Proceedings of ASIALEX 2015
Hong Kong
PREFACE

The 9th International Conference of ASIALEX was held at the Hong Kong Polytechnic University on 25-27th June 2015. The theme of the conference is *Words, Dictionaries and Corpora: Innovations in reference science*.

The response to the call for papers was encouragingly strong, and seventy-five percent of presenters contributed a full paper. The range of the topics covered by the conference is rich and will certainly deepen our understanding of semantics, dictionaries and computational lexicography. The proceedings have six sections: 1) General lexicographic studies; 2) Pedagogical lexicography; 3) Corpus and computational lexicography; 4) user studies; 5) Terminologies and lexical studies and 6) e-dictionaries.

The ASIALEX 2015 was sponsored by the Department of English, and also the PolyU Funding for Prestigious International Engagements. I am grateful to the generous institutional support which made this international event possible. Teamwork, patience and tireless effort are the key to the success of the conference and the Proceedings. I would like to express my particular thanks to the organizing committee: Prof Huang Churen, Professor Martin Warren, Miss Ye Meng, Mr Liu Liming, Mr Jamie Mckeown. We were a small team running a big conference.

LI Lan  
Convener of ASIALEX2015, Hong Kong  
June 2015
CONTENTS

PREFACE

KEYNOTE SPEECHES

Thirty Years of User Studies – And what we still need to find out
Hilary Nesi 1

A New Concept and a New Format in Bilingual Lexicography: Reflections on the Making of A Comprehensive Chinese-English Dictionary
CHAN Sanwai 9

SECTION 1 GENERAL LEXICOGRAPHIC STUDIES

Adaptability of the Bilingual Theory to Dictionary Making And the Genesis of a New Kind Dictionary
Ali Aycan KOLUKISA 22

Bilingual Dictionary and Encyclopedia – Where goes the line? Cultural specificities as sources of dilemmas and solutions proposed
Mats-Peter Sundström 35

Roles That Labels in Dictionaries Are Expected to Play
Masaaki Ogura 44

Strict Corpus-based Lexicography and the Development of Dictionaries in the South African Indigenous Languages
MV Mojela 48

Comparing Japanese-English Dictionary Entries of English Loanwords with Corpus-Derived Lexical Profiles
Keith Barrs 55

The Possibility of Building a Dependency-based Japanese-English Construction Dictionary
Masanori Oya 61

The Crowdsourced Compilation of the China English Dictionary
QIN Xiaohui 74

A Dictionary of Hong Kong English vs The GloWbe Corpus?: Triangulating sources of lexicographic evidence
Vincent B Y Ooi 80

Proverbs in A Comprehensive Indonesian-English Dictionary: Critical evaluation on cultural-bound phraseology
Dora Amalia 90

Cuyonon and Tausug Cultural Dictionaries: Databank for determining mutual intelligibility of two indigenous Philippine languages
Rosario B. Dizon, Mary Ann S. Sandoval, Myrna Halim, Lydia L. Mata and Ser Robert David B. Dizon 97
Forms of Address in French-Korean and Korean-French Dictionaries from the Perspective of a French Learner of Korean
NAM Kilim 114

The Development of the Dictionary of the Usage Frequency of Basic Korean Vocabulary and Its Application to Korean Lexicology
KIM Jinung and SEO Sang-Kyu 124

Problems and Countermeasures in Chinese Lexicography Modernization
KANG Shiyong 133

Where Do We Go from Here?
Ari Kernerman 140

SECTION 2 PEDAGOGICAL LEXICOGRAPHY

Lexicography as a Teaching Tool: A Hong Kong case study
James Lambert 144

Grammatical Items Used in EFL Dictionary Examples: From the Japanese EFL Learner Perspective
Yasutake ISHII 154

Where is the place for slang? An analysis of labeling of restrictive word usage in monolingual English learners’ dictionaries
Susanna Bae and Young-kuk Jeong 162

The Research on the Name Information in the Commercial Press Learner’s Dictionary of Contemporary Chinese
HE Shan 173

Two Innovations in Pedagogical Dictionaries
Geoff Toister 182

A New Trend in Lexicography: Bringing Semantic Prosody into Learner’s Dictionary
Jusong WANG 189

Anglocentrism in the Oxford Advanced Learner’s Dictionary of Current English
Hai XU and Martin Weisser 195

Pragmatic Information in Learner’s Dictionaries: Redefined
Anmin WANG 204

SECTION 3 CORPUS AND COMPUTATIONAL LEXICOGRAPHY

Crowdsourcing Microdata for Cost-Effective and Reliable Lexicography
Martin Benjamin 213

Comparison of Dictionary Use and Corpus Use: Different effects on learning L2 phrases
Yoshiho Satake 222

House M.D. Corpus Analysis: A Linguistic Intervention of Contemporary American English
Locky LAW 230
Building an English-Filipino Tourism Corpus and Lexicon for an ASEAN Language Translation System
Charmaine S. Ponay and Charibeth K. Cheng 250

Japanese Idioms in Corpus Data and Dictionaries: Variability and Patterning
Priscilla Ishida 261

A Corpus-based Difficulty Level Evaluation of Example Sentences with the CEFR Levels in CALD4
Yoshiyuki Notohara 273

A Corpus-based Study of Noun-adjective Continuum in Japanese and English
Masayuki HIRATA and Alex Chengyu FANG 281

Lavender Filipino: Computational Models of Twitter Swardspeak
Nathaniel Oco, Raymart Fajutagana, Christine Mae Lim, Judi Diane Miñon, Julie-Ann Morano and
Ryan Christian Tinoco 302

The Old and New Tagalog Orthography: A computational comparison
Manolito Octaviano Jr., Nathaniel Oco and Rachel Edita Roxas 311

FilCon: Filipino sentiment lexicon generation using word level-annotated dictionary-based and
corpus-based cross lingual approach
Darlan Keen Domingo, Nyker Matthew King, Jerome Lorenzo Lopez, Alexandra Mondares and
Charmaine Ponay 317

SECTION 4 USER STUDIES

An Empirical Study of English Corpus as a Reference Tool for PhD Students
LI Lan 331

Using Dictionary for Learning English: a Case Study of Non-English Major Students at People’s
Public Security University of China
LI Xiu-ping 344

The L1 Guide to the Usage of an EFL Dictionary: The case of the Japanese Guide to the Usage of
Oxford Wordpower Dictionary
Shigeru Yamada 352

Meeting More Readers’ Needs: the YES-CEDICT Chinese Dictionary
Xiaoheng ZHANG 355

Lexicographic Needs in Turkic Republics
Baktygul KULAMSHAEVA 361

SECTION 5 TERMINOLOGIES AND LEXICAL STUDIES

A Contrastive Study of Phraseological Units among Andalusians and the Fluminense people
Rita Giovana M. Ferraro 369

Phraseology and Terminology in LSP: Implications for lexicography
Jirapa Vitayapirak 377

Equipping Students with Lexicographical Means to Tackle Collocation Problems
LIU Liming 382
Compilation of Traditional Building Construction Terminologies in Tamil
Saravanan RAJA 391

A Study of The Prototypical Definition of the “walk” Headword Based on the Characteristic of its Non Self-Sufficiency
FENG Haixia 403

A Comparative Study of War Metaphors in Business and General News Articles Using a Dictionary-based Database
ZHONG Xian 411

A Diachronic Comparison of the Strategies Adopted by The New York Times in Translating Expressions with Chinese Characteristics
LIU Meiyan 425

The Effects of Task Types on Incidental Vocabulary Acquisition through L2 Listening
YUAN Ximing 435

SECTION 6 E-DICTIONARIES

Improving an Electronic Dictionary for Morphological Analysis of Japanese: Use of historical period information
Tomoaki KOUNO and Toshinobu OGISO 441

Portable Electronic Dictionary (PED) Use in Spanish as Second Language Students in Japan
Lidia SALA CAJA 450

The Design of an App for Learning Chinese Idioms
Shan WANG and Yujia WU 460

An Electronic English Frequency Dictionary of Age Groups
LI Hanhang John 469

LINE Dictionary: More than a dictionary
Tae-hoon LEE and Eun-joo JEON 481

Spelling Variation Problem in Persian: Automatic Development of a Spelling Database for a Digital Dictionary
Saghar Sharifi, Marziye Sanaati and Masood Ghayoomi 488

Towards Automatic Translation from Sanskrit to English
Chitresh SONI, Girish Nath JHA, Naveen Kumar JHA 498
Thirty Years of User Studies – And what we still need to find out

Hilary Nesi
Coventry University, UK
h.nesi@coventry.ac.uk

Abstract
The main - perhaps the only - reason for investigating dictionary use is in order to improve users’ consultation methods, and to meet their consultation needs more closely. This is easier said than done, of course, given that there are so many different types of user consulting dictionaries in so many different contexts, for so many different purposes, and with such differing levels of knowledge and expertise. Moreover although the research area is still relatively young (very few empirical studies were conducted before the 1980s) it spans a period of great technological change and has experimented with a wide range of methodologies, so that studies purporting to address similar research questions, in similar contexts, have sometimes arrived at very different conclusions.
Focusing particularly on dictionaries for learners of English, this paper will trace developments in user studies over the past thirty years, and will attempt to identify conclusions we can all agree on, claims that remain contentious, and important questions that still remain to be answered. It assumes that the ‘perfect’ dictionary consultation is the one which provides the best answers, in the least obtrusive way.

