s)he touches and (s)he reaches”. Compiler also gave an inflected verbs in them.
Table 7 Examples about inflected interlingual homograph

<table>
<thead>
<tr>
<th>Headword</th>
<th>Arabic script</th>
<th>Pronunciation</th>
<th>Etymology</th>
<th>Definition</th>
<th>P.</th>
</tr>
</thead>
<tbody>
<tr>
<td>dikr</td>
<td>دكر</td>
<td>bi-kesri'd-dâl</td>
<td>Arabî</td>
<td>Yâd edip anmak manasınıadr.</td>
<td>272</td>
</tr>
<tr>
<td>diğer</td>
<td>دكر</td>
<td>bi-kesri'd-dâl ve fethî'l-kefû'l-Fârisî</td>
<td>Farsî</td>
<td>Diğer gibi bu dahi gayrî ve yine manasına.</td>
<td>272</td>
</tr>
<tr>
<td>değer</td>
<td>دكر</td>
<td>bi-fethateyn kef-i Fârisîyle</td>
<td>Türkî</td>
<td>Dokunur ve ulaşır manasına ve eder manasına da gelir.</td>
<td>272</td>
</tr>
</tbody>
</table>

5. Conclusion

This paper describes the dictionary which named Eser-i Şevket, its structure and some examples about their macro and microstructures. These examples show that compiler of Eser-i Şevket wanted to do an extraordinary and unique dictionary in his own time because there is not any other dictionary like that. For these reasons this dictionary is very important for Turkish lexicography.

References

An Example of Interlinear (Subline) Dictionary Kitabü’l-müsellese

Tuğba POÇAN
Başkent University, Ankara, TURKEY
tugba_sarial_28@hotmail.com

Abstract

The religious factor was more effective in establishing and writing the dictionaries in the Middle Eastern culture at middle age period than the basic reasons such as trade, migration and religion. In Turkish culture, initial studies and works of the interlinear dictionaries were established in order to translate and spread the meaning of the sacred book (Qur'an) with Turkish provisions.

Kitabü’l-müsellese, written in interlinear technique, is distinguished from other religious works in this context in its period, as the subject was consisting of daily events and words. The Manuscript work was written in 9th century by Kutrub, scholar of linguist Sibeveyh. As of 16th century; it was translated into Turkish again by interlinear dictionary writing technique. The latest form of the dictionary is in the form Arabic into Turkish translation with the interlinear provisions, included Persian words enriching the latest form, turning out the work into Arabic-Persian-Turkish translation.

Müselles means triple. In this dictionary, new meanings arise when vowels of first letters of words changes. But this does not continue on regular basis. The interlinear translation is done by giving meaning of each word underneath of it. But this case is not also continued as regular rule. For some of the words, translations are given as sentence form.

In this paper; the writing technique of interlinear will be analyzed; by Kitabü’l-müsellese, the positive and negative aspects of this technique. The work was established in the form of translation from Arabic to Turkish and also to Persian. In this respect, the work has also acquired a multilingual feature. From this perspective, it is possible to detect words that have lost meaning. At the same time, there have been some negative reflections of interlinear on the work of the translating technique. Linguistic methods will be used for analysis.

Keywords: Glossary, interlinear, Arabic-Turkish, Kitabü’l-müsellese, multilingualism

1. Introduction

Kitabü’l-müsellese is an Arabic dictionary written by Kutrub in the 9th century. When it reached the 16th century, the work was transferred to the 16th century Turkcic language by means of interlinear technique. In this period, Turkish was under the influence of Arabic and Persian. One of the most intense Arabic influences in Turkish, which began to be influenced by
Arabic and Persian in acceptance of Islam, Turks who used many alphabets previously changed their alphabet with this effect and used Arabic alphabet for many years.

The Kitabü’l-müsellese was also transferred to the Turkic language in the 16th century, when the Turks used the Arabic alphabet. The work is a manuscript. Although there are many copies in Arabic, there are very few copies in Turkish by using interlinear technique. The copy we are dealing with is a complete numbered copy.


The main introduction to the work begins with the text above. In order to understand the Qur’an with the influence of Islam in the Karahanlı Turkish period, it is quite often confronted in the translation of the text to the dialectic translation technique. But the Book-Kitabü’l-müsellese is a non-religious work. This is very important in this respect. Besmele, the name of the work and the article is given, then it is seen that the translation has been passed. The first word is al-ğamru. While the Arabic dictionary was being prepared, this work also became a first product. The first word of the word was moved to the conquest, with the fetha, kesre and the zammey and three new meanings of meaning were provided from the same word. The Kitabü’l-müsellese is the first to be revealed in this sense. For this reason, it is named as three, the split, the triple meaning muselles.
el-ğamru āb-ı bisyār çok şu / so much water
el-ğamru kīne gine kīne / anew grudge
el-ğamru ġāfil ve ebleh gine ġāfil dāhu ‘akulsız/ anew unwary and foolish
el-łacru dāmen pīr amen gönelle etegi / shirttail
el-licru ġurad ‘akil / mind
el-łicru nām-ı merdiyyet ādem adı / human name

In addition to the Arabic and Turkish dialects, some of the points were also given in Persian. In this sense, the dictionary also has a multilingual feature. In this sense Arabic-Persian and Turkish are separated from each other in order to distinguish between Arabic words normal, Persian words italic, and Turkish words written in black are written in bold. If we look at the work in this sense, it is observed that the work actually has a triple feature in terms of both movement and language. But Persian provisions do not match regularly. At some point, the antagonist emerges. It is possible for the author to consider this as a step in the language teaching process. For those who can not fully understand the antecedents of translation, Persian words also have an auxiliary role.

The system in the original Arabic section of the text is the new triple word syllables created by the different movement of the first syllable of the word invented by Kutrub. This order was given in the 9th century. As far as the 16th century is concerned, it has been given the ability to translate underline.

2. İnterlinear Dictionary Writing Technique

The first translation activities in Turks are seen in Uighurs who have changed religions. Much of the Uyghur translation corpus is known to be made up of religious works. The religion of the contents of the works was born from the necessity of fulfilling the requirements of religion and fulfilling the necessities of the community that a religious-changing society is now a member. In Muslim Turks, the first known translations are translations of the Holy Qur'an from the same necessity. The first translations of the Qur'an to the Turkic language were made in Eastern Turkism and probably in the 13th - 14th centuries, just under the Arabic original text.

Throughout history, Turks who have embraced the beliefs of Heavenly God, Manihaism, Buddhism, Christianity and Judaism have gradually expressed this religious interest with the spread of Islam in Turkestan region and after Karahanlı rulers adopted Islam as official religion, members of Islam religion have begun to belong to Islam. The new acquaintances needed the interpretation of the Holy Qur'an to fulfill the requirements of this religion and to understand Islam. Translation activities started in the period of Karahanlılar continued to the Turks who settled in Anatolia and today various translations of Qur'an which have been preserved in many libraries have been reached.

It was founded on the basis of recognizing the Qur'an in the eastern first translations. The Qur'an is based on the principle of one-sided technical Arabic words and phrases. This creates a very difficult situation in terms of the structure of the Turkic people. Because the syntax of Turkic and Arabic, Persian syntax is different from each other. Turkish syntax is based on subject, predicate, object, which is based on the subject, object, predicate, while Arabic and
Persian are Indo-European language features. At this point in the point of interpersonal translation, the historical point of Turkic syntax is born.

Kitabü’l-müşellese is remarkable in that it is a non-religious text in this sense. The use of sentence structures and the use of other grammatical structures in the 16th century Turkish.

2.1 Word by Word Translation

It is the method that is often applied in the Western Quran translations. As opposed to the word of the line, the word is the clearest, most kein meaning is chosen. In this respect, individual responses were important. It is the most commonly used response method in the booklet.

2.2 Sentence Form Translation

One of the remarkable points is that there is a lot of confusion in the sentence structure in the booklet. The Arabic word provides much more comprehensive mental settlement. The correspondence of the word is exemplified by being widely explained. This allows the dictionary user to be assimilated by the dictionary user.

It is noteworthy that this sentence is quite functional if the sentence form in the booklet is handled. The existence of canonical and transposed cues draws attention. Especially since the use of the overturned sentence structure seems to be a characteristic of a dialect of Indian-European language family at first glance, it may be more correct to link the cause of it to another place since it is the only Arabic word.

For example, the definition of the Kaf mountain, which is a mythological mountain in the opposite direction, is made. When you look at the sentence form, you come up with a structure that is the basic syntax and the side syntax is at the end.
In another example, the same structure emerges again. The samples can be reproduced in this way.

2.3 Idioms

Idioms are also one of the elements that are referenced without responding to the underwriting. A language is the last and most difficult step in teaching. Molded expressions can sometimes burn ratios. It has been observed that idiomatic expressions are also used in some points in the subterranean technique.

There are many more idiomatic examples than the examples given.

2.4 Synonym

The synonyms are given in the general sense of the word 'gine' as a sub-word.

2.5 Strengthening

Reinforcement is a frequently used structure in Kitab al-musellese. It is made especially using the word 'katı/ strict'”. Apart from these, there are other structures for reinforcement.
2.6 Opposition

Contradiction is given in some conjunctions or in sequential order of words.

fa’ad yitirmek hem yitik bulmak / lose anf find
isrâh gizlemek hem âşker etmek / hide and reveal
sehre tölmak hem boşalmak /fulling and ejaculate
serrâ’e baylık hem ucuzlık / wealth and sale
mülevvâtu gice gündüz / night and day

2.7 Hendiadyoins

The duplicates provide a harmony of the text. It is quite functional in this sense.

ağdeyn gice gündüz/ night and day / Area of use is still available in Türkic Turkish
äkbele ögen ögen gitti/ going back/ It has fallen from use.
‘as’as önün önün gitti ve kaçan kaçan gitti/ Went ahead by ahead / Though the usage area of the building has decreased, Turkey is encountered in Turkish dialects.

3. Positive Aspects of İnterlinear Dictionary Writing Techniques

Nowadays, the method of using interlinear technique is almost unused. The translation of the texts defines independent dictionaries. But there are some positive aspects of roundtrip translation. At the beginning of these, just under the meaning of the response to the meaning of the flow of the impression of the word is not perceptible at that moment allows. It saves time without losing time without intervention. It keeps it interesting. In this way, the text six translation offers the convenience of the opposite. A rapid flow is achieved without breaking the tube.

4. Negative Aspects of İnterlinear Dictionary Writing Techniques

The use of the interlinear technique arises at this point as some textual difficulties arise because the text we are dealing with here is a manuscript. The most important of these is the length of the gap to be given to the line. In Kitabû’l-mûsellese, many sentences in Arabic are given in the form of long sentences in the form of sentences. This caused some shortcomings due to the space limitation between the two lines. For example, when a word is spoken, the extension of the words below other words is an example of this because of the limited space.
At some points they are written on top of the page, between them, or on the edges to save some space. This again caused some difficulties in reading.
In addition to this, the interlinear translations depend on the knowledge of the translator. However, if there are provisions from the dictionaries, wealth can be born in terms of meaning. The spoken word can be varied. By trying to be identified from different dictionaries, more information about wordness can be obtained.

5. Conclusion

The book is a handwritten manuscript with a very rich source in terms of the writing technique. The word is presented in the form of sentence structures and long word groups at the same time, which are not given as words only. This also adds value in terms of syntax. But the fact that this technique contains long words has some disadvantages. At the beginning of these is the lack of space. So the author tried to write some expressions in some empty spaces while translating. Sometimes the promise has gone on and has come under another word. In this case, there was a disturbance in the separation of words. Some parts of it have very difficulty in some points that have reduced the legibility rate.

At the same time, the translator has already pushed the translator down a bit and has already caused the conversation to succumb to the existing line. However, there is no clear criterion to test the vocabulary of the person who gives meaning to the words.

Apart from that, if we take it from a positive point of view; such a translation provides convenience to the reader. It provides a very efficient use of time. The reader locks the text to prevent it from deviating to other parties. So the flow moves faster and easier. Words can be perceived more clearly.
References


Çelebi, M (1981). Kutrub, Hayattı, Eserleri ve Kitâb Al-azminâ. Unpublished Associate Professor Thesis. Atatürk University Faculty of Islamic Sciences, Erzurum


Kutrub b. Ahmed el-Basrî (1556), Kitabü’l-Müsellese


Steingass, F. J. A (1892) Comprehensive Persian-English dictionary, including the Arabic words and phrases to be met with in Persian literature. London: Routledge & K. Paul


On the Presentation of Semantic Prosody in Advanced English-Chinese Learner’s Dictionaries

Lu Bai
Nanjing University
bailunju@126.com

Semantic prosody is one of the contributions of corpus linguistics. Explicit knowledge about it can help learners produce expressions natural to native speaker of a language. Studies have found that the language performance of Chinese advanced English learners reflects an insufficient awareness towards the semantic prosody information of an English word. As an indispensable tool for learners to master a foreign language, English learner’s dictionaries are supposed to attach importance to semantic prosody and provide adequate information in the entry so as to enhance the vocabulary learning outcome. The treatment of semantic prosody information in monolingual learner’s dictionaries is satisfactory compared with that of English-Chinese learner’s dictionaries, whose conventional English-Chinese defining method failed to reveal the subtle difference between an English word and its Chinese “equivalent(s)” in the light of semantic prosody. This deficiency is likely to mislead and cause unnecessary confusion for the user. Taking several verbs and phrasal verbs as examples, the present study attempts to construct a presentation system to reveal semantic prosody information of English words in advanced English-Chinese learner’s dictionaries. It is hoped that this study can provide some theoretical and practical implications for learner lexicography.

Keywords: Semantic prosody, advanced English-Chinese learner’s dictionary, presentation

References

于屏方 杜家利. 汉、英词典对义位语义韵律的处理[J]. 辞书研究, 2010(5):31-41
The Effectiveness of a Bilingualized Dictionary in Helping Learners Determine Article Use

Alice Yin Wa Chan
Department of English, City University of Hong Kong
enalice@cityu.edu.hk

This paper reports on the results of a study which investigated the use of a English-Chinese bilingualized dictionary in the determination of English article use by advanced Hong Kong Cantonese ESL learners. A homogenous group of 30 English majors in a local university participated in the study, which consisted of an article selection task with the use of a bilingualized dictionary. A total of seven nouns each used in three different contexts were included. In at least one of the sentence contexts the noun was used as an uncountable noun, and one in which the noun was used as a singular noun, a plural noun, or a countable noun. The participants had to determine which article to use for the target nouns in each sentence context. They were also required to do a questionnaire after completing article selection for each noun. The aim of the questionnaire was to examine the problems that the participants encountered, the examples/definitions they relied on in making a decision, as well the strategies they used when they were in doubt. The results of the study show that the provision of dictionary examples showing different choices of articles with the same noun in different contexts causes confusion, and learners often use inappropriate structures as models for article selection. Chinese translations are sometimes sources of learner problems. It is suggested that learners be alerted to the subtle differences between the different uses of a noun, and that lexicographers be more learner-centered in their provision of dictionary information.

Keywords: article use, bilingualized dictionary, learner problems
A Discourse Approach to Critical Lexicography: Methods, Principles and Applications

Wenge Chen
School of International Languages, Xiamen University of Technology
chenwenge@xmut.edu.cn

Since the critical turn in the late 1990s and early 2000s, applied linguistics research has witnessed a burgeoning body of literature on, but critical lexicographic studies are obviously a neglected issue.

Dictionaries, with their descriptive, didactic and ideological functions, are extremely important social discourses. They enjoy an almost unquestioned authority in public, academic and legal fields. Researchers and judges frequently cite dictionaries to underline what a word or concept means or how it is used. Dictionaries are so useful that people are reluctant to criticize them. From a critical perspective, however, the view of dictionaries as authority in words’ meaning and usage has to be challenged. Dictionaries are in fact closely related to power and politics and are important carriers of ideology. Though linguists and lexicographers have to date contributed a lot to the development of lexicographical theories, most tend to see the dictionary as a tool or an end-product. The dictionary as a social phenomenon and as a discursive phenomenon has received undue attention. Though there has been a wealth of literature of lexicographical ideology, these studies are mostly sociological. There is a paucity of literature that sees dictionaries as discourse formation in their own right.

This paper argues that lexicographical discourse, like news discourse and other discourse in general, is never an instrument neutrally and objectively conveying information, but is a site of ideological struggle. To reveal how meaning is negotiated in the dictionary and how these struggles impact on the compilation of a dictionary, the present paper proposes critical lexicographical studies from the perspective of discourse (Critical Lexicographical Discourse Studies or CLDS for short). It first reviews the varied views of the dictionary. Then it deals with the theoretical foundations of the Critical Lexicographical Discourse Studies, defining the several related concepts of CLDS, namely critical, power, text, ideology, and discourse. It then elaborates the principles of Critical Lexicographical Discourse Studies and proposes an analytic model drawing inspiration from Hornscheidt’s (2008) research agenda for a critical research agenda of lexicographical studies. Finally, in four separate case studies, it briefly exemplifies how Critical Lexicographical Discourse Studies can be applied and conducted. The paper also sums up the strengths and limitations of the approach and provides some suggestions for future critical lexicographical studies from a discourse approach.
Coding Transitivity Construction in English Monolingual Dictionaries: the Syntax-Semantics Perspective

Ziyue Chen
Guangdong University of Foreign Studies
262651327@qq.com

Embarking on the conflicting perspectives concerning the transitivity coding practice in English monolingual dictionaries (MLDs), this article illustrates the transitivity coding system Subject-Predicate-Object-Adjunct-Oblique object (SPOCAOBL) at the level of the sentential construction. Through observing the concordances of twenty six verbs on the transitivity mistakes made by Chinese intermediate and advanced EFL learners in Chinese Learner English Corpus, it was found that Chinese EFL learners’ needs for transitivity construction are not satisfactorily met by the current online ‘Big Five’ MLDs. Based on the usage-based evidence in British National Corpus (BNC), we selected “fear”, ”develop”, ”take” through stratified sampling of the twenty six verbs ranked in terms of the frequency list of Corpus of Contemporary American English (COCA) and classified the prototypical and extended event structures with reference to their distinctive meanings in the linking verb construction, transitive construction, ditransitive construction and intransitive construction. Adopting the collostruction analysis, we then identified the meanings of participants in each slot of argument structure constructions. As a direct correspondence between constructions with a highly abstract meaning and the meaning of a verb or verb class in a given syntactic pattern could not be expected, the coding of an argument structure construction in an abstract use was separately dealt with therewith. Finally, we propose a “three-tier” model to code situation types of transitivity constructions. The main goals of this study are not only to maximize the usefulness of coding system for Chinese EFL learners, but also to provide some insights into reasonable and comprehensive representation as regards the coding of transitive use.

Keywords: transitivity coding system, SPOCAOBL, sentential construction, collostruction analysis

References


The Treatment of Figurative Collocations in Paper and Electronic Learner’s Dictionaries

Yuanjun Dai
Xinghai Conservatory of Music
douglasengine@hotmail.com

Figurative collocations have figurative meanings which are not calculated from their co-occurring words. Presenting figurative collocation meanings through some form of linguistic analysis and analogy to enhance learners’ metaphor awareness may facilitate retention of unknown figurative language. The paper examines the treatment of figurative collocations in a range of five leading learners’ dictionaries, including the COBUILD, the Longman Dictionary of Contemporary English, the Oxford Advanced Learner’s Dictionary, Macmillan English Dictionaries and the Cambridge International Dictionary of English to address the following questions:

1. How do learners’ dictionaries treat figurative collocations? Is there any difference between paper and electronic version of the same learner’s dictionaries in the treatment of figurative collocations?

2. Do these dictionaries help raise metaphorical awareness of users?

3. How is conventionality of metaphorical expressions dealt with?

The examination shows that the leading learner’s dictionaries, in lack of cognitive metaphor awareness, are insufficient in the presentation of metaphorical information. Learner’s dictionaries should go beyond overt collocational information and delve into the underlying cognitive metaphorical motivation. One exception, the Macmillan English Dictionary, takes to the other extreme. By drawing on the cognitive linguistic metaphor studies, it relies too heavily on “generalization commitment” and relegates linguistic facts to a secondary role, where the conventionality of metaphorical expressions is not given its due attention.

In light of these findings, it is important and necessary to incorporate conceptual metaphor information into learner’s dictionaries and give due attention to conventionality of figurative collocations to facilitate Chinese EFL learners’ figurative collocation learning and production.