Keywords: users, research, English language learners

1. Introduction

The study of dictionary use should be an indispensable part of the dictionary-making process; there is no point in changing dictionary design if the changes do not make dictionary consultation more useful, efficient and pleasurable, and these effects can only be monitored by examining user attitudes and behaviour. Although lexicographers and teacher informants might be able to deduce indirectly what lexical information specified groups of users need (c.f. Tarp 2014), without empirical investigation it is impossible to tell how these users might respond to different methods of information provision, affecting, for example, defining language, the macro and micro-structure, and the encoding of grammatical and phonological information. In many ways the process of dictionary creation is analogous to the process of language course design. Designers start by devising a sort of content ‘syllabus’ based on perceived user needs, but the success of the product depends on user-appropriate communication of this content: the pedagogical methods, the learning materials, the mode of delivery. Dictionary evaluation, like language course evaluation, must be made in terms of what end-users think and do.

The study of dictionary use is an art rather than a science, however. We expect the findings reported by corpus linguists and language engineers to be replicable by other researchers, and we expect to be able to build on these findings to create ever more informed linguistic descriptions. Studies of dictionary use, on the other hand, are heavily influenced by local circumstances and the methodological choices of individual researchers. Findings are often open to interpretation, are difficult to replicate, and may or may not be generalizable to other contexts. This is because of variation in learner backgrounds (for example their first language, proficiency level and dictionary skills), the research instruments (whether the study investigates user preferences, user consultation practices or dictionary usefulness in a particular kind of language task), and the actual dictionaries under investigation.

Perhaps we should accept that it is impossible to prove beyond doubt that one approach to dictionary design is superior to another, given that the variables in contexts of dictionary use are complex and difficult to isolate for controlled experimentation. However, as with other
types of pedagogical research, there are also serious resource issues that impede the development of best practice. Unlike dictionary development projects, user research is rarely funded from the public purse, and although it is good that publishers sometimes support the piloting of dictionary products, too much commercial involvement can run the risk of compromising study design. The lack of resources particularly hinders user studies, where edictionary software and hardware are needed to replicate natural dictionary access conditions.

All these considerations reduce the impact of user studies, despite their very great informative potential. Dictionary use researchers are also often teachers, and the insights they gather regarding user behaviour will feed back into classroom practice (see, for example, Bae 2014), but Lew and Dziemianko (2006: 277) claim that ‘few modifications to the learners’ dictionary design are supported by published results of experimental research on how learners really use dictionaries’. This paper examines trends over 30 years of research into the use of dictionaries for learners of English, aiming to identify some common ground on which teachers and lexicographers can build, and some potentially fruitful future research directions.

2. *Some areas of general consensus*

2.1 *Monolingual v. bilingual dictionaries*

A large proportion of the research into dictionary use has focussed on monolingual dictionaries for advanced learners of English. This is partly because each new edition of these dictionaries has tended to introduce some innovation worthy of investigation, such as the restricted defining vocabulary introduced by Longman in 1978, for example, or the ‘extra column’ introduced by COBUILD in 1987. It is also because, as has often been noted (e.g. by Lew 2011, 2015; Nesi 2013), participants in studies of dictionary use tend to be linguistically advanced university staff and students rather than a true cross-section of the dictionary-using public. The researchers themselves are based in universities, and they make use of the resources nearest to hand. Also, in most educational contexts printed monolingual dictionaries are more readily available to researchers as class sets; in multilingual research settings the alternative to a set of monolingual dictionaries would usually be a number of different bilingual dictionaries, of differing styles and quality, perhaps in languages unknown to the researcher.

Many studies have reported that educators hold monolingual advanced learners’ dictionaries in higher esteem than bilingual dictionaries (for example Waring 2001; Boonmoh & Nesi 2008; Chan 2011, Bae 2014), although bilingual dictionaries do seem to enjoy high status amongst Japanese teachers, according to research emanating from Japan (see for example Nakao 1998). These preferences may distort the findings of research into dictionary use, especially if the researcher is also a teacher of the research participants, as is often the case. Respondents in questionnaire surveys may exaggerate their use of monolingual as opposed to bilingual dictionaries in order to present themselves as good students, and in observational studies they may hide from sight their bilingual dictionary use. This bias can affect learner behaviour even in studies where the researchers state no preference for any particular dictionary type; Nesi and Haill (2002), for example, found that participants in their naturalistic study tended to report on their use of monolingual rather than bilingual dictionaries, presumably because they felt that bilingual dictionaries would be of little interest to their teachers.

There is, however, overwhelming research evidence that learners prefer to use bilingual dictionaries (Tomaszczyk, 1979; Baxter, 1980; Bensoussan, Sim & Weiss 1984; Atkins & Varantola 1998; Komuro & Yamada 2000; Thumb 2004; Wingate 2002; Yoon 2014). Learners have difficulty with the metalanguage of monolingual dictionaries, and lack the grammatical knowledge to make sense of grammar codes (Neubach & Cohen 1988; Nesi & Haill 2002; Thumb 2004; Miller 2009), and although there is some evidence that monolingual dictionary use increases at more advanced proficiency levels (Battenburg 1991;
bilingual dictionary use also persists at advanced levels, often in combination with other types of reference material (Atkins & Varantola 1998; Yoon 2014). Particularly positive findings have been reported in user studies of bilingualised dictionaries, by Thumb (2004), Didac et al. (2006) and Batia Laufer and her associates (Laufer & Melamed 1994; Laufer & Hadar 1997, Laufer & Levitzky-Aviad 2006), although this type of dictionary is not available for some languages and it does not seem to have been very widely investigated.

2.2 Print v. electronic
In the past many students simply bought the print dictionaries that their teachers recommended, and teachers thus had considerable power over dictionary ownership and use (see, for example, Chi 2003). The debate over whether to recommend bilingual or monolingual dictionaries may no longer have much relevance, however, given that pocket electronic dictionaries (PEDs) typically contain both bilingual and monolingual sources (old and new), and users now increasingly access free online ‘aggregate’ dictionary sites which provide links to multiple types of reference tool, including bilingual sentence banks.

Nevertheless, as in the pre-electronic era, much metalexicographic attention still tends to focus on the use of monolingual advanced learners’ dictionaries, on CD-Rom (e.g. Nesi 2010, Dziemianko 2012), in PEDs (e.g. Chen 2010) and online (e.g. Dziemianko 2010, 2011). Prestigious monolingual dictionaries are certainly easier to work with for research purposes, and teachers’ and researchers’ suspicions of popular bilingual e-dictionary content have been justified, at least in some cases. Scanning mistakes occur when converting old paper-based bilingual dictionaries to electronic form, for example, (Boonmoh et al. 2005; Nesi & Boonmoh 2009), and there are wildly inaccurate automatic translations of web content in some popular online dictionaries (Nesi 2012). Mair (2007) blames automatic translations in aggregate dictionaries for the spread, in China, of “absurdly crude English mistranslations in bizarrely inappropriate contexts”.

Entirely excluding less prestigious bilingual e-dictionaries from experimental user studies seems to be a mistake, nevertheless, because these are the dictionary types which are accessed most frequently around the world under non-experimental conditions. The developers of the aggregate dictionary Jin Shan Ci Ba, for example, claim that it is used by more than twenty million people in mainland China (Nesi 2012). Research findings indicate that bilingual e-dictionaries are chosen in preference to monolingual e-dictionaries not only because they are easier to understand but also because they are more in evidence on aggregates sites, and because they often contain additional features such as corpus data, wordlists, language tests, games, and personal organizer functions. Thai research participants in Nesi and Boonmoh’s study (2009), for example, using the default settings for local English-Thai and Thai-English dictionaries in their PEDs, and claimed never to consult the monolingual dictionaries content. Similarly students introduced to the Macmillan learners’ dictionary on CD-ROM (Nesi 2010) acknowledged the superiority of its lexicographical content, but still opted for the bilingual e-dictionaries on their hard drives and in their PEDs because they were easier to access, and contained more extra facilities. Yoon (2014) found that her Korean participants generally chose to use the Naver online Korean-English dictionary as opposed to the Longman Dictionary of Contemporary English.

Although in the early days many teachers were against e-dictionaries, the new generation of digitally-savvy teachers seems to be reconciled to their use, and more than half the teachers in Bae’s survey (2014) reported that they no longer used print dictionaries at all. When e-dictionaries first became available teachers were worried about a loss of depth of learning if dictionary consultation became faster and easier (Taylor and Chan 1994, Sharpe 1995, Stirling 2003). This seems to be because of ‘involvement load’ (c.f. Craik & Lockhart 1972), a hypothesis Koyama and Takeuchi (2004) cite to explain why their research participants retained information better from paper dictionaries than from PEDs. Although easy access to
e-dictionary information seems to encourage learners to make more frequent consultations (Nesi 2000; Ronald & Ozawa 2008, Lew 2013) it may also discourage them from recording consultation outcomes (Boyd 2011) and thereby consolidating the information they have acquired. However the majority of empirical studies (e.g. Nesi 2000; Koyama & Takeuchi 2003; Kobayashi 2007; Xu 2010; Chen 2010, 2012; Dziemianko 2011, 2012) have found little or no difference in task performance when learners use dictionaries in paper or electronic form.

Interestingly, although CD-Roms are frequently used in studies of e-dictionary use, this seems to be the format least preferred by teachers and learners. In Bae’s survey of Korean teachers (2014) over 70% were found never to have used a CD-Rom dictionary, and in Boonmoh and Nesi’s survey of Thai students (2008), only about a quarter acknowledged ownership of a monolingual dictionary CD-Rom, even though the majority owned a print copy of the Longman Active Study Dictionary which came with a copy of the CD-Rom attached to the front cover. This lack of interest in CD-Roms may partly be due to access problems; many of the students in Boonmoh and Nesi’s survey did not have their own computers, and even laptop users might find CD-Roms less portable than print dictionaries, or mobile dictionary devices. Portability is becoming more and more important; modern learners expect to be able to consult online sources on the move, via their smart phones, and this in turn has increased the range of activities which can feasibly be assisted by dictionary consultation; dictionaries with audio functions can now support conversations and service encounters. Experiments with dictionary users are lagging behind in this respect as they still generally focus solely on reading and writing tasks.

2.3 Dictionary skills
There is now abundant empirical evidence to suggest that dictionary skills can be taught effectively (Kipfer 1987; Bishop 2001; Carduner 2003; Chi 2003; Lew and Galas 2008; Miller 2009; Bae 2014; Liu 2014). A number of writers have developed taxonomies of the reference skills needed for successful dictionary consultation, for example Scholfield (1982), Bogaards (1993), Nesi (1999), Lew and Galas (2008) and Bae (2014). Tono (2001) complained about the lack of any reliable way to measure dictionary using skills, but this problem has been partially addressed by Gavriilidou (2013), who presents a standardised skills assessment instrument based on suggestions in the prior literature. Gavriilidou’s instrument seems more appropriate for print dictionary use, however, with references to buying dictionaries in bookshops and attending to alphabetical order and letter distribution. Lew (2013) provides a more thorough account of dictionary skills in the context of online dictionaries, building on previous taxonomies and the attempts by Engelberg and Lemnitzer (2009) and Pastor and Alcina (2010) to classify electronic dictionary search techniques.

3. Future directions
User studies have tended to concentrate on the skills at the centre of Nesi’s taxonomy (Stage Three: locating entry information), and in recent years the range of research methods to explore user behaviour in this area has greatly expanded. Now, in addition to questionnaires, interviews, and paper-based tests, researchers use think-aloud techniques (e.g. Thumb 2004; Nesi & Boonmoh 2009), online quizzes (e.g. Nesi & Tan 2011), log-file analysis (e.g. De Schryver & Joffe 2004; Verlinde & Binon 2010; Schoonheim et al. 2012; Koplenig et al. 2014) and eye-tracking (e.g. Tono 2011; Lew et al. 2013). Still largely missing from this area of investigation, however, are longitudinal studies which take into account the way users’ strategies develop as they increase their knowledge of the language and their reference tools. Most studies only take a ‘snapshot’ of user skills, and conflicting results from different studies may simply be due to the differing positions of participants on a learning curve, disguising the fact that they are heading in the same direction.
Studies of the way users process dictionary entries have the potential to feed back into future dictionary design. The fact learners’ dictionaries are now less heavily encoded suggests that some feedback from experimental findings has already occurred, despite the claim of Lew and Dziemianko (2006:277) to the contrary. Equally important, however, are the skills areas on either side of the central information location stage. Nesi’s six-part taxonomy (1999) includes ‘before study’ and ‘before consultation’ stages, involving the interpretation and recording of entry information and a general understanding of lexicographical principles, and post consultation stages involving the recording and evaluation of dictionary information. These areas have been less well investigated, as Tarp (2014) points out, although the ability to evaluate consultation situations and to apply dictionary information to new contexts seems to be particularly important now, as users gain easier access to a far wider range of resources, often of very varying quality. Whereas in the past the authority of dictionaries was rarely questioned, digital resources require ‘a more pragmatic and less ideological or dogmatic view of dictionaries’ (Lew and de Schryver 2014: 342). Bae and Nesi’s (2014) study of dictionary-related online queries suggests that UK/Irish users of Yahoo! Answers feel more empowered to challenge lexicographical sources than Korean users of Naver Knowledge-iN. Researchers should perhaps keep a more watchful eye on users’ evaluative behaviour, with a view to building evaluative strategies into dictionary training programmes, as Bae (2014) has done.

User studies matter. Monolingual learners’ dictionaries have evolved over time to meet user needs, possibly in response to findings from such studies, but as we have seen monolingual dictionaries are not the first choice for most learners, and while the monolingual dictionaries have been evolving many more bilingual and multilingual dictionaries have become available to users in digital form. Perhaps future user studies will have some influence on the future design of some of these products, but an immediate concern is to investigate users’ strategies for coping with the massively increased range of consultation options. E-dictionary consultation is rather a private matter, especially if it is conducted via mobile devices such as smart phones, but if teachers can find out what resources learners naturally turn to, and why, they will be in a better position to guide them in their information selection, helping them to apply information appropriately, and adopt, when necessary, a more critical stance.

References


Lauffer, Batia and Linor Melamed. 1994. Monolingual, bilingual and 'bilingualized' dictionaries: Which are more effective, for what and for whom? In *EURALEX '94


Tarp, Sven. 2014. Detecting user needs for new online dictionary projects: Business as usual, user research or …? In Wörterbuchbenutzungsforschung. 5. Arbeitsbericht des wissenschaftlichen Netzwerks Internetlexikografie, eds. Carole Tiberius and Carolin Müller-Spitzer, 16-26. Mannheim: Institut für Deutsche Sprache (OPAL - Online publizierte Arbeiten zur Linguistik).


Waring, Rob. 2001 Oxford Learner’s Dictionaries: Getting your students to use their dictionaries effectively. Tokyo: Oxford University Press Japan


A New Concept and a New Format in Bilingual Lexicography:

Reflections on the Making of A Comprehensive Chinese-English Dictionary

Chan Sin-wai
Department of Translation
The Chinese University of Hong Kong

Abstract
This paper is a reflection on the compilation of my newly published dictionary entitled *A Comprehensive Chinese-English Dictionary*, which focuses on the holism and interconnectedness of dictionary entries. Most Chinese-English dictionaries on the market follow a certain method for arranging their entries, such as those based on romanization, number of strokes, or radicals. Currently, the most popular, as well as the most widely used, method is the Hanyu Pinyin romanization. What is most undesirable about this method, however, is that entries are listed alphabetically according to the romanization of the first character, which results in the separation of terms related to the entries (such as negatives and adjectives) due to the variations in the positions of the key characters, making it impossible to see their holism and interconnectedness.

In my dictionary, the holism of entries can be clearly observed in the normal and reverse order arrangements for both single-character and multiple-character entries. In addition, all entries are arranged according to the four tones of Hanyu Pinyin for easy consultation. At the same time, the method of semantic analysis in linguistics has been applied in the reverse part, arranging multiple-character entries with key characters in different positions according to Hanyu Pinyin, which systematically shows the interconnectedness of the listed entries. It is hoped that the application of the concepts of holism and interconnectedness, and the arrangement of entries according to Hanyu Pinyin, will represent a breakthrough in dictionary format and a whole new concept in bilingual lexicography.

Introduction
This year marks the 200th anniversary of the birth of the first Chinese-English dictionary compiled by Robert Morrison in 1815. The present work, *A Comprehensive Chinese-English Dictionary*《順逆序漢英大詞典》, celebrates this anniversary by introducing a new concept and format in the compilation of a new-style Chinese-English dictionary. The new concept is semantic, grouping expressions with related meanings under the same headwords; the new format is holistic, arranging entries in both normal and reverse sequences to give a larger context for users to understand the semantic, pragmatic, and cultural connotations of the headwords in their entirety. This paper is my personal reflections on the making of this dictionary which comprises 60,000 single-character headword entries and multiple-character sub-entries, 420,000 English equivalents, 2 million words, and 3,500 pages.

A New Concept: Semantic interconnectedness
The need for a new concept in the compilation of Chinese-English dictionaries is obvious, as all dictionaries since 1815 have used either the method of Romanization or other methods of consultation that are mainly phonetics-, stroke-, or radical-based, but not linguistics-based or, in particular, semantics-based. The emphasis that this dictionary places on the semantic interconnectedness of its entries is not only its major characteristic, but this also presents a totally novel concept in bilingual lexicography.
This new orientation towards semantic relations of the entries arises out of the over-emphasis on the use of Romanization systems as the sole or main consultation method. In fact, the use of Romanization methods in arranging entries in Chinese-English dictionaries targeted mainly at foreigners and learners of Chinese and applied ever since the publication of Morrison’s first Chinese-English dictionary. This was closely related to the development of China’s relations with foreign countries since the early sixteenth century (Chien and Creamer 1986). Later, the foreigners who entered China for culture, missions, technology, and business, before and after the Opium War (1839-1842), found it necessary to learn both spoken and written Chinese. It was therefore no mere coincidence that Robert Morrison (1782-1834), a Scottish missionary, became the first person to compile *A Dictionary of the Chinese Language*, which was published from 1815 to 1823 in three parts. The entries in this dictionary were arranged according to his own Romanization system. Thomas Francis Wade later refined Morrison’s scheme, and Herbert Allen Giles, through his *Chinese English Dictionary*, further popularized Wade’s system, to form the Wade-Giles system, which became the dominant Romanization system for Mandarin. Seen in a historical context, Chinese-English dictionaries in the nineteenth century were mainly produced by missionaries such as Medhurst (1848), Williams (1874), and Goodrich (1891) to help promote Christianity and enable foreigners to learn Chinese language. Dictionaries produced in the twentieth century, on the other hand, were oriented towards language learning and academic research, such as those by Fenn (1926) and Mathews (1931). When China began to open up to the rest of the world in the 1970s, world news media started to replace Wade-Giles spellings with Pinyin, and academia soon fell into step as well (Yu et al 1999). The majority of bilingual dictionaries published in the present century use the Hanyu Pinyin Romanization (Cao 2002; Wu 2003, 2004). Taiwan, however, is an exception as the Wade-Giles system has been used from 1949 to the present. A typical dictionary that used this Romanization was *《遠東漢英大辭典》* (*The Far East Chinese-English Dictionary*) compiled by Liang Shih-ch’iu (1997).

While the Romanization of the first character of an entry is a convenient way to find the word we want to look up in a dictionary, it has the disadvantage of separating semantically related entries due to the variations in the positions of the key characters, making it impossible to recognize their semantic relationships. This dictionary, therefore, arranges all entries according to the tone order of the Hanyu Pinyin Romanization for easy consultation, while it applies the method of semantic analysis in linguistics in the reverse part to arrange multiple-character sub-entries with key characters in different positions according to Hanyu Pinyin, which systematically shows the interconnectedness of the entries listed.