Keywords: figurative collocation, learner’s dictionaries, metaphor, conventionality, metaphor awareness
This paper aims at describing the process of KBBI (Indonesian Comprehensive Dictionary) in utilizing the wealth of local languages to enrich Indonesian vocabulary by KBBI online application. KBBI is a descriptive dictionary that was firstly published in 1988 by Ministry of Education and Culture. Although KBBI is a descriptive dictionary, people consider it a standard reference since it is an official and government’s product. KBBI is not only a dictionary which records all facts of language in society, but it also unites at least 646 local languages in Indonesia. Local languages play role in enriching Indonesian vocabulary. Local languages provides new concepts that are not available in Indonesian. In addition, they are also used as equivalents for modern terms in particular fields. For example, the Javanese words unduh and unggah are used for download and upload. There are five criteria considered in determining whether a word can be accepted as a member of Indonesian language. They are a) unique in concept; b) subject to Indonesian grammar and spelling, c) positive in connotation, d) euphonic, and e) highly-used in frequency. The standardization process is held through Terms Commission Meeting, convened twice a year, consisting experts from various disciplines, practitioners, user, and language advisor. In order to obtain users’ participation in enriching the Indonesian vocabulary, KBBI facilitates them an online application called KBBI Daring. KBBI Daring was developed during the making of KBBI Fifth Edition in 2016. By this application, users are able to propose or change entries, senses, or illustration. This crowdsourcing application showed a positive impact in enriching Indonesian language.

**Keyword:** local language, crowdsourcing, online KBBI dictionary, vocabulary enrichment

**References**


University medical students in China form a large group of ESP learners. A survey is conducted among 72 undergraduate medical students at Fudan University to examine their use of dictionaries for ESP learning purposes. Meanwhile the survey aims to find out user expectations and potential responses to HTF, an online bilingualized English-Chinese Health TermFinder. The findings show that the dominant majority of the students never even used any specialized medical dictionary (monolingual or bilingual), and unlike English majors at the same university (Ding, 2015), very few of them use more than one dictionary for their English learning. An analysis of their answers to the open questions leads to the conclusion that the Chinese medical students have rather unrealistic expectations for dictionaries in their ESP learning, though they hardly know how to use dictionaries, whether general or specialized ones. In spite of the prevailing lack of knowledge about dictionary use, however, the students demonstrated genuine eagerness to try the HTF, a promising sign that with proper guidance they may form a potential user group to be benefited from such a specialized medical termfinder in the near future.

References


---

1 HTF is a lexicographical project currently being developed at the linguistics department at Macquarie University, Sydney. Meanwhile, College of Foreign Languages and Literature at Fudan University, Shanghai is involved in the bilingualization of HTF into Chinese. Students from the Medical School of Fudan University are envisioned as the test users of the bilingualized HTF, hence, also subjects of the current study.
For EFL learners, knowing what a word means is easy, but figuring out how a word is used is demanding. Complex-Transitive Argument Structures (CTASs) are a case in point. They are conventionalized but unpredictable, concerning item-specific knowledge with respect to the co-occurrence of one word with a particular grammatical construction. Recent research on argument structures shows that it is implausible to focus on the semantics of verbs alone to explain the syntactic realization of verbs and that both verbs and constructions contribute meanings and arguments to the whole construction (Goldberg, 1995; Hoffmann & Trousdale, 2013; Perek, 2015). As one of the most useful and essential tools for foreign language learning, English learner’s dictionaries (ELDs) are supposed to take into consideration the encoding needs of EFL learners. In this study, we conducted a detailed analysis of CTASs and examined how they are treated in the six leading ELDs, i.e. the ‘Big Six’. It was found that CTASs are not sufficiently included in ELDs. Existing ELDs tend to focus mainly on the semantics of verbs, unaware of the possible roles of constructions. In addition, for some verb-specific constructions, the words used in each slot may demonstrate idiomatic preferences or restrictions. For example, in the OPINION sense of hold, words that can be used in the adj slot in the construction HOLD n adj are restricted to accountable, liable and responsible (Francis, 1996, p. 280). However, some ELDs like OALD8 fail to fully convey such restrictions to users, only taking responsible into account. We discussed similar findings in existing ELDs and provided recommendations on how the entries for CTASs may be improved in future ELDs.

**Keywords:** complex-transitive argument structures, usage-based, constructions, English learner’s dictionaries

**References**


Confucius once said, “In all affairs, if you plan ahead you can be successful, and if you don't plan ahead, you will fail.” From ancient times to the present day, dictionary-making has been related to the concept "lexicographical planning" in China. Yongle Dadian, commissioned by the Yongle Emperor of the Ming Dynasty, and Kangxi Dictionary, ordered by the Kangxi Emperor of the Manchu Qing Dynasty, are two typical examples in ancient China. After the founding of P. R. China, three lexicographical plans at the national level have so far been issued by the government, i.e. the 1st National Lexicographical Plan (1975-1985), the 2nd National Lexicographical Plan (1988-2000) and the 3rd National Lexicographical Plan (2013-2025). A few scholars have mentioned the issue of lexicographical planning from the perspective of lexicographic history. But the relevant discussions have so far been rather sporadic and unsystematic. The article will first discuss the concept "lexicographical planning" in the context of modern China. Just like language planning can be defined from different perspectives, the definition of lexicographical planning may vary from one scholar to another. It will give a working definition of "lexicographical planning" based on the question: “Who plans what for whom and how?”. It will then trace the development of three National Lexicographical Plans since 1975, focusing on the discussions of the positive and passive roles lexicographical planning plays in the compilation and publishing of dictionaries. Based upon these considerations, reflections on the significance of and new approaches to national lexicographical planning in the Internet Era will be presented.

Keywords: lexicographical planning, positive role, passive role, reflections, the Internet Era

References

FSPOST: A Part-of-Speech Tagger for Filipino

Matthew Phillip Go¹, Nicco Nocon², Nathaniel Oco¹ & Rachel Edita Roxas¹
¹National University, Manila, Philippines | ²De la Salle University, Manila, Philippines
matthew.go123@gmail.com | noconoccin@gmail.com
nathanoco@yahoo.com | rachel_roxas2001@yahoo.com

This paper discusses an early stage of research towards development of a Filipino part-of-speech (POS) tagger using the Stanford POS Tagger. The Stanford POS Tagger is considered as one of the POS taggers achieving the highest accuracy on covered languages such as English, Chinese, Arabic and German (Toutanova et.al, 2003). On the other hand, the Filipino language is very limited in terms of the number, and accuracies of POS taggers developed for it. Stanford POS Tagger’s flexibility to be trained on different languages presents a significant opportunity for Filipino to have a high accuracy POS tagger which is a very fundamental tool towards natural language processing (NLP) researches as POS tags is considered a basic feature to be used to relate words together and making sense out of them (ex. naliligo ‘taking a bath’ and tumatayo ‘standing up’ are both imperfective verbs). Sample NLP researches that uses POS tags are: information extraction, knowledge base building for AI, grammar checker, sentiment analysis, and others. This research examines how a machine learning based POS tagger can be utilized for the Filipino language having its own unique linguistic phenomena: high morphology, free-word order structure, borrowed and derived words (Go & Borra, 2016). The POS tagger uses 7,384 complex manually-tagged Filipino sentences consisted of 183,533 tokens (25,674 are unique) that are translated sentences from English Wikipedia for its training and testing. The Filipino Stanford POS Tagger achieves 93.9% accuracy. This is 9.15% higher than the state-of-the-art Filipino POS tagger – SMTPOST (Nocon & Borra, 2016).

Keywords: Part-of-speech Tagger, Natural Language Processing, Filipino, Tagger Performance

References


In this paper, we present an experiment using topic modeling as an approach to assist in analyzing opinions and ideas related to the disaster risk reduction strategies. These opinions and ideas were gathered from local communities that have been victims of Philippine disasters. We used paper and pen surveys and a participatory toolkit called “Malasakit,” which is a Filipino term for “sincere care”. Malasakit is a customizable multilingual participatory assessment platform that collects quantitative assessment, qualitative feedback, and enables peer-to-peer collaborative evaluation of qualitative responses to reveal novel insights (see http://opinion.berkeley.edu/pcai/v21/). Malasakit extends upon our previous work developing the Collaborative Assessment and Feedback Engine (CAFE), deployed to crowdsource priority policy issues in California and Mexico, and to enable participatory evaluation of the effectiveness of development programs in rural Uganda (see cafesystem.org). A total of 124 responses to the question “How could your barangay help you better prepare for a disaster?” were collected. In order to better understand key topics emerging from participants, five topic models, each with twenty words, were generated and labeled. Analyses show that respondents from these disaster-stricken communities focus their suggestions on community-wide communication during evacuation (sample keywords: communication, system, community, flooding, evacuation, items); early warning, equipment, and supplies in preparation for rivers overflowing (sample keywords: flood, warning, equipment, evacuate, sound, food); allotted funds and equipment for immediate community-wide response and recovery (sample keywords: cleanliness, clear, family, unity, fund, budget, boat); providing relief to those displaced by flooding (sample keywords: relief, assistance, help, aid, service, river, flood, high); and responsibility of community officials (sample keywords: seminar, cleaning, repair, goods, distribution, required, proper). Challenges encountered when labeling the topic models include words with different senses (e.g., give - distribute vs. conduct); spelling variations; and words with various inflections. Future direction include presenting the results to community officials towards policy making, and developing a lexicon for the disaster domain. The study can also be extended to related domains, such as maternal health care, civic engagement, education, among others.

**Keywords:** Topic Modeling, eParticipation, Disaster Preparedness
A Review of Definition Research in Chinese Dictionaries for Foreign Learners (CDFFLs)

Qiping Guo
School of Foreign Languages & Cultures, Southwest University of Science and Technology
guoqiping1@hotmail.com

This paper reviews the definition research in CDFFLs in the past 30 years. At macro-level, some scholars have advanced macro definition models. Yihua Zhang (2008, 2010, 2011a, b) proposes the multi-dimensional definition model based on user’s perspective and the most recent developments in linguistic theories. Xiaoling Weng (2011) researches the meta-language-based definition model using The Commercial Press Learner’s Dictionary of Contemporary Chinese as a case study. Yongan Gao (2012) suggests that the definitions for a ploysemy in CDFFLs should take the internal relationship of word meaning system into consideration as well as the explanations of its literal meaning. Following the example of Collins Cobuild Advanced Learner’s English Dictionary, Yumin Xu (2008) adopts the full-sentence definition model for most Chinese words in Learner’s Dictionary of Contemporary Chinese (Elementary Level). Xiaoqi Li (2007) points out that the goal of definitions in CDFFLs is to aid dictionary users in understanding the meanings of Chinese words, mastering their usage rules and using the definienda properly in specific contexts.