Taking the multiple-character entries of the single-character entry *shi* 是 as an example, the first few entries in normal order are as follows:

*shi fei bu fen* 是非不分 cannot tell black from white; confuse right and wrong; confuse truth and falsehood; fail to distinguish good from bad; fail to distinguish right and wrong; make no distinction between right and wrong;

*shi fei bu ming* 是非不明 have no sense of right or wrong; unable to tell right from wrong;

*shi fei dian dao* 是非颠倒 confound right and wrong;

For entries in the reverse order, the emphasis is on the key characters *shi* 是 and *fei* 非 of the multiple-character entry *shi fei* 是非. Regardless of the positions of *shi* 是 and *fei* 非, entries with *shi* 是 and *fei* 非 are semantically related to the multiple-character entry *shi fei* 是非, and they are arranged in Hanyu Pinyin Romanization according to the tone of the first character in an entry. The following are examples of the multiple-character entry *shi fei* 是非 in reverse order with variations in the positions of the key characters of *shi* 是 and *fei* 非.

(1) *shi fei* 是非 (*A B*)

*ban nong shi fei* 搬弄是非 carry tales; create troubles and dissensions; indulge in tittle-
tattle; make mischief; make mischief through tittle-tattle; sow discord through
gossip; stir up trouble by gossip; tell tales;

(2) _shi_ 是 _fei_ 非 ( _A_ _B_ )

_ jin shi zuo fei_ 今是昨非 come to realize how wrong one has been all these years;
realize how one has been wrong; wake up to one’s past folly after realizing what
is right today; things of the present are right and those of the past are wrong;

(3) _shi fei_ 是非 ( _A_ _B_ )

_ re shi fei_ 誘是非 incur unnecessary trouble; provoke a dispute; stir up ill will; stir up
trouble;

(4) _shi shi fei fei_ 是是非非 (AABB)

_ shi shi fei fei_ 是是非非 gossips; scandals;

Other examples that could be cited include:

(5) _B_ _A_

_[tu li 图利]_ desire to make money; plan to make money;
唯利是图 be bent solely on profit; be interested only in personal gain; blind to all but
one’s own interest; care solely for profit; have an eye to the main chance; plan only how
to get money; pursue profit as one’s only aim; put profit-making first; scheme after
nothing but gain; seek nothing but profits; seek only profit; intent on nothing but profit;

This keyword approach to entry treatment is useful in dealing with the following variations.

(1) Variations in an Idiom

Variations in the positions of the keywords in an idiom will not result in the separate
appearances of these entries as the meanings of the expressions are the same. _Da yu qing pen_大
雨傾盆, for example, means: “it is raining like billy-o; it rains cats and dogs; rain
pitchforks; the rain falls in sheets; the rain is pelting; the rain teems down”. This idiom can
also be expressed inversely as _qing pen da yu_傾盤大雨, or “torrential rain”. Another example
is _bu shi yi ding_不識一丁, which means “illiterate; not to know a single word”. This
expression can also be written as _yi ding bu shi_一丁不識: “not know B from a battledore; not
know beans; not know chalk from cheese; not know one’s ABC”.

(2) Variations in the Positions of the Key Words in Common Expressions

With this semantics-oriented approach, expressions with the same keywords are listed under
the relevant entries. Take _da xiao_大小 under the headword _da_ 大 as an example. If _A_ is _da_ 大,
_B_ is _xiao_ 小, then the normal sequence for _da xiao_ 大小 as a compound expression can be
_A_ _B_, such as _da xiao bu yi_大小不一 (not uniform in size); _da xiao he shi_大小合適 (size to
fit); _da xiao shi zhong_大小適中 (moderate size; right size); _da xiao xiang tong_大小相同 (of
uniform size); and _da xiao you zhi_大小由之(can fit any size).

Entries in the reverse-order sequence of _da xiao_ 大小 (AB) can have more variations.

(a) _A_ _B_ _

_da chun xiao ci_ 大醇小疵 sound on the whole though defective in details; with great
purity and small flaw;

da _jing xiao guai_ 大驚小怪 bark at the noon; be surprised at sth normal; get excited
over a little thing; great alarm at a little bogey; make a fuss about nothing; make a
great ado over sth; make a rare fuss over sth; much cry and little wool;
da _ti xiao zuo_ 大題小做 little about a major issue; make little of; treat major issues
light;

(b) _A_ _B_ _

_hua da wei xiao_ 化大為小 turn big issues into small ones;
_ke de xiao_ 可大可小 changeable; elastic;
_mei da mei xiao_ 没大沒小 impertinent; impolite to an elder; imprudent; show no
respect for one’s elders;
yi da qi xiao 以大欺小 bully the weak for being strong; the big bullies the small;
qi da ba xiao 七大八小 objects of various sizes thrown together;
qi da jiu xiao 棄大就小 exchange the great for the small; leave big shots alone and go
for the small fry;
yi da ya xiao 以大壓 小 the big coerce the small;
yi da yi xiao 以大易小 exchange the great for the small;
you da you xiao 由大由小 may be large or small;
yi da ya xiao 以大壓 小 the big coerce the small;
yi da yi xiao 以大易小 exchange the great for the small;
you da you xiao 由大由小 may be large or small;

(c) AB
yuan da xiao 原大小 life-size;

(d) _ AB
ge zhong da xiao 各種大小 every shape and size;
sha jiao da xiao 夾角大小 corner dimension;
mo li da xiao 磨粒大小 abrasive grain sizes;
qi shi da xiao 氣室大小 air cell size;
shi wu da xiao 實物大小 life-size;
zen ren da xiao 真人大小 life-size;

(e) _ _ A B
chuang hu de da xiao 窗戶的大小 size of a window;
mao xi guan da xiao 毛細管大小 capillary dimension;

(f) _ B _ A
qi xiao jiu da 棄小就大 lose a fly to catch a trout; one has to make sacrifices in order
to succeed;
wei xiao shi da 為小失大 lose a lot to save a little; lose a pound in trying to save a
penny;
yi xiao bo da 以小博大 throw out a spar to catch a herring;
yin xiao shi da 因小失大 lose a big opportunity because of a trifle consideration; lose a
great deal through trying to save a little; lose much because of a small thing; lose
the greater for the less; lose the main goal because of small gains; pay too big a
price for mere trifles; penny-wise, pound-foolish; suffer a big loss for a little gain;
try to save a little only to lose a lot;
you xiao bian da 由小變大 change from being minor to being major; from weak to
strong; grow from small to big;
you xiao dao da 由小到大 grow big from being small; grow from small beginnings
into a mighty force; grow from small to big; grow in size;

(g) _ B _ A
jian xiao bu jian da 見小不見大 fail to see the wood for the trees; strain at a gnat and
swallow a camel;

Another example is the sub-entry che ma 車馬 (chariots and horses), whose reverse-order
entries are as follows:
bi che nu ma 弊車駑馬 a decrepit cart drawn by a lean horse;
che shui ma long 車水馬龍 be crowded with people and vehicles; be thronged with
visitors; endless stream of carriages and horses; heavy flow of traffic; heavy
traffic; incessant stream of horses and carriages;
qian che wan ma 千車萬馬 a thousand coaches and ten thousand horses;
su che bai ma 素車白馬 plain cars and white horses － in a funeral procession;
xian che nu ma 鮮車怒馬 lead a luxurious life; new carriage driven by fat horses －
lavish service;
xiang che bao ma 香車寶馬 fragrant carriage and precious horse －the beautiful
carriage of women;

(3) The Treatment of Sentence Patterns
This keyword approach allows us to treat sentence patterns in a systematic manner. The main idea is to identify the keywords in a sentence pattern and give their common collocates to form entries.

\( \text{dui ... biao shi 对...表示} (\text{show...to...}) \)
\( \text{dui ... biao shi fan dui 对...表示反对 declare oneself against...;} \) show one’s objection to...;
\( \text{dui...biao shi tong qing 对...表示同情 express sympathy with...;} \) show sympathy toward...;
\( \text{dui ... biao shi wei wen 对...表示慰問 extend one’s sympathy to...;} \)
\( \text{dui ... biao shi yi han 对...表示遺憾 proffer regret at...;} \)

\( \text{cong...dao...從...到... (from...to...)} \)
\( \text{cong gu dao jin 從古到今 from ancient to modern times;} \)
\( \text{cong ri chu dao ri luo 從日出到日落 from dawn till dusk;} \)
\( \text{cong sheng dao si 從生到死 from the cradle to the grave;} \)
\( \text{cong tou dao jiao 從頭到腳 from head to foot;} \)
\( \text{cong tou dao wei 從頭到尾 from beginning to end; from first to last;} \)
\( \text{cong wu dao you 從無到有 grow out of nothing;} \)
\( \text{cong xiao dao da 從小到大 develop gradually; expand from small to big;} \)
\( \text{cong zao dao wan 從早到晚 from dawn to dusk; from morning till night;} \)

\( (4) \text{ The Treatment of Set Expressions with a Certain Pattern} \)

This refers to set expressions with key words in a relatively fixed pattern, such as \( \text{bu...bu...不...不...} \) (neither...nor...; not...or...). Expressions with \( \text{bu...bu...不...不...} \) include the following:

\( \text{bu bei bu kang 不卑不亢 neither cringing nor arrogant; neither haughty nor humble; } \)
\( \text{neither haughty nor pushy; neither servile nor overbearing;} \)
\( \text{bu fei bu fa 不悱不發 will not explain to sb not determined to learn;} \)
\( \text{bu fen bu qi 不憤不啟 would not explain unless one is desperately anxious to learn;} \)
\( \text{bu jiao bu zao 不驕不躁 free from arrogance and rashness; guard against self-conceit } \)
\( \text{and rashness; not proud or touchy;} \)
\( \text{bu kang bu bei 不亢不卑 in a happy medium between pride and humility; neither } \)
\( \text{haughty nor humble; neither overbearing nor servile; neither proud nor humble; } \)
\( \text{neither supercilious nor obsequious;} \)
\( \text{bu lun bu lei 不倫不類 (1) beyond standard; neither fish nor fowl; nondescript; (2) } \)
\( \text{without sense or order;} \)
\( \text{bu pian bu dang 不偏不倚 fair to all; without wavering to one side or the other;} \)
\( \text{bu pian bu yi 不偏不倚 avoid leaning to either side; even-handed; free from any bias;} \)
\( \text{hold the scales even; impartial; not to take sides; not to throw one’s weight either way; } \)
\( \text{partial to one; show no partiality to either side; unbiased; without partiality;} \)

\( (5) \text{ The Treatment of Infixes} \)

This keyword approach allows entries with infixes to be placed in the same heading, such as \( \text{bu zhu 不住} \) do sth well, and \( \text{做不好} \) cannot do it well.

\( (6) \text{ The Creation of New Sub-entries for Easy Consultation} \)

For easy consultation, it is necessary to create new sub-entries to put entries with the same key words under the relevant headwords. Take \( \text{bu zhu 不住} \) as an example. \( \text{不住} \) can be defined as “ceaselessly; continuously;” it can also be translated as “cannot”. The following are examples of the new entry \( \text{bu zhu 不住} \) when used as a suffix.
Another example of entry creation is *qi ba* 七八 ((1) seven or eight; (2) seventy or eighty per cent; (3) probably), which illustrates to some extent the use of numerals in some set expressions. The following are examples:

- *qi ba cheng* 七八成 (1) seventy or eighty per cent; (2) extremely likely; possibly; probably;
- *qi ba yue de nan gua – pi luo xin bu luo* 七八月的南瓜—皮老心不老 pumpkins in the seventh or eighth month of the lunar year — (fig) the skin is old but the heart is young; (fig) old in age but young at heart;
- *jia qi jia ba* 夹七夹八 at random; cluttered; confused; incoherent; talk incoherently;
- *luan qi ba zao* 亂七八糟 a fine kettle of fish; a glorious mess; a nice kettle of fish; a pretty kettle of fish; all in a tangle; at sixes and sevens; chaotic; higgledy-piggledy; in a clutter; in a little; in a mess; in a muddle; in a pickle; in a state; in an awful mess; in complete confusion; in wild disorder; jumbled; make a mess of sth; make hay of; rough-and-tumble; topsy-turvy; upside down;
- *luan qi ba zao de higgledy-piggledy*;
- *luan qi ba zao shuo yi tong* 亂七八糟說一通 make a chaos of utterances;
- *fang jian luan qi ba zao* 房間亂七八糟 one’s room is a dump; one’s room is a mess; one’s room is topsy-turvy;
- *gao de luan qi ba zao* 搞得亂七八糟 make a mess;
- *jiang ji hua gao de luan qi ba zao* 將計劃搞得亂七八糟 mess up one’s plans; throw one’s plans into confusion;
- *xin li luan qi ba zao* 心裏亂七八糟 feel all hot and bothered; feel very perturbed;

Holism in Chinese-English Dictionaries

It goes without saying that all Chinese-English bilingual dictionaries on the market, including the popular ones such as those compiled by Hui Yu 惠宇 (《新世紀漢英大詞典》(A New Century Chinese-English Dictionary) (2003)) and Wu Guanghua 吳光華 (《漢英綜合大辭典》(A Comprehensive Chinese-English Dictionary) (2004)), do not truly present the headwords in a context large enough to see how these single-character entries combine with other single or multiple characters to form character combinations that are semantically related to the headwords. In other words, the headwords and their sub-entries do not give a general background to understand the lexical, semantic, and cultural aspects of the entries. It is therefore necessary to treat headwords in bilingual dictionaries in a holistic manner.
This dictionary addresses the issue of holism through the arrangement of the single-character headwords and multiple-character sub-entries in both normal and reverse sequences. This dictionary may not be the first dictionary to put entries in reverse order (Yu et al 1986; Cao 2002), but it is probably the first dictionary to use the reverse order at multiple linguistic levels to allow the incorporation of pre- / post-modifiers, collocates, and measure words.

For single-character or headword entries, entries in the reverse order are listed after the last multiple-character entries, which are arranged in normal sequence. Taking the entry of *pai* 排 as an example, its definitions are as follows:

**pai** 排
1. arrange; arrange in order; put in order; sequence
2. line; rank; row
3. a clip of; a line of; a rank of; a row of
4. platoon; rehearse; raft
5. blow down; blow off; discharge; drain off
6. discriminate against; eject; exclude; get rid of; reject; repel; shut out;

The entries of *pai* 排 in the normal sequence are as follows:

- **pai ban** 排班
  1. fall by rank; fall in line
  2. arrange turns of work

- **pai zi** 排字
  1. compose; typeset; typewrite
  2. compositor; typesetter

- **pai zi gong** 排字工
  1. compositor; typesetter

- **pai zi gong ren** 排字工人
  1. compositor

- **pai zi ji** 排字機
  1. composing machine; compositor; typesetter; typesetting machine

- **she ying pai zi ji** 撮影排字機
  1. photographic typesetter

- **yi tai pai zi ji** 一台排字機
  1. a typesetting machine

- **pai zi jia** 排字架
  1. composing frame

- **pai zi pan** 排字盤
  1. composing stick

- **dian nao pai zi** 電腦排字
  1. computerized typesetting

The entries in the reverse sequence are as follows:

- **ai pai** 挨排
  1. arrange; make arrangements for

- **an pai** 安排
  1. arrange; arrange for; arrangements; cuddle up; find a place for sth; fix
  2. lay on sth; lay out; make arrangements for; manage; plan; tee up

- **bian pai** 編排
  1. arrange; lay out; make up sth

- **bing pai** 冰排
  1. ice floe; ice raft

Multiple-character sub-entries are also arranged in both normal and reverse sequences, such as *xiao hua* 消化.

- **xiao hua dao chu xue** 消化道出血
  1. alimentary canal haemorrhage

- **xiao hua li** 消化力
  1. digestion

- **xiao hua liang hao** 消化良好
  1. eurpsia

- **xiao hua qi guan** 消化器官
  1. digestive organs

- **xiao hua zheng chang** 消化正常
  1. eupepsia

- **xiao hua ye** 消化液
  1. digestive juice

- **jian xing xiao hua** 鹼性消化
  1. alkaline digestion

- **ren gong xiao hua** 人工消化
  1. artificial digestion

- **suan xing xiao hua** 酸性消化
  1. acid digestion

- **xi jun xiao hua** 細菌消化
  1. bacterial digestion

- **xu yang xiao hua** 好氧消化
  1. aerobic digestion

- **yan yang xiao hua** 厌氧消化
  1. anaerobic digestion

This normal-reverse arrangement to show headwords in a holistic manner is fully justified.
First, characters are no longer understood in a unidirectional manner. Take the character da 答 as an example. If entries are arranged in a normal sequence, they would be as follows:

**da【答】**

1. answer; reply; (2) reciprocate; return a visit;
2. da an [答案] answer; key; solution;
3. ti gong da an 提供答案 furnish an answer; supply an answer;
4. zheng que da an 正确答案 correct answer;
5. da bai [答拜] pay a return visit; return a courtesy call;
6. da bian [答辩] reply in support of one’s ideas; reply to a charge;
7. da bian quan 答辩权 right of reply;
8. kou tou da bian 口頭答辩 verbal defense;
9. zuo kou tou da bian 做口頭答辩 make verbal defense;
10. da ci [答詞] answering speech; reply; speech in reply; thank-you speech;
11. da dui [答對] answer; reply;
12. da fu [答覆] answer; reply;
13. dian hua da fu 電話答覆 answer by telephone;
14. han hu de da fu 含糊的答覆 dubious answer;
15. jian duan de da fu 簡短的答覆 brief answer;
16. jin zao da fu 尽早答覆 make a reply as early as possible; reply at one’s earliest convenience;
17. ju li da fu 據理答覆 give a reasonable answer;
18. ken ding de da fu 肯定的答覆 affirmative answer; affirmative response;
19. kou tou da fu 口頭答覆 answer orally; give an answer by word of mouth; oral answer;
20. li ji da fu 立即答覆 immediate answer; prompt answer;
21. ma shang da fu 馬上答覆 make an answer straight off;
22. man yi de da fu 滿意的答覆 favourable answer;
23. ming que de da fu 明確的答覆 categorical answer; distinct answer;
24. mo hu liang ke de da fu 模糊兩可的答覆 ambiguous answer;
25. qi dai da fu 期待答覆 expect an answer;
26. que ding de da fu 確定的答覆 decided answer;
27. shan shuo qi ci de da fu 閃爍其詞的答覆 evasive answer;
28. shou dao da fu 收到答覆 get a reply; have a reply; receive a reply;
29. shu mian da fu 書面答覆 give an answer in writing;
30. yao qiu da fu 要求答覆 want an answer; request a reply;
31. zhan ding jie tie de da fu 斬釘截鐵的答覆 categorical answer;
32. da hua [答話] answer; reply;
33. da juan [答卷] answered examination paper;
34. wan cheng da juan 完成答卷 finish one’s answer to an examination paper;
35. da li [答禮] return a salute;
36. da lu [答錄] answer and record;
37. da lu ji 答錄機 recorder; recording machine;
38. duo sheng dao da lu ji 多聲道答錄機 multichannel broadcast recorder;
39. li ti sheng da lu ji 立體聲答錄機 stereo recorder;
40. da xie [答謝] acknowledge; express appreciation; reciprocate;
41. da yun [答允] undertake;

When entries in the reverse order are also listed, then the lexical contents of the headword 答 are greatly enhanced.
Second, this holistic approach to entry presentation provides information on the semantic hierarchy of a headword. More specifically, it shows both the superordinates 上義詞 and the hyponyms 下義詞.