At micro-level, some scholars have presented word-class categorization definition models. Yongan Gao (2007) proposes a definition model for segregatory words based on their usage framework. Ying Zhao (2013, 2015) suggests one for Chinese specific words from the perspectives of the hierarchical structure of the internal morphemes of basic meanings and frame semantics. Hui Wang(2015) puts forward a definition model that requires paying particular attention to learners’ errors, applying a variety of definition methods, avoiding synonym definition, choosing typical exemplification and setting up words discrimination column. Lastly, the shortcomings and deficiencies in the extant definition research in CDFFLs have been explicitly pointed out.

Keywords: CDFFLs, definition research, review

References


Computational Representation of a Model of Entries in a Business English Learners’ Dictionary

Jianing He & Mosi He
School of English for International Business, Guangdong University of Foreign Studies
M.A. of Division of Psychology & Language Sciences, University College London
jianing.he@gdufs.edu.cn | mosi.he.15@ucl.ac.uk

A number of lexical databases for English (e.g. WordNet, VerbNet, DANTE) have been built which contain rich information of the general English lexicon. However, there is not yet a lexical database for business English. In this paper, we discuss the methods, techniques and tools of computational representation of a general model of typical entries in a business English learners’ dictionary (BELD), which is an encyclopedic learners’ dictionary. The entries in such a dictionary should contain the hierarchical information of both the English language and business knowledge. The objective of this dictionary is to serve the special needs of learning business English. Business English major is a new one in China and a BELD should be compiled to meet the needs of students of business English major for using the dictionary. The target BELD in this study is an English-English-Chinese bilingualised dictionary with students of business English major as potential users. Some samples of dictionary entries are illustrated in XML (eXtensible Markup Language) with an XML editor (e.g. XML Copy Editor) to show how the general model is compatible with specific entries with different types of headwords, and how this model can represent an existing dictionary entry or a newly-constructed entry. This study is helpful for building an electronic business English lexical database which in turn is the basis for compiling an electronic BELD. For further research, programming languages (e.g. PROLOG) is to be employed to manage the above-mentioned database, generating automatically different sub-types of a BELD with the desired information modules in accordance with the specific needs of the users for decoding and encoding English for business purposes.

Keywords: business English, lexical database, BELD, dictionary entry model, XML

References
The aim of this study is to examine how nineteenth-century dictionary-makers dealt with cultural aspects in Chinese bilingual dictionaries. Many bilingual dictionaries from that era were compiled by and for Western scholars and missionaries, who during their sojourn in China, not only learned the language but were also introduced to Chinese culture and customs. These are reflected in the dictionaries they compiled.

The starting point for this study is the Dutch-Chinese Dictionary with Transcription of Chinese Characters in Zhangzhou dialect compiled by Dutch sinologist Gustaaf Schlegel (1840-1903). From the many comments, explanations and example sentences, it is clear that he expected his users to know little about China. For terms which Schlegel perceived no direct Chinese equivalents, he explains the cultural differences at work in the interpretation of the terms in China and the West. To show users how words are used in Chinese, he quotes liberally from Chinese source texts, such as the Book of Odes, The Analects, Wonders Old and New, and Kangxi Dictionary.

Here I will examine entries about specific cultural aspects which Schlegel explains. I will compare these with entries in earlier and later bilingual dictionaries, including those compiled by Robert Morrison (1782-1834), Walter Henry Medhurst (1796-1857), Paul Hubert Perny (1818-1907) and Wilhelm Lobscheid (1822-1893). Taking into account the background of each compiler, I will analyze the differences and similarities of the entries and determine the interpretation of cultural aspects for each of them. My findings will give an insight into the interpretation of cultural aspects as part of the complex process of compiling bilingual dictionaries.

Keywords: bilingual dictionaries, China, translation, culture
Definitions of Chinese four-character idioms in Chinese-French dictionaries are a challenging task for lexicographers and raise a big issue for current researches. Many relevant researches probing into the problems of definition for Chinese four-character idioms are based on several randomly selected examples, while few studies investigate a multilingual corpus to testify the quality of definitions for four-character Chinese idioms and then to give suggestions on their improvement. This article investigates in two Chinese-French dictionaries definitions for 100 idioms, whose first character is 大(da). The results show that some of these definitions are not usage-based by offering inaccurate, subjective, low contextual information. To solve these problems, this article suggests improved definitions of four-character Chinese idioms by investigating and analyzing their real use in the Chinese corpus, their equivalent translation(idioms or expressions) in the French corpus of the corpus tool-Word Sketch Engine. It is proved that based on corpus, the modified definitions could better cater for the L2 learners’ needs because they provide more objective, accurate, highly contextual information for Chinese four-character idioms.

**Keywords:** corpus-based investigation, definition, four-character idioms, Chinese-French dictionaries, improvement
Issues in Bilingual Lexicography with Pashto and English as Pairs

Zafar Iqbal & Saida Jamal
Imperial College of Business Studies, Lahore, Pakistan
drizqbl@gmail.com

Lexicography with all its taxonomic variations in general and bilingual lexicography in particular has been the outcome of some substantive, pragmatic desiderata in real life situations in the respective contemporary times. This is exactly the case of Pashto-English, English-Pashto lexicography that witnessed its genesis in the eighteenth century in the Indo-Pak subcontinent. The aim of this paper is to investigate the original rationale behind the production of such dictionaries and to critically evaluate their development ever since their appearance in reality and the pattern of their progress they pursued over the years in different sets of socio-cultural background. After highlighting the actual challenges faced by this kind of lexicography, the study will propose the modern principles of bilingual lexicography which if followed in the case of Pashto-English lexicography can potentially ensure a better generation of dictionaries for a new kind of dictionary users. As many as ten dictionaries as available in the Pashto Academy in Peshawar University, KPK have been chosen for a comprehensive analysis of their mega-, macro-, and microstructures.

Keywords: lexicography, Pashto-English bilingual lexicography, modern principle of bilingual lexicography, bilingual lexicography, Pashto English as pairs, megastructure, macrostructure, microstructure, dictionary users, taxonomic variations, Pashto Academy Peshawar, Kpk
A thorough grasp and analysis of lexicographic literature can be highly helpful to uncover the “history, present and mystery” of lexicographic studies. However, most current studies on this topic only draw statistical summaries of literature works but fail to respond to the advanced methodology in this continually expanding digital universe, and therefore, make their results less convincing and updated. The visualizing tool CiteSpace developed by Prof. Chen Chaomei uses many innovative techniques and algorithms for conducting systematic and visualized surveys of literature works. It has established itself as an excellent tool for researchers to keep up with and locate hot issues and trends in the dissemination and spread of scientific information. Based on such a powerful knowledge mapping software, the present paper tries to offer dynamic visual analysis of the collected data resources from International Journal of Lexicography, bearing the intention to reflect the most important issues and trends in the field of lexicography over the past few decades. To be detailed, the study conducts quantitative and qualitative analysis from several dimensions, including the co-occurrence of keywords and authors of literature data, the visual mapping of scientific knowledge in lexicographic discipline over time, the researching hotspot, front, strength, representativeness and key nodes about literature in this field. Finally, the highest keyword-time trends and the strongest citation bursts are also drawn out. With the help of these visual data, the paper attempts to conclude the developing trends of lexicographic research overall and evaluate the limitations of lexicographic studies in the mainland of China.

**Keywords:** CiteSpace, knowledge map, lexicography, International Journal of Lexicography

**References**


Subject Fields in EFL Dictionaries and ESL Textbooks

Takashi Kanazashi
College of Commerce, Nihon University, Japan
fredericfchopin@gmail.com

This paper aims to improve the labelling systems in EFL dictionaries. The EFL dictionaries put subject field labels to some subject-specific headwords, but the way the labels are used are a matter of controversy, because inconsistency has been found within a dictionary as well as between dictionaries (cf. Hartmann (ed.) 1996, Kanazashi et al. 2009).

The selection of subject field labels in alphabetical dictionaries tends to be different from the division of chapters and sections in thematic dictionaries like Gakken editorial team (ed.) (2002) or a number of ESL textbooks. Gakken’s dictionary, Japan As It Is: A Bilingual Guide, 4th Edition, presents a good example of thematic arrangement in its 5 parts, each divided into 12 to 41 chapters.

Today, many ESL textbooks, particularly ones intended for Japanese university students, are divided into chapters according to the themes. However, not all ESL textbooks are thematically arranged in the traditional sense. Matsumoto (2017) can be regarded as a strategy learning textbook in which contents are arranged according to the question patterns or strategies to answer the questions in the TOEIC® test rather than the subject field.

Taking the latter two (thematic dictionaries and textbooks) into consideration, this paper will discuss how the EFL alphabetical dictionaries could devise a better way of handling subject field labels. In Kanazashi et al. (2009), only Longman Advanced American Dictionary, 2nd Edition, seems to be criticized for lack of consistency, but in actual fact, other EFL dictionaries published by Longman, OUP, Macmillan, and Merriam-Webster seem to have the same problem.

Keywords: EFL dictionary, ESL textbook, subject field, thematic arrangement

References
A Study on the Method to Select Headwords for the Development of a L2 Learner of Korean Synonym Dictionary: Based on the Errors of the Learner’s Corpus

Jeong-hyun Kim
Yonsei University
jhkim234@gmail.com

The purpose of this study is to re-establish the concept of synonyms for students learning the language as a foreign language. As such, the study is based on the practical viewpoint that in order to develop a synonym dictionary, not only synonym groups from a theoretical perspective but also from a practical perspective must be reviewed, with a focus on frequently made errors by learners. Moreover, by comparing and reviewing the meaning error list for synonyms in the learner’s corpus and the existing list in the synonym dictionary for Korean learners, this study seeks to select a new list of synonyms.

For the study, learner’s corpus of 279,389 phrases were used. They were categorized by different meaning error type as shown in the text. A list of 253 corresponding pairs that can be replaced by the meaning errors of the learners was compared with the list in the synonym dictionary for Korean learners. This showed that only 165 pairs corresponded. This indicates that the difficulties of vocabulary learning that actual learners experience cannot be overcome by existing synonym dictionaries.

The error in meaning were as follows: 1) Errors in vocabulary of comprehensively similar meaning that indicate the same concept or same object; 2) Hypernym vs. hyponym; 3) Proper words that encompass a broad meaning and words based on Chinese characters that have a more specific meaning.

Error number 1 corresponds with the existing list, while errors 2 and 3 are those where words are not listed in the synonym dictionary but students frequently make mistakes about. These words need to be added as headwords in the synonym dictionary.

The result of study are expected to offer practical help to language learning, by using learner’s corpus unlike existing studies to select head words when developing a synonym dictionary for learners.

Keywords: Synonym, Dictionary, Headwords, Learner’s Corpus, Error

References

Japanese Words in the Dictionarium Annamiticum Lusitanum et Latinum (1651)

Emi KISHIMOTO
Osaka University
emik@aioros.ocn.ne.jp

The Dictionarium Annamiticum Lusitanum et Latinum (1651), the Annamese-Portuguese-Latin dictionary, was compiled by Alexandre de Rhodes (1591–1660), a French Jesuit, who worked in Vietnam for approximately 20 years. He is famous for his influence on the modern Romanization of Vietnamese. However, the vocabulary of the Dictionarium has not been studied in detail.

This paper focuses on several Japanese words found in the Portuguese and Latin translations of Rhodes’s dictionary, such as Catana (Japanese sword), Bonzo (Buddhist monk), and Xiru (Japanese soup). The preface states that Latin words were added using two previous Annamese dictionaries—Annamese-Portuguese by Gaspar do Amaral, S. J. and Portuguese-Annamese by António Barbosa, S. J.—which are unavailable today. The Japanese words were possibly included because Rhodes studied Chinese, Vietnamese, as well as Japanese and may have referred to the Dictionarium Latino Lusitanicum, ac Iaponicum (1595), a Latin-Portuguese-Japanese dictionary and/or the Vocabulaio da lingoa de Iapam (1603-1604), a Portuguese-Japanese dictionary, which were printed by Jesuits in Japan. However, when comparing examples of Japanese words in Rhodes’s dictionary with the earlier Japanese dictionaries, there is no distinct evidence of a close relationship between them. Moreover, there are differences in meaning and spelling in several words between them.

Researching the historical background revealed that there were two main towns in Annam—Tourane and Faifo—where Catholic missionaries lived with hundreds of Japanese, and the number of Christians among them increased after the Japanese government began persecuting Christians in Japan in 1614. This implies that Rhodes included the Japanese words in the dictionary because several Japanese loanwords already existed in Annam at the time.

Keywords: Jesuits, Japanese, Portuguese, Annamese (Vietnamese)

References
Developing a Bi-Directional Ilocano-English Translator for the Travel Domain: Using Domain Adaptation Techniques on Religious Parallel Corpora

Ariel Neimond A. Lazaro, Nathaniel Oco & Rachel Edita Roxas
National University
ariel.11580@yahoo.com | {naoco, reoroxas}@national-u.edu.ph

Current work is on the development of a bi-directional basic travel expression translator for a low-resourced language. Domain of the study are English and Ilocano - a Malayo-Polynesian language used in Northern Philippines and with wider communication status. We addressed the challenge of limited available travel parallel corpora by using domain adaptation techniques on existing Bible editions to augment an in-domain 89-sentence corpus. Through keyword search, we selected sentences that are relevant to the domain and we manually simplified keywords. The resulting corpus has a total of 20,964, 2,964 of which were used for testing. Automated testing resulted with 44.38 BLEU score for English-Ilocano and 46.19 for Ilocano-English. The higher BLEU score of the latter can be attributed to the fact that conjunctions in Ilocano can have multiple translations in English depending on how it was used. Philippine languages also enjoy free-word order, which explains the result. The work can be extended by bootstrapping from a small travel expression corpus and integrating it in a public service project such as the ASEAN Machine Translation project.

Keywords: Statistical Machine Translations (SMT), BLEU, Moses SMT, Domain, Translation Model

References
Ang, et al. (2015), FEBSMT (Filipino-to-English Bidirectional Statistical Machine Translation Using Feedback)
Defense Language Institute Foreign Language Center, Ilocano Cultural Orientation (2006), Koehn, et al. (2003), Statistical Phrase-Based Translation, University of Southern California.
Lim, et al., (2008), Domain Adaptation Through Phrase Generalization for Improved Statistical Machine Translation Quality

1 Data based on Ethnologue: https://www.ethnologue.com/language/ilo
Features of *The English-Chinese Dictionary* Edited by Prof. Lu Gusun

**Ang Li**
Shenyang University
anglee555@aliyun.com

*The English-Chinese Dictionary* edited by Prof. Lu Gusun, the first comprehensive bilingual dictionary completely compiled by Chinese scholars has 200,000 entries with 15 million words. Since its publication it has exerted great influence on the learners of English in China and has been one of the reference books in the United Nations. Its influence is hard to measure. The features of the dictionary are mainly discussed in the paper which is composed of three parts. The first part explores the spirit of a craftsman as a dictionary editor. The second part discusses the main features of the dictionary respectively. The third part talks about the meaning of the dictionary in the present time as well as its shortcomings. The main features of the dictionary are 1) search rate, 2) authority, 3) independence, 4) usefulness, 5) allusion, 6) commonness and 7) perfection. The paper is not a simple demonstration of how great the dictionary is, but an attempt to make scholars to think over how to edit or compile a better large-size dictionary in the future. This paper is written in honor of Prof. Lu Gusun as a homage to him and his great work.

**Keywords:** *The English-Chinese Dictionary*, Prof. Lu Gusun, large-size dictionary, features
The Presentation of Metaphorical Information in Chinese-English Dictionaries: An Empirical Study

Dehui Li
Xiamen University
jessica8604@126.com

Cultivating the metaphorical intelligence of English of foreign language (EFL) learners is an increasingly urgent task in China, which should be one of the important productive services Chinese-English dictionaries (C-E dictionaries) are destined to do. The statistical study and questionnaires of the C-E dictionaries popular among Chinese dictionary users, however, reveal that they fall short of the expectations in many aspects. A pre-test was taken of the target constructions among intermediate and high EFL learners. The close study of the structures, lexical meanings and syntactic features of metaphorical constructions leads us to a tentative approach to the presentation of the metaphorical information in C-E dictionaries. The compounds in sub-entries, categorized into constructions of different types based on corpus and C-E dictionaries at hand, bear metaphorical meanings through the integration process and exhibit potentiality of metaphorization of the one-morpheme headwords. Clues for entry and sub-entry collection, equivalences and illustrative examples are drawn from evidences from parallel corpus showing the collocation patterns and prosodic preferences of some randomly selected compounds. The learning effect of sample entries based on this proposed presentation of metaphorical information in microstructure are testified by the language production tests and questionnaires among EFL learners in China.
The Back-of-the-book Indexes in the Chinese Lexicographical Books

Huixian Li & Qin Dai
Center for Lexicography, Guangdong University of Foreign Studies
lihuixian1002@126.com | qudyqin@qq.com

The study focus of this paper is the status of the back-of-the-book indexes in the Chinese lexicographical books with relevant insights gained during the research process. The back-of-the-index is an effective tool for knowledge and information obtainment and retrieval, and it enjoys a long history of development in the western countries; while in China, research works show that there is a lack in this field as standard principles for the back-of-the-book index have not completely established which remain open to many doubts and disputes. This paper shall re-define the concept of back-of-the-book index from a lexicographical perspective, holding that references, cross-references, terminologies and a list of dictionaries involved in the book (if any) should all included into the index, and it also means such inclusion should take both the original words and phrases extracted from the book and the generalized concepts abstracted from the book into account. Indexes should be an organic entity as a complementary compositional part of the book, they are methods of evaluating quantitatively the quality of a book; furthermore, they are vital ways to achieve information extension and expansion. Another important element regarding indexes is book readers and users. The existence of index is to facilitate reading, understanding, retrieving and further information exploration for readers and users to let them find what they want in an efficient and enlightening way. And it conforms to the concept of user-friendly and user-oriented. At this point a question should be asked that who is responsible for the indexing work, the paper will try to look into the viewpoints of mainland scholars to make feasible index-making suggestions.

Keywords: Back-of-the-book indexes, lexicographical books
Coping with Medical Polysemy in Online Medical Dictionaries: A Case Study

Jinyan Li
Guangdong University of Foreign Studies
lilacpark@163.com

Abstract
Polysemy is a pervasive phenomenon, even within a specialized domain. The polysemy of (semi-)technical medical terms has widely been confirmed, and that of semi-technical medical terms are known to pose difficulties for the target users. The present paper is intended to deal with the definitions and presentation of the polysemous term, factor, and factor-related noun phrases in online medical dictionaries. My choice of factor owes to its high frequency in medical domain, its generic feature and the potential differences in sense extension compared with specific nouns. Based on the examination of definitions and lexicographical treatments of factor in current online medical dictionaries, its qualia roles (proposed by Pustejovsky) were analyzed, in order to systematically categorize the prototypical conceptual information for conceptual description and definition of factor and its noun phrases. Corpora, such as MedAca (Medical English discourse of Academia) Corpus, E-C Parallel Corpus built by Southern Medical University, as well as the medical subsets from BNC and COCA were consulted for the meaning components, collocations, and the linguistic behavior of the term. The semantic extension of factor was charted and definitional templates for factor and factor-related noun phrases were finally proposed.

Keywords: Specialized lexicography for learners, definition, online dictionary

References
Bilingual Glossaries in Chinese Almanac: Space for Informal Language Learning

Michelle Li
Caritas Institute of Higher Education
mli@cihe.edu.hk

The Chinese almanac was listed as one of the Hong Kong’s intangible cultural heritage. Originally designed as a calendar for the farmers, the almanac gradually included more and more elements such as (in)auspicious dates, proverbs, diseases, religions, sciences, etc., making the book an interesting miscellany for many Chinese families. This paper studies one of the most noticeable features in the almanac, namely the Chinese-English glossary. The inclusion of such glossary in the almanac can at least be dated back to the middle of the 19th century, probably due to the need to communicate with the English. For example, in an 1855 almanac published in Fuzhou, there was a brief description of the phonetics of the Fuzhou dialect and a list of English words and phrases transliterated in the Fuzhou dialect. The items in the glossary were numerals and words related to trade. Similar kind of glossary can still be seen in some almanacs sold in Hong Kong, despite the fact that most Hong Kong people no longer learn English this way. Back in the 19th century, not many Chinese could receive English education, therefore this kind of Chinese-English glossaries and other phrasebooks provide an alternative for the Chinese to learn English, or more precisely Pidgin English. As an important lingua franca for China Trade, Chinese Pidgin English was mainly learnt informally through almanacs and phrasebooks. This significance of this study are two folds: culturally it reveals a special, lesser studied aspect of almanac as a cultural heritage; linguistically the bilingual glossaries in the almanacs can serve as additional sources for examining Chinese Pidgin English, as well as a means to understanding the informal transmission of English in China.