For instance, gou 狗 is a superordinate. Entries relating to gou put in the normal sequence could be gou dan bao tian狗膽包天, gou er狗兒, gou fei狗吠, gou gou ying ying狗苟繡營, gou jiao狗叫, gou niang yang de狗娘養的, gou pa shi狗爬式, gou pi狗屁, gou shi狗屎, gou shui shi狗瘦主人羞, gou tou shu qi狗偷鼠竊, gou tui zi狗腿子, gou wo狗窩, gou xie狗血, gou yao狗咬, gou zai dui狗仔隊, gou zai zi狗崽子, gou zao狗蚤, gou zhi狗彘, and gou zui狗嘴.

With the addition of entries in the reverse order, a number of hyponyms are included, providing a large lexical context to understand the character gou.

- ba er gou 巴兒狗 (1) lapdog; (2) Pekingese; (3) flatterer; sycophant; toady;
- ba er gou 叭兒狗 (1) lapdog; (2) Pekingese;
- bai gou 白狗 white dog;
- ben gou 笨狗 big mastiff;
- feng gou 瘋狗 mad dog; rabid dog;
- ha ba gou 哈巴狗 (1) Pekingese; (2) sycophant; toady;
- hai gou 海狗 fur seal; ursine seal;
- kan men gou 看門狗 watchdog;
- la chang gou 腊腸狗 sausage dog;
- lang gou 狼狗 wolfhound;
- lie gou 獵狗 hound; hunting dog;
- lou shui gou 落水狗 dog in the water;
- xi shi gou 西施狗 shitzu;
- xiao gou 小狗 doggi;
- xiao ling gou 小靈狗 whippet;
- ye gou 野狗 wild dog;
- yi tiao gou 一條狗 a dog;
- yu gou 魚狗 kingfisher;
- zou gou 走狗 flunkey; lackey; running dog;

Another example is bu 布. As a headword, it has

- yi ceng bu 一层布 a ply of cloth; a thickness of cloth;
- yi kuai bu 一块布 a piece of cloth;
- yi kun bu 一箇布 a bale of cloth;
- yi ma bu 一碼布 a yard of cloth;
- yi pi bu 一匹布 a bolt of cloth; a roll of cloth;
When *bu* 布 is combined with *liao*料, we have *bu liao* 布料, which could have the following measure words:

- *bu liao* [布料] cloth;
- *yi duan bu liao* 一段布料 a length of cloth; a piece of cloth;
- *yi fu bu liao* 一幅布料 a breadth of cloth; a piece of cloth;

The display of measure words in individual entries does not mean that the measure words do not exist as independent entries.

Take *yi ceng* 一層 as an example.

- *yi ceng* 一層 (1) one floor; one story; (2) a stratum; (3) a bed of; a blanket of; a cloak of; a coat of; a curtain of; a deck of; a film of; a flake of; a floor of; a layer of; a level of; a mantle of; a ring of; a story; a story of; a veil of;

Fourth, the cultural aspects of a character can also be shown through the reverse-order arrangement.


- *ai shen*愛神 Amor; Cupid; Eros; God of Love; Venus;
- *an shen*安神 (1) calm one’s nerves; (2) relieve uneasiness of body and mind;
- *ao shen*怒神 Goddess of the Earth;
- *bai shen*拜神 worship gods;
- *cai shen*財神 (1) God of Wealth; (2) money-makers;
- *cao shen*操神 tax on one’s mind; bother; trouble;
- *chu shen*出神 abstractedness; abstraction; aphelxia; be lost in thought; in a trance; spellbound;
- *chuan shen*傳神 lifelike; vivid;
- *ding shen*定神 (1) collect oneself; compose oneself; pull oneself together; (2) collect one’s thoughts; concentrate one’s attention; take a grip on oneself;
- *fei shen*貴神 (1) may I trouble you to do sth; would you mind doing sth for me; (2) waste of energy;
- *fen shen*分神 give some attention to;
- *feng she*丰神 manner;
- *gui shen*鬼神 ghosts and gods; spirits; supernatural beings;
- *hai shen*海神 Neptune; Poseidon;
- *huo shen*火神 God of Fire;
- *jing shen*精神 (1) consciousness; dauber; gist; mind; spirit; (2) essence; gist; spirit; substance;
- *lao shen*勞神 a tax on one’s mind; bother; trouble;
liu shen 留神 careful; keep one’s eyes on the ball; keep one’s eyes peeled; look out for; look sharp; take care;
men shen 門神 door-god;
ing shen 凝神 with concentrated attention;
nü shen 女神 goddess;
qi shen 棄神 cultivate one’s mind; discipline one’s mind;
qing shen 青神 call up an evil spirit;
qiu shen 求神 beg the gods;
ru shen 人神 man and God;
ru shen 如神 like god;
ru shen 入神 captivated; deeply absorbed in; enthralled; entranced; spellbound; with ecstasy;
shan shen 山神 mountain deity;
shang shen 傷神 be nerve-racking; overtax one’s nerves;
shi shen 死神 a pale horse; dead; Mr Mose; old floorer; old man Mose; old Mr Grim; the great levelleer; the great whiper-in; the graim monarch; the Grim Reaper;
si shen 祀神 worship gods;
song shen 送神 send off the gods after the offering of sacrifices;
tao shen 淘神 bothersome; trying;
ti shen 提神 arouse; elate; give oneself a lift; refresh oneself; stimulate;
tian shen 天神 deity; god;
tiao shen 跳神 sorcerer’s dance in a trance;
tong shen 通神 capable of buying the gods;
wen shen 瘟神 God of Plague;
wu shen 巫神 sorcerer; wizard;
xin shen 心神 mind; mood; state of mind;
xing shen 形神 body and spirit;
xing shen 醒神 induce resuscitation;
xiong shen 凶神 demon; evil spirit;
xiong shen 兇神 demon; fiend;
yan shen 眼神 (1) expression in one’s eyes; gleams of the eyes; light; (2) eyesight;
yang shen 揳神 have mental relaxation; repose; rest to attain mental tranquility;
yi shen 一神 monotheistic;
yi shen 怡神 inspire peace and harmony in one’s mind;
yi shen 頤神 have a mental relaxation; rest one’s mind;
you shen 有神 (1) full of spirit; (2) miraculous;
zao shen 灶神 kitchen god;
zou shen 走神 absent-minded.

Conclusion
The concepts of semantic interconnectedness and holism are the two linchpins of this dictionary. They offer changes to what has been practised for two centuries in the world of Chinese-English dictionaries. It is hoped that this dictionary will herald a new trend in the compilation of printed Chinese-dictionaries that are both phonetics-based and semantics-based, offering an easy consultation method and a wealth of semantic information for users to know the linguistic and cultural connotations of words in a holistic context.
References


Goodrich, Chauncey (1891/1965) A Pocket Dictionary (Chinese-English) and Pekingese Syllabary, Hong Kong: Hong Kong University Press．


Williams, Samuel Wells (1874) A Syllabic Dictionary of the Chinese Language; Arranged according to the Wu-Fang Yuen Yin, with the Pronunciation of the Characters as Heard in Peking, Canton, Amoy, and Shanghai, Han Ying Yinfu 《漢英韵府》, Shanghai: American Mission Press, Available from: http://archive.org/details/syllabicdictiona00willrich．


Adaptability of the Bilingual Theory to Dictionary Making
And the Genesis of a New Kind Dictionary

Ali Aycan KOLUKISA
Nevşehir Hacı Bektaş Veli University, Turkey
akolukisa@nevsehir.edu.tr / kolukisa_aycan@hotmail.com

Abstract
Dictionaries have been good references for a long time. When learning a new language, they become much more important to the users of them. Since the innovation of West and Hornby’s Learner’s Dictionaries, the main characteristics and the methods used in Learner’s Dictionaries has remained almost the same. Actually these methods and techniques are the things what really make them the best in class. But it is not so easy to find a Learner’s Dictionary when you are learning a foreign language other than English. There are of course dictionaries in other languages, but most of them are bilingual. Some people find bilingual dictionaries very useful especially at the beginning of their learning journey. But as users increase their competency in the use of the foreign language, the more it becomes obvious that bilingual dictionaries have a number of salient handicaps. After a while, many users abandon bilingual dictionaries and prefer the use of monolingual dictionaries written in the target languages. This research tries to understand the problematic nature of bilingual dictionaries in a deeper sense and suggests a new method to solve the problems identified. In drawing upon theories of bilingualism, from Weinreich (1953) to Pavlenko (2009), a new method of dictionary making is developed, which, it is hoped, will start a new era for bilingual dictionaries.

Keywords: bilingual dictionary, bilingual theory, polysemic method, concept, user.

1. Introduction

Weinreich (1953: 9-10) introduces three different types of bilingualism. These are called in Weinreich as “A. Coordinative”, “B. Compound” and “C. Subordinative”. And he gives examples as follows. If we summarize Weinreich’s three types of bilingualism: A-type of bilingualism, which is called “Coordinative”, has two different concepts and two different signs in each language which are connected separately in a bilingual mind. B-type of bilingualism, called “Compound”, consists of one set of concepts equal to each other in two languages which can also be regarded as ‘one concept’, and two different signs in each language. And finally C type of bilingualism, called “Subordinative”, is composed of one concept, one sign in mother tongue, and another sign that is learned in the target language as an equivalent of the sign in the mother tongue. In Woutersen et al. (1994), Weinreich’s these tree types of bilingualism are modelled as seen in the adapted figure at the left. However in Woutersen “Coordinative” type is named as “Coordinate”.