Keywords: almanacs, bilingual glossaries, Chinese Pidgin English, informal learning
The Use of an Online Dictionary by EFL learners: A Case Study

Qian Li¹, Liang Li² & Lixin Xia¹
¹Guangdong University of Foreign Studies | ²Jilin Normal University
lqchristina@gdufs.edu.cn | cdhuiyi@aliyun.com | 492130980@qq.com

Most previous studies on dictionary use in the last three decades were focused on the use of paper dictionaries, despite the fact that the move from print to digital dictionaries has been quite vigorous (Lew, 2015). This study aims to explore the process of online dictionary use by two EFL learners. Finger-driven Tracking method which is similar to eye-tracking method was adopted to record the looking-up process in text reading. Two participants of different proficiency levels (advanced vs. intermediate level) were asked to accomplish two different tasks (reading comprehension task & translation task). They were asked to move their mouse in the entry lines when they consulted the online dictionary for different types of tasks. Their moving tracks were recorded by computer for research analysis. A retrospective method was used as well to explore what the participants thought for certain mouse movement. It was found that the participants of different levels focused on different entry information and adopted different consultation strategies in the looking-up process. In addition, the looking-up processes in the reading comprehension task was different from that in the translation task. Participants tended to go through all the senses in the translation task, while they preferred to choose one of the first three senses in the reading task. This study discussed the reasons for the differences in the use of online dictionary by EFL learners. Suggestions for further research were proposed as well.

**Keywords:** dictionary use, online dictionary, looking-up process

**References**


Xiqin Liu¹, Dongping Zheng² & Yushuai Chen³

¹ South China University of Technology; Guangdong University of Foreign Studies
² University of Hawai‘i at Mānoa | ³ Jinan University

flxqliu@scut.edu.cn

EFL learners’ motivation for mobile technology use deserves in-depth investigation. This study aims to explore the latent classes of smartphone dictionary users among Chinese university EFL learners in terms of motivation for using such dictionaries. Under the framework of Gavrilidou’s (2013) Strategy Inventory for Dictionary Use, this mixed methods inquiry employed 11 semi-structured interviews followed by a confirmatory questionnaire survey (N=659). The former revealed four tendencies of smartphone dictionary use and six tendencies of users’ viewpoints about dictionary benefits. On the basis of these tendencies, an online questionnaire was designed and administered nationwide. By latent class analysis with Mplus, a model was identified, with three groups labeled Customisation (33.3%), Utilitarian (14.8%) and Ordinary (51.9%). The first two groups respectively exhibited concern for customised dictionary features and for utilitarian purposes like exams, but the last group demonstrated concern for neither. Multinomial logistic regressions showed male users and non-key university undergraduates were more likely to fall into Customisation while English majors more apt to be Ordinary. The demographic variable of gender was found to be more powerful predictors of class membership than major and university. The findings from converging data complement existing literature on MALL with new evidence from lexicography. The study has implications for both lexicography and EFL teaching, illuminating tangible pivots for e-dictionary customisation and education in e-dictionary use. Methodologically, it shows that latent class modelling can be a practical approach to user or learner classification, facilitating researchers, lexicographers and teachers to understand and scrutinise problems from a new perspective.

Keywords: smartphone dictionary use; Chinese EFL learner; motivation; mobile assisted language learning (MALL); customisation

References
The English translations of major political concepts assume particular importance in diplomatic and international communication of China. Taking The Governance of China in 2014 and Davos Speech delivered by President Xi in 2017 as the data, the author conducts a tentative research into the lexical collocation networks of “community of shared future” in Chinese-English transcultural discourses visualized by GraphColl and further interpreted through graph theory. By combining Corpus Linguistics and discourse analysis, the researcher interprets the concordance and collocation data generated by AntConc, and collocation networks of “community of shared future” visualized via GraphColl in LancsBox, which serves as a priming study for further demonstration of lexical connection between “community of shared future”, “Chinese Dream” and “National Rejuvenation”. According to results from AntConc, there is a major lexical transition from “community of common/shared destiny” to “community of shared future” which are used respectively in the translations to President Xi’s notion of “mingyungongtongti/命运共同体” in 2014 and 2017. The reason for the change in diction would be illustrated indirectly through the collocation networks of “future” with “bright” as a commendatory adjective of positive implications. Different from the previous study of the two-way collocation, GraphColl in LancsBox visualizes the collocation networks of each word in “community of shared future”. With double clicks on each word, graphs generated by GraphColl prove the collocation pattern of “community”, “shared”, “common” and “destiny” as a specific cluster. Furthermore, double click on “shared” in the corpora of The Governance of China extends the lexical connections with “Chinese Dream” (dream/ Chinese/ realization/ realize) and “National Rejuvenation” (nation/ rejuvenation/ Chinese/ youth).

Keywords: community of shared future, collocation network, GraphColl, graph theory

References
This paper is an attempt to reconsider to what extent sense ordering by prototypicality is applicable in English-Chinese dictionaries for advanced learners (hereafter ECDAL) and how presentation of senses can be improved to enhance learning of polysemous words. After reviewing polysemy from the perspective of The Prototype Theory, this paper points out that the polysemous category should be holistically treated in ECDALs in view of relatedness between senses and a multi-dimensional internal structure. However, compilers of ECDALs tend to adopt a modular approach to polysemy for the sake of accessibility and arrange senses linearly or hierarchically due to flatness of the dictionary text. Empirical findings show that most users are only concerned with the sense that fits in certain context and relatedness between senses often goes unnoticed. In addition, the sense-centered addressing structure in ECDALs distance senses from each other and makes it even harder for users to perceive the implicit relatedness. The author of this paper proposes that senses be ordered by frequency to meet users’ searching needs and a graphic menu provided as an aid for the holistic learning of polysemous words.
Counting and Expressing Numerals in Asian Languages: How People View Their World

Noa Nishimoto
Center for South East Asian Studies, Kyoto University
noa@cseas.kyoto-u.ac.jp / triathlon.nishimoto@gmail.com

Numbers are unlimited; however, numerals are limited. Human beings recognize and work with numbers, which is one of the most important activities in our daily life. Although the methods of counting and numerals as a grammatical category differ across languages, we view the world and employ natural and easy techniques to count objects, in general. This paper mainly discusses the methods of counting and expression of numerals in Asian and some other characteristic languages and attempts to find the universality and diversity of humankind’s view of the world through these languages. Similar methods of counting are used worldwide; the numerals themselves reflect the way we count and method we employ for counting. For counting, we use pebbles, shells, notches, strings, and body parts. In Balinese, lima indicates “hand” and “five.” This method of counting using fingers is natural and obvious; the corresponding words denoting “hand” in the same language group (Austronesian) as Balinese are limo in Toraki (Sulawesi) rima in Rapanui, lima in Kavaran (Formosan), dima in Ivasay, rima in Tahitian, and so on. In ancient Japan, thread, fiber, and mote were used to express small numbers and grains, ditches, moistures, and rivers were used for bigger numbers. The English word “calculation” originated from the Latin calculus, which means “pebble.” Body parts and easily available objects such as pebbles, branches, and grains are also used for measuring or surveying. Finally, we should remember that several living languages do not have numerals; many depict only 1 and 2. Some examples are Kawi (old Javanese), Chiquitos, Tacanas, Pirahã, the Gudang dialect, and many of the aboriginal languages in Australia and American Indian. However, this does not imply these speakers are primitive or ignorant of arithmetic; they do recognize numbers and apply them in everyday life.

Keywords: Numerals, methods of counting, humankind’ view of the world

References
Conant, L.L. (1895). The number concept: its origin and development, The Eco Library.
Re-examining the Big Five English Learner Dictionaries for Semantic Prosody

Vincent B Y Ooi
National University of Singapore
vinceooi@nus.edu.sg

As a central concept in corpus linguistics, semantic prosody is claimed to be inadequately incorporated into learner dictionaries by Wang and Cui (2015). While Wang and Cui’s suggestions hold true for dictionaries that are not corpus-based, can the same be said of the ‘Big Five’ online learner English dictionaries (LDOCE, OALD, Macmillan, Cambridge, and M-W Learner’s – cf. Lew 2011) that are compiled on the basis of corpus evidence? In this paper, we take 10 well-known examples of semantic prosody from the corpus linguistic literature (Louw 1993, Sinclair 2004, Hunston 2002, Stewart 2010) in order to examine whether their respective analyses harmonize with those in the Big 5 dictionaries. Examples for examination include utterly(adj), set in(intr v), undergo(v), unleash(v), affect(v), recipe for(n), naked eye(n), sit through(v), brook(v), and bent on(idiom).

The respective semantic prosodies for these lexical items are then checked against the recent GloWbe corpus (Davies and Fuchs 2015) whose composition makes it suitable to compare 20 English-speaking countries that exhibit various types of lexical priming (Hoey 2005), of which semantic prosody is a central element. The GloWbe corpus indicates the pluricentric nature of English nowadays and its study will widen the international base of potential learners of English.

An illustration comes from utterly, which is listed as having an overwhelmingly negative prosody in the corpus linguistic literature, but rather attenuated in the OALD definition as follows:

[(used for emphasis) completely]: We're so utterly different from each other.
She utterly failed to convince them.

The sense of negativity can only be inferred from the chosen example sentences selected by the OALD.

However, LDOCE does explicitly indicate this overwhelmingly negative sense, in both example and definition:

[completely – used especially to emphasize that something is very bad, or that a feeling is very strong]: You look utterly miserable.

In addition, an analysis of the GloWbe corpus shows not only negative semantic prosody examples but also positive ones, e.g.:

(Canada-GloWbe): Housewife is equal parts multilayered and beautifully nuanced - an enthralling debut told in an utterly original voice.

(Australia-GloWbe): One thing you have to say about styling in the 80s... utterly fearless. I mean think about it
We therefore propose that the Big Five dictionaries should provide a more harmonized and inclusive picture of English to learners, thus justifying the “International Student’s Edition” label often advertised for such dictionaries.

**Keywords:** semantic prosody, Big Five learner dictionaries, corpus linguistics, English lexicography

**References**


The Effectiveness of Multimedia Lexical Knowledge Representation as a Learning Tool

Takeshi Sato & Tyler Burden
Tokyo University of Agriculture and Technology | Meisei University

tsato@cc.tuat.ac.jp | tylerburden1@gmail.com

This study examines the effectiveness of multimodal representation of lexical information, which would lead not merely to a reference tool but also to a learning tool of English as a foreign language (L2). Knowledge representation under a multimedia environment where verbal and visual knowledge can be concurrently displayed has been recognized as an important factor of longer retention of the target L2 knowledge (e.g. Lindstromberg & Boers, 2008; Sato, Lai & Tyler, 2014). Internet-based dictionaries, however, might not have made full use of their potential and rather keep the traditional representation as a reference tool. Their knowledge representation depends mainly on verbal one, or several traditional images are added as glosses, although multimodal lexical knowledge presentation can make salient lexical items and their linguistic features, which would be a prerequisite of a good L2 learning tool (Pachler, 2001). This study, therefore, develops a multimodal material for L2 phrasal verbs. Learning the multiword units can be important for successful L2 use because they appear frequently on daily basis, whereas they are difficult to learn due to the fact they are polysemous including physical and metaphorical senses. This feature is recognized as challenging by the learners (Garnier & Schmitt, 2016). This material is designed to display the schematic images (Lakoff, 1987) for each verb and preposition as well as verbal knowledge, and then to radicate the images which are different from each word. By learning the phrasal verbs with this material, Japanese L2 learners, this study hypothesizes, could have a profound understanding of the semantic structure of each phrasal verb. Then this leads to successful learning in their L2 text comprehension and production with the target phrasal verbs. The findings would suggest the optimized lexical knowledge representation as a learning tool under a multimedia environment.

Keywords: learning tool, multimedia, phrasal verbs, polysemy, schematic images

References


The selection of good illustrative examples is crucial for pedagogical lexicography. Historically speaking, the first learner’s dictionary, the *Idiomatic Syntactic English Dictionary* (ISED) (1942), contained many short phrasal examples, and this tradition was succeeded by the *Oxford Advanced Learner’s Dictionary* (OALD) until the *Longman Dictionary of Contemporary English* (LDOCE) was published, which drastically increased the number of full sentence examples. Since then, monolingual dictionaries seem to focus on full sentence examples and the difficulty levels of illustrative examples in Big 5 (OALD, LDOCE, CALD, COBUILD, and MEDAL) are increasing as they went through new revisions. On the other hand, many bilingual dictionaries in Japan still take a balance between phrasal examples and full sentence examples, with specific target users in mind. It is also a major concern how we should control the difficulty levels of examples in a dictionary according to the proficiency levels of target users (Kawamoto and Tono 2015).

This paper critically reviews the status of dictionary examples in monolingual and bilingual English learner’s dictionaries. First, the overall lexical properties of illustrative examples in Big 5 were examined based on Big 5 example database. Also, approximately 2,000 examples were extracted from both monolingual and bilingual dictionaries for learners at different proficiency levels for further analysis. Linguistic properties of illustrative examples, such as phrasal vs. sentence, the number of words in examples, the vocabulary levels of words used in examples, the ratio of proper nouns, as well as types of dictionaries (monolingual vs. bilingual) were set as predictor variables in a regression model in order to explain the levels of dictionaries (beginning, intermediate, or advanced). Those factors relevant to the selection of dictionary examples for different levels of learners were evaluated. The results seem to indicate some useful patterns as well as unsystematic nature of example selection.

**Keywords:** illustrative example, pedagogical lexicography, Big 5, bilingual dictionary
Zongzi, a Prospective New Word from Chinese Entering the OED?

Tai-yuan Hugo Tseng
Soochow University (Taipei)
hugozeng@scu.edu.tw

Zongzi, one of the most distinctive traditional food items in the Chinese-speaking world, has not yet found its way into the English lexicon. In contrast to baozi and jiaozi, another two members of the zi-family already entered in the Oxford English Dictionary (henceforth OED), zongzi has not been included in the OED but seems a likely prospect.

Zongzi is variously known in the West as rice dumpling, sticky rice dumpling, or Chinese tamale, all of which are semantically vague and may allow different interpretations.

No major monolingual English dictionaries have entered zongzi, even the most comprehensive ones on both sides of the Atlantic. Interestingly, Grolier International Dictionary, a small-size, lesser-known Australian dictionary of international English, has it codified in a different form (Tsung Tzi rice dumpling).

A few hundred loanwords from Chinese are listed in the OED, where some traditional Chinese delicacies have their names. I argue that, based on corpus evidence and following a similar pattern of baozi and jiaozi, the inclusion of zongzi is well in sight.

CNN, the New York Times, and other major English-language media, when introducing this traditional Chinese food, has consistently called it zongzi (or zong for short). Furthermore, the corpus evidence from Google shows that zongzi has returned over one million hits, a solid proof that this term is in wide, current and continued use.

Baozi (in its earlier form of pao-tzu) was not added to the OED until 2005, jiaozi not until 2006, and bao, the clipped form of baozi, not until 2016. In this paper, I will show that zongzi will follow suit and soon earn its seat in the OED.

Keywords: zongzi, OED, Chinese word, new word, loanword
On the Lexicographical Characteristics of The Chinese-English Dictionary

Jiangbo Wan
Fudan University
jbwan@fudan.edu.cn

In the new era of globalization, while China is developing the Chinese discourse system to further enhance the national cultural soft power, it is imperative for our lexicographers to compile a high-quality dictionary between Chinese and English – the world’s two most widely-used languages – to meet the needs of the times. The Chinese-English Dictionary (hereinafter abbreviated as CED) is born into such a historical period. Its chief editor, late Fudan University Professor Lu Gusun (陆谷孙), says in its “Preface” that in the context of globalization the dictionary is intended to serve the Chinese readers who are interested in – but with inadequate linguistic knowledge of – their language (and probably, the target language), the foreign readers who want to learn about the most commonly used Chinese characters and phrases, as well as to serve the communication between different languages, especially the translation practice between Chinese and English. In August 2015 – two centuries after the birth of the Chinese-English dictionary, the first volume of CED was published after 15 years’ editing. It is expected that after the publication of its second volume CED will include 15 million Chinese characters, with over 320,000 entries and over 500,000 definitions. Particularly CED covers the following lexicographical characteristics: first, CED is designed and built as an “unabridged” Chinese-English dictionary; second, besides offering information on pronunciation and meaning of a lemma, CED also provides its readers with morphological exegesis of some Chinese characters; third, CED follows “descriptivism with a grain of salt” as its compilation guideline; fourth, CED defines lemmas and translates citations both semantically and pragmatically, striving to go “beyond equivalence”.

Keywords: The Chinese-English Dictionary (CED), unabridged, morphological exegesis, descriptivism with a grain of salt, beyond equivalence

References
As an important component, vocabulary play a significant role in teaching Chinese as a foreign language. In advanced Chinese language learning, the amount of vocabulary the language learner acquired turns out to be the key for them to raise their communication competence. This paper examines the homonyms from The Graded Chinese Syllables, Characters and Words for the Application of Teaching Chinese to the Speakers of Other Languages that underwent definition changes in the fifth and seventh Contemporary Chinese Dictionary (CCD). First, this paper compares the senses of all advanced-level words in the fifth and seventh edition of CCD and analyzes 5 homonyms in detail. The change in definition is mainly due to increase or deletion of contain contents. Secondly, this paper examines the 5 homonyms in four learner's dictionaries and compared them with CCD. It finds that different dictionaries deal with them in different ways. Third, this paper uses the CCL corpus to do sense annotation to the 5 homonyms. The sense frequency information shows that if the senses are arranged by frequency, the order will be quite different from the existing dictionaries, which shows the importance of sense frequency in learner's dictionaries. In sum, this study leads to a new direction to deeply examine the handling of homonyms in international Chinese teaching.

**Keywords:** International Chinese teaching, advanced level, homonym, learner's dictionary, sense frequency

**References**


Parallel Corpus on Different Language Representations of the Same Chinese-English Concepts

Xiaofang Wu, Houwei Ge, Yu Wang & Li Wei
Chuzhou University
js.xiaofang@163.com

Corpus lexicography involves the studies of corpus-based lexicographical theories and techniques. Under the theoretical framework of corpus lexicography, this paper will conduct a discussion on the development and application of the parallel corpus on the different language representations of the same concepts (DLROSC) in Chinese and English. The expression of the DLROSC in this paper is used to designate the linguistic phenomenon that different cultures have different language expressions, i.e. lexical expressions, phrasal expressions and clauses, to represent the same concepts which include the basic nominal concepts and event concepts in our daily life. For the language representations of the same concepts, the languages of different cultures share great commonness because human beings possess the similar physical structures and interact with the similar real world. In a meanwhile, there also exits certain differences.

If these differences are collected, classified, annotated with cognitive motivations and eventually developed into Chinese-English bilingual parallel corpus, it will undoubtedly be a great benefit for English foreign language learners. This paper therefore will first attempt to explore how to develop the parallel corpus on the DLROSC in Chinese and English which is labeled difference types and cognitive motivations, and then to discuss how to apply this corpus. For the development of the parallel corpus, the dedicate effort is supposed to be taken into the following works: (1) the collection of enough linguistic data on DLROSC in Chinese and English; (2) the analysis, classification, cognitive motivation annotation of the data; (3) the alignment of the data and the input of data. For the application of the corpus, we will follow the example of Corpus of Contemporary American English.

Keywords: DLROSC in Chinese and English; parallel corpus; development; application

References
Middle Construction: Difficulties for L2 Acquisition and Adaptation to Dictionary Representation

Cheng Xiong
Guangdong University of Foreign Studies
cheng@dwelt.in

Traditionally, English middle construction is believed to take an active verb morphology and a passive argument structure, thereby posing a challenge for L2 acquisition. Latest lexical approaches (inter alia, cf. Ackema & Schoorlemmer 1994, Marelj 2004 and Lekakou 2005) have favored the analysis that middle construction has a modal reading, a view yet to be made available to EFL learners. However, current ECLDs have paid little attention to learners’ needs in the representation of middle construction. In light of status quo, the present study is conducted from user’s perspective, one that heeds the process of language acquisition.

This research studies the semantic features and the selectional constraints of middle construction alone with other factors affecting its acceptability. With the help of corpus data and questionnaires, this research aims to identify the difficulties underlying the acquisition and use of middle construction among Chinese EFL learners. A survey is led to analyze the representation of middle construction in three learners’ dictionaries, revealing room for improvement. In the end, specific sample representation is proposed on the basis of Multidimensional Definition Theory (MDT).