Afterwards, Potter et al. (1984) proposed the “concept mediation” and “word association” model. And with the effect of the research seen in Potter et al. (1984), Kroll & Stewart (1994: 158) proposed “Reversed hierarchical model of lexical and conceptual representation in bilingual memory”, which represents those two models within the same model, not separately as seen in Potter et al. (1984). It can be said that this was a very salient theory to be able to understand a bilingual mind, and this theory also led Pavlenko (2009) to reach her own model. On the other hand, Kroll & Stewart (1994) suggests that
the bilingual mind keeps the lexical items separately however the concepts related to these lexical items are just stored in one place. As seen in the adapted figure at the left, the storage of L1 represents the lexical items in mother tongue and the storage of L2 represents the lexical items in target language. The biggest storage at the bottom represents the conceptual memory. As one can see from the figure, L1 and L2 are separate storages whereas the concepts are stored only in one room. Moreover it is assumed in the model that the lexical link from L2 to L1 is stronger than the lexical link from L1 to L2, and the link between L1 and conceptual memory is also regarded to be stronger than the link between L2 and conceptual memory (Kroll&Stewart, 1994: 158). The weaker link from L1 to L2 is shown in the figure by the broken line, to reflect the fact that during L2 learning, bilinguals associate every newly learned word in the L2 with its L1 translation equivalent (Heredia&Cieślicka, 2014: 27). And this also explains why the L2 to L1 link is stronger, because every L2 word is mapped onto its L1 equivalent (p.27). The link between L2 and the Concepts is also shown by the broken line in the figure, to be able to reflect the bilingual’s inability to directly access the concepts of the words stored in L2 (p.28).

However, these facts can only be valid at the beginning of the learning process of L2. Indeed, it is thought that it is not easy to apply this strength-degree mechanism of Kroll&Stewart (1994) to advanced bilinguals. Furthermore, it is not possible to think the same process as valid for all kinds of words. For cultural words for example, it would be very absurd to expect an exact translation. While learning Japanese, the words samurai and tsunami might be a good example of it, as stated in Kolukisa (2015a: 665-666).

On the other hand, in Pavlenko (2009) a modified version of Kroll&Stewart’s RHM is seen. The model is called as “Modified Hierarchical Model (shortly ‘MHM’)” in Pavlenko (2009: 147). There are differences between the models of Kroll&Stewart’s “Reversed Hierarchical Model (shortly ‘RHM’)” and MHM. We can briefly say that the most important part of Pavlenko’s modification is the separate room for shared concepts. Moreover, it is seen that “the distributed conceptual model” of De Groot (1992) and “The shared asymmetrical model” of Dong et al. (2005) have been the base of Pavlenko’s modification.

At the left side, an adaptation of Pavlenko’s Modified Hierarchical Model is seen. The most outstanding part of this model is the separation of the rooms for the categories. There is also a separate room for the shared categories. Thus it enables to separate the specific concepts according to the languages. It is also important to be able to show the process of vocabulary development in L2. About conceptual transfer of MHM, Pavlenko says:

’...distinguishing feature of the MHM is the recognition of the phenomenon of conceptual transfer, which in turn is predicated on the differentiation between semantic and conceptual levels of representation’. The term semantic representation refers here to the largely implicit knowledge of: (1) the mapping of words and concepts determining how and which particular concepts are expressed by a particular word via polysemy or metaphoric extension and (2) connections between words, which account for phenomena such as collocation, word association, synonymy and antonym.” (Pavlenko, 2009: 148-149).

In regards to specific language or cultural words, Pavlenko (2009, 147) also says “only one language may have the necessary word forms, while activation of lexical links in the other language would fail”. In this case, it is seen that the bilinguals prefer almost lexical borrowing or loan translation...
(Pavlenko, 2009: 147). She also gives the example of “chashka” in Russian, and “cup” in English, to be able to show the restructurering observed at the conceptual level (see. Pavlenko, 2009: 149). Furthermore, in Pavlenko’s original figure, the boundary of L2 specific category room is shown by the broken line to show the development process of the L2 categories.

The models mentioned above are useful in order to understand the bilingual mind, however, it is thought that their area of usage could possibly be expanded, if they are developed. In this paper, it is thought that after the required revisions and explanations about the polysemy, concept and word meaning, it will become possible to apply the Pavlenko’s model to Lexicography.

2. Polysemy, concept and the word meaning

The concept of polysemy relates to the phenomenon of a single word containing more than one meaning. In Carter & McCarthy (1988: 19) it is written that when Ogeden and Richards produced the Basic English word list in 1930, they thought that the core meaning of the words that they offered for the project, was more important and prior than the others. As seen in Ogeden and Richards, in most cases, the meaning of a word is divided into two. In short, a word was seen to have two meanings: “denotation” and “connotation”. In Akimoto (2002: 107) “denotation” is defined as a meaning which is regarded as common by a great number of people, and “connotation” is defined as the secondary meaning that is identified by the culture and individuals. In Cowie (2009: 64) also it is said that “denotation constitutes the basic conceptual level of meaning, connotation the equally vital associative level”. And for the term “connotation”, Jackson (1988: 58) says “connotation relates to the associations that a word has over and above its denotation”. So from these descriptions it is possible to think “denotation” as a “core meaning” of a word, and “connotation” as the meanings which stratify over that core meaning. We will call all these connotations that stratify over the core meaning as “expanding meanings” in this paper.

Demirici (2014: 198) says, when a word has come into existence, the meaning that it represents is considered to be its real or basic meaning, and this real or basic meaning can be found at the top or fist rank of the definitions in dictionaries. After calling them as core and expanding meanings in this paper, we can briefly say that core meaning is generally represented by ① and found at the first rank, whereas expanding meanings are shown by ②, ③, ④, ⑤, etc. in dictionaries and found at previous ranks as seen in figure 1.

In the book of “Key Terms of Semantics”, concepts are defined as mental representations of knowledge about categories of entities and experiences (Murphy & Koskela, 2010). And as an example elephant is given:

“The concept ELEPHANT incorporates information about the size, body shape, colour and typical habitats of elephants. The concept allows us to identify and categorize things in the world as elephants, label them by the linguistic form elephant and make INFERENCES about them” (Murphy & Koskela, 2010: loc.373).

From the description, it is possible to say that concepts allow us to recognize the real world around us. And also in Jackendoff (2012: 73) language is regarded as a system that links concepts and thoughts with
pronunciations. So it means that the words belong to real world whereas the concepts belong to our mental world. Murphy G.L. (2002: 385) also says, “By concept, I mean a nonlinguistic psychological representation of a class of entities in the world”. It can be said that our approach in this paper and the descriptions seen in Murphy G.L. (2002) are very alike. It is known through experiments that while we are learning our mother language, it is possible to build up concepts first even though we do not have the words to label at early ages (see also, Bloom, 2001: 1101). To be able to reach the same conclusion it might be felicitous to make an inference from our experiences of seeing toddlers giving different sound-sets, or words in toddlish, to label the particular objects around themselves just before they start to speak. If we had learned the words first, then labeling the objects differently in toddlish would not be possible. However this can be said only for the early stages of our first language acquisition process. Because, when we grow older, we can hear and memorize the words before even an exact existence of concepts. We can even keep them in our minds for a later reference or to make a future combination with a concept. Magical spells that have no meaning might be a good example of it.

Figure 2, illustrates the concept. However, different to the previous researches (see. Pavenko, 2009: 151), figure 2 is the illustration of a monosemic concept. In this paper, two different types of concepts are proposed. These are “monosemic” and “polysemic” concepts, representing the “monosemous” and “polysemous” words. In order from top to bottom, Figure 2 represents the external part, internal part and the inner side (i.e. the ‘content’) of a monosemic concept. It is thought that, every concept is like a container and the core-part is always filled with mental-liquid like water which symbolized the mental equivalent of what it represents in the real world. So, for a monosemic concept, as long as it represents only one thing, it has only one core-part always filled with water. I do think that concepts in our minds are three dimensional objects as illustrated in figure 2, whereas the lexical meanings of words have only two dimensions. Although we symbolize the concepts in our minds as cylindrical containers, it is just a symbolization to be able to make a generalization to explain how the process works. On the other hand, it is not actually possible to say that every concept is the same in shape. Figure 2 is only an illustration of a monosemic concept. However, most words in the real world have polysemy.

According to this paper, just like the words in real world, it is thought that also the concepts have polysemic structure as seen in figure 3 below. Figure 3 is the illustration which shows the inside of a polysemic concept. It is suggested that the core part of a concept is always filled with water for its identity, however the other parts are ready to be filled with water as seen in figure 3. In other words, the core part of a concept always stays filled with water what shows the unmarked and most admitted shape of that concept just like the reflection of the core meaning of a word. However, the water inside the core part of a concept is unlimited, so that it can fill any other part of that concept. When a polysemous word is used under a real speech situation in real word or in a context, then the selection is made and the selected part of the concept is filled with water. When this happened, also the core part of the concept stays filled with water, because, it is thought that every expanding area gets its origin from the core meaning (see. Kolukisa, 2015b: 153). There is another illustration of this process at the left in figure 4. Figure 4 shows just what we explained here. You may also wonder how a polysemic concept is linked to a polysemous word and how this mechanism works. We will explain it with the help of figure 5 below. However, briefly we can say that our approach is ‘cognitional’ rather than directly ‘cognitive’. This is due to the fact that, in some aspects our approach is very close to cognitive linguistics, whereas in some parts it is very close to ‘psychology’ and ‘psycholinguistics’. In this paper it is proposed that when the
words are used under a real speech situation meanings of the words become clear according to the function of the model that I call “Access and Fetch”. Access and Fetch Model (shortly ‘AFM’) is originally created and explained here with the help of figure 5 below, to be able to make the relations between concept and word clear.