Major findings include: 1) Chinese EFL learners have difficulties in acquiring middle constructions due to insufficient language input, overgeneralization and crosslinguistic expressional differences. 2) The representation of middle constructions in current ECLDs is not appropriate or satisfactory enough. Features of middle constructions are not shown clearly while some necessary information is not provided. 3) The representation of middle constructions in learners’ dictionaries can be improved by attending to difficulties identified.

Keywords: middle construction, dynamic modal, ECLD, MDT

References


Lekakou, M. (2005). *In the middle, somewhat elevated: the semantics of middles and its cross-linguistic realization* (Doctoral dissertation, University College London (University of London)).

Using Think-Aloud Protocol to Investigate the E-Dictionary Use by Chinese Learners

Tingting Xu
International Cultural Exchange School, Fudan University
tingtingxu@fudan.edu.cn

Currently, the knowledge about how Chinese learners use E-dictionaries and what information they need is mainly based on data collecting from surveys and interviews, which is insufficient to reveal the actual use states. This study aims to deepen the understanding of dictionary use behavior by investigating the Chinese learners’ mental processes when they consult E-dictionaries for encoding and decoding. The main research methodology is the Think-Aloud Protocol, which is known as a valid instrument of deriving information about cognitive processes, supplemented with a questionnaire on dictionary use habits and preferences.

Nine Chinese learners were recruited from preliminary, intermediate and advanced Chinese classes at Fudan University. They were asked to vocalize their thought processes while they were engaged on a reading comprehension task and a writing task with the help of E-dictionaries installed on their smart phones or tablets. The introspective data were gathered by the researcher through audio recordings and observational notes, and later transcribed and analyzed.

The study has confirmed learners’ heavy dependence on bilingual dictionaries regardless of the L2 proficiency levels in both reading and writing tasks, and the general awareness of the collocation and grammar information. Moreover, the results demonstrate that learners looked up metaphorical meanings, rhetorical devices and information about appropriateness more often and regarded them as more useful than researchers had known. Some other issues show the connection between the success/unsucces of look-ups and the L2 proficiency levels. Learners with lower Chinese proficiency encountered more problems in selecting the right entries since they had weaker ability of recognizing word boundaries in written Chinese. Advanced learners applied more strategies to selecting the right meaning of polysemous words and looked up not only the lexical meaning but also the specific usage of the meaning.

Keywords: Think-aloud, E-dictionary Use, Dictionary Look-up Strategy, Lexical Information

References
Atipat Boonmoh (2012). E-dictionary Use under the Spotlight: Students' Use of Pocket Electronic Dictionaries for Writing, Lexikos 22, 43-68
EFL Dictionaries Should Treat Meaning More Properly

Shigeru Yamada
Waseda University
shyamada@waseda.jp

There are several factors that contribute to the difficulty of describing and grasping meaning properly in recent EFL dictionaries. For this, corpus basis, frequency principle, full-sentence definitions (FSDs), and defining vocabulary (DV) are chiefly held responsible. The use of corpora in dictionary compilation is standard, and the elicited frequency information provides guidelines about the following decisions: identification of headword items, senses, and patterns; inclusion of them; and the ordering of senses. The context-sensitive analysis of corpus data tends to divide the meaning of a word into (too) fine senses that are arranged in order of frequency. As a result, related senses of a headword are scattered. Large corpus data has made it possible to identify usage patterns of words. If a word is predominantly used in a phrase rather than in isolation, it can happen that only the phrase is treated without the word itself being defined. This creates an unfortunate situation where users can understand the meaning of the phrase but are left in the dark about the meaning of the central constituent of the phrase (e.g., kilter in out of kilter). FSD can be detrimental to the description of meaning. In a usual FSD, the if-clause shows the headword in its typical context(s) and the main clause deals with its meaning. However, this definition method, placing emphasis on usage, can make it difficult for users to understand the exact meaning of the headword. DV restricts the range of vocabulary that lexicographers can use for definitions. The resultant definitions can be imprecise because of the constraints on word choice. In the era of electronic lexicography, which makes space limitation less of a problem and allows for flexible presentation, these problems related to meaning can be solved practically.

Keywords: defining vocabulary, EFL dictionary, frequency, full-sentence definition, sense discrimination.

References

Study of Confucian Words in Chinese-English Dictionaries: Based upon Phraseology in Contrast

Rong Yang
Sichuan International Studies University
804739912@qq.com

Phraseology in contrast and the present situation of Confucian words’ definition in Chinese-English Dictionary provide the theoretical guidance and research basis for the study of the Chinese-English equivalence of Confucian words. Based on the theory of Phraseology in contrast's extended unit of meaning and BNC and modern Chinese corpus, this paper studies the form, meaning and function of 明明(adverb) and “obviously” from colligation, semantic preference and semantic prosody three levels. The results show that: compared to obviously, "it is obvious" and 明明 are more equivalent. The study not only provides a data reference for improving the quality of Confucian words’ definition in Chinese-English Dictionary, but also holds great significance for the construction of cultural power and promotion of the Confucian classics’ "going out".

Keywords: Confucian words; colligation; semantic preference; semantic prosody; corresponding units of meaning

References
Assisting Non-Native Chinese Learners: Robert Morrison’s Pioneering Efforts in Wuche Yunfu (1819)

Ying Ye & Xiangqing Wei
Nanjing University | Nanjing University of Chinese Medicine
Nanjing University
clace_ye@163.com | dicweixiangqing@163.com

The present paper examines the lexicographic efforts Robert Morrison put into his compilation of Wuche Yunfu (五车韵府) (1819), to assist those non-native Chinese learners in his times. It explores how Robert Morrison provided his target users with different access to both the linguistic and cultural information of the Chinese language. More importantly, based on the detailed textual analysis, the authors of the present paper argue that Morrison's Wuche Yunfu (五车韵府), as a matter of fact, did well serve the purpose of helping those non-native Chinese learners in 19th century with their second language learning. Through his tremendous efforts, Morrison demonstrated how a bilingual dictionary could be learner-friendly and learning-oriented. The present study aims at depicting the design features of Wuche Yunfu (五车韵府), which are considered the neglected legacy that Morrison left for his successors. It is hoped that Morrison’s lexicographical practice will inspire bilingual lexicographers today to produce in the future better learner’s dictionaries for non-native Chinese learners.

Keywords: learner-oriented, learner-friendly, bilingual dictionary, Wuche Yunfu, Morrison

References
A Second Bite at the Ultimate Dictionary

Xuhua Zhang
Fudan University
zhangxuhua@fudan.edu.cn

There is little doubt that phraseology is at the heart of all language use. Although did not use a unified terminology, an increasing amount of corpus-based research into English (e.g., Biber & Conrad, 1999; Cowie, 1999; Wray, 2005) has evidenced that multi-word expressions are used in a broad range of everyday situations. The overall finding is that the phraseological tendency in language use is much greater than previously assumed. However, there has been no principled way to prioritize the inclusion of multi-word expressions with constituency variation or positional variation in pedagogical material or bilingual dictionaries. Moon (2008) is a good practice to discuss the treatment of phraseology in dictionaries.

This study is inspired by the hypothesis of ‘ultimate dictionary’ made by Sinclair, Jones, & Daley (2004: xxiv), in which Sinclair postulated that “a dictionary containing all the lexical items of a language, each one in its canonical form with a list of possible variations, would be the ultimate dictionary”. We tried to find the canonical form of multi-word expressions and the related full phrasal expressions with the help of online corpus and search engine (http://corpus.byu.edu/) and software packages (WordSmith Tools and ConcGram). It first observes the phraseological behavior of “touch” in the Corpus of Contemporary American English (COCA) and its Chinese equivalents in the Modern Chinese Corpus (MCC). It then considers the ways in which their collocates and phraseological patterning are represented in dictionaries. The feasibility of recording lexical items and their variations is discussed at the end of the paper.

Keywords: phraseology, corpus, ultimate dictionary

References

Form-Concept Mapping and the Organization of Semantic and Contextual Information in *The Chinese-English Dictionary (Unabridged)*

Cuilian Zhao  
Fudan University  
wyzcl@126.com

The process of identifying and arranging meanings and contextual information is a complex task for dictionary writers. Increased attention to context and to the mental lexicon on the part of a reader/user gives them an opportunity to introduce new data *en masse*. Dictionary writers can organize the entry structure in a more logical and effective way by mapping the dictionary senses (or translational equivalents or definitions in the case of a bilingual dictionary) and contextual information (provided in examples) onto the reader/user’s mental lexicon, enabling him/her to better understand the entries by way of accessing prestored conceptual nodes, store newly learned chunks as conceptual nodes, or use them appropriately in context. Entries of the *Greater China Chinese-English Dictionary* follow a form-concept mapping framework wherein established meanings are rearranged and contextual data are organized.
The Influence of English-Japanese Lexicography Upon Early Monolingual English Learners’ Dictionaries

Lianzhen Zhao
China Pharmaceutical University
buddy_zlz@163.com

Though the first generation of monolingual English learners’ dictionaries (MLD) emerged in the mid-20th century in Asia, their origin is often attributed to the research into vocabulary control, pedagogy grammar, and phraseology by the three founding fathers, Michael West, Harold Palmer, and A. S. Hornby. Thus, the lexicographical activities in Asia, especially in Japan, during that period and their potential influence on the early MLDs are under-explored in the historical research of MLDs. In consideration of this, the present study attempts to investigate the impact of English-Japanese lexicography on the genesis of early MLDs in relation to the formation of their central defining features. It first sketches the early history of MLDs, in respect to West and Endicott’s A New Method English Dictionary (NED), Palmer’s A Grammar of English Words (GED), as well as Hornby’s The Idiomatic and Syntactic English Dictionary (ISED). Based upon this, the study argues that it is the encoding functions of GED and ISED that mark the MLDs’ break with native speakers’ monolingual English dictionaries. The study proceeds to gives an in-depth analysis of the design features that aim at the encoding functions in GED and ISED. Subsequent to this, the study traces the sources of Palmer’s and Hornby’s lexicographical conceptions and views on the acquisition of foreign vocabulary. In so doing, the emphasis is put on the influences on the two pioneers from English-Japanese dictionaries produced in that period. The study concludes with a discussion on the interplay between bilingual and monolingual learners’ dictionaries and a prospect for the compilation of future bilingual learners’ dictionaries.

Keywords: influence, English-Japanese lexicography, early monolingual English learners’ dictionaries

Reference