Murphy G.L. (2002: 389) says that there is a one-to-one mapping between words and concepts, and he figures it as: “Word 1 <====> Concept 1”, “Word 2 <====> Concept 2”, “Word 3 <====> Concept 3”, “Word 4 <====> Concept 4”, “Word N <====> Concept N”. We can briefly say that our approach in this paper is very alike. However, our explanation about how the concepts and the words are combined when the polysemy has taken place, is different. Figure 5 explains how the concepts and words are combined. The model is created to demonstrate the polysemy seen at both lexical and conceptual level. Unlike the present studies, it is thought that polysemy is not a phenomenon seen only at lexical level. It is claimed out in this paper that it also occurs at conceptual level. As seen in figure 5, every expanding meaning is a layer constructed over the core meaning. We named the core area here with the letter “X” and the other expanding areas stratifying over the core meaning as “x+a1”, “x+a1”, “x+a1”, “x+a1” as explained in Kolukisa (2015b: 153). “X” is just a symbol and may change to “Y”, “W”, or etc. as the words and correspondingly the concepts change. On the other hand, although Kolukisa (2015b: 153) explains the polysemy seen at both conceptual and lexical level, there is not an exact explanation of how these two are related and work. In this paper, it is thought that we can explain the relations and procedures seen at both conceptual and lexical levels thoroughly by using AFM. In this model, the word itself actually is just like an empty dummy container, as long as it is not in use. And when we use the word in sentence or in a real speech situation than just like the process seen at phoneme restoration effect in psycholinguistics, it is thought we just send an access signal to our mind, (lit. to ‘conceptual memory’) so that to be able to find and fetch the suitable meaning of a polysemous word. In other words, when we hear or start to utter a word, our minds find the concepts of these words first and work like a computer by analyzing and selecting the area of each concept which is to be filled with water according to the context or situation. And after the selected area of the concept is filled with water, then the same signal is loaded up with the corresponding area code of the polysemous word. Thus the meaning has become selected and coded at lexical level. For instance, before using the word “dream”, we cannot know which area is active, unless it is used under a real speech situation or in a sentence. Let’s see how it really works with the help of word “dream”.

Firstly we should have a look at the dictionary to be able to see how this word is explained in own mother language. An explanation from the The American Heritage Dictionary for the word “dream” is seen in figure 6. However, as a matter of convenience, we will handle only the noun form of it here. In figure 6, the core meaning is symbolized with “X” and
Now, suppose that we’ve just used the word “dream” in a sentence as follows: “Being a doctor was my dream”. When we hear this sentence, a signal is sent to our mind first to detect which words are in use, and in what combination they are used. Then the Access and Fetch procedure takes place. It actually starts and finishes in a very short time, so that we can easily understand what the context really is. Here let’s focus on just the word “dream”. The process is illustrated in Figure 7 below.

In Figure 7, the Access and Fetch mechanism is illustrated when it is in use. After the signal is sent, our mind starts to analyze all the areas of the word “dream” to be able to find the most suitable one for the context or speech situation. At this point, we generally refer to all the meaning areas of that word, some of which are even not included in dictionaries. Then after making an elimination, just the suitable area of the concept is filled with water. Thus a ‘fetch’ signal is sent back from the filled area of the concept to the equivalent meaning area of the word “dream”, correspondingly. As a result of this process, only the meaning “an ambition; aspiration” has become selected. So, the Access and Fetch – procedure becomes completed and then we can say that the word “dream” is Accessed and Fetched.

Actually when we use a polysemous word in our daily life, our brain tries to apply every area to the context at the background, to be able to find the most suitable one for the situation, however we do not realize it for it is a very fast process. We can also say that the connections with other words are very identificatory on finding the suitable meaning area of a polysemous word. From this point of view, it is supposed that our brain does not need to make a very heavy process when we use the collocations of the words. Because, only one fixed meaning area of a word is used when we use a collocation.

3. Monosemic method -in use today, vs polysemic method -for the future of bilingual dictionaries

Only in a very few research the terms “Monosemic Method” and “Polysemic Method” are being used. For example, the term “Monosemic Method” is seen in studies such as Gobbo (2009: 98), Miháltz&Prószyk (2004: 176); and the term “Polysemic Method” in Kornai (2008: 488). However, our approach to these terms would be a little bit different than the previous studies, although there are also some points in common (see. Katz&Fodor, 1963; and Kornai, 2008). The approach of this paper to these terms are explained separately in detail below.

3.1. Monosemic method

Monosemic method discussed here is the method of giving the equivalent of a headword in a target language. This method is seen almost in every bilingual dictionary of today. The method is preferred and used for example, while explaining a headword of L1 in L2, or conversely, while explaining a headword of L2 in L1. Every headword is
explained with its equivalent translation in the target language. As an example, we can give the noun forms of the word “oxygen” in English for L1, and the word “kislorod (кислород)” in Russian for L2. The monosemic method used in a bilingual dictionary is symbolized in figure 8. In figure 8, the English word “oxygen” is equalized to Russian equivalent “kislorod”. Making “one-to-one” matching of L1 and L2 is a very simple method used in bilingual dictionaries of today. Actually we should admit that this method works very well if the words in L1 and L2 are both monosemous. If you look up the word “oxygen” in a monolingual dictionary such as The American Heritage Dictionary, Concise Oxford English Dictionary or Oxford Dictionary of English (2nd Ed.), etc., you will probably find only one definition for the term “oxygen”. And this will be almost the same if you look up the word “kislorod (кислород)” in a Russian monolingual dictionary. So it could be roughly said that, Monosemic Method is useful while explaining monosemous words. It might be easy to find monosemous words if you are looking up a word in a dictionary of special terms of a specific field. On the other hand, it is not very easy to find monosemous words in a normal monolingual, and also in a bilingual dictionary. Even the word “oxygen” has more than one meaning if you look it up in Meriam-Webster Dictionary or OED Online. For the real situation at present, Monosemic Method explained in this paper is almost being used to explain every word in bilingual dictionaries, however, the biggest problem of this method is the polysemy of the words seen at definitions. As stated in Kolukisa (2015b: 148) “although bilingual dictionaries of today have found some solutions to the polysemy of the headwords, they seem to have ignored the same polysemic problem when it is seen at the definition words” (Kolukisa, 2015b: 148). In figure 9, Monosemic Method Used in Bilingual Dictionaries is illustrated. Monosemic Method could be effective to equalize the monosemous words in L1 and L2, however, when the same method is used for explaining the headwords with polysemous words in the definitions, it causes some problems and conceptual deviations.

3.2. Polysemic method

Polysemic Method that we will discuss here is a new method for bilingual dictionaries. This method is seen first in Kolukisa (2015b), however, there is no special term in Kolukisa (2015b) to name this method. Polysemic Method called in this paper, is a special method proposed for the polysemous words used at definitions. It is main aim is to prevent the conceptual deviations and ambiguities, while making an equalization between the headwords and the target words. For example, suppose that we look up the adjective form of the word ‘pretty’ in Kenkyusha’s English – Japanese Dictionary For The General Reader (3rd Ed.), you will probably find definitions as seen in figure at the left. According to this bilingual dictionary, the word ‘pretty’ has four different
meanings in Japanese. Now let’s look at the first core meaning in 1a. According to 1a ‘pretty’ is equal to the Japanese words seen below. And at the first rank of 1a, ‘Pretty’ is equalized to Japanese word Kirei (na)’.

Pretty 1a. = Kirei (na)
Pretty 1a. = Kawaii
Pretty 1a. = Karen (na)
Pretty 1a. = Kogirei (na)
Pretty 1a. = Utsukushii

question here is, “Do you really think that we can use ‘pretty’ everywhere that we can use ‘kirei’ in Japanese?” The bilingual dictionary says they are equal, meaning we should be able to use them as direct equivalents. Now, at this point, let’s have a closer look at the definitions of the word ‘kirei’. It is better to see how the word ‘kirei’ is treated in a monolingual dictionary of Japanese. In the dictionary of Digital Daijisen the word ‘kirei’ is defined with seven different meaning areas.

So, when a monolingual method is used in a bilingual dictionary, although it is not realized very clearly till now, the equalization of L1 and L2 actually has been made as in figure 10. Then the word ‘pretty’ can be seen to have seven meanings at once. However, the word ‘pretty’ doesn’t have so many meanings in English. A similar figure is also seen in Kolukisa (2015b: 156), to be able to show the conceptual deviations between the Turkish word ‘güzél’ and the Japanese word ‘kirei’. As understood from the figure 10, although there are some overlapping meaning areas inside the Japanese word ‘kirei’, we cannot say that the whole parts of ‘kirei’ are overlapping. The meaning areas “x+a3,” “x+a4,” “x+a5,” and “x+a6” do not match with the meaning of ‘pretty’. However, while making an equalization with the monosemic method in a bilingual dictionary, it is not possible to show both the overlapping and non-overlapping parts separately. Moreover, it is found out that most of the missuses of the target words are caused by the non-overlapping meaning areas of the words in definitions (see. Kolukisa, 2013). On the other hand, it may be that the editors and the writers of the bilingual dictionaries should have realized this handicap, because in almost every dictionary, and especially in the bilingual ones, there are
lots of sample sentences to be able to show the correct usage of these words. However, showing and giving examples only for the overlapping parts do not explain the non-overlapping parts. Here to be able to explain Polysemic Method, we should first see how we make an equalization while we are learning a new vocabulary in target language by the help of a bilingual dictionary.

In Figure 11, we see the word ‘pretty’ explained in the dictionary and at the right of the figure the meaning areas that we arranged according to the dictionary definitions are being monitored. Figure 12 shows the equalization at the lexical level and also the overlapping and non-overlapping parts of the words ‘pretty’ and ‘kirei’. As understood from Figure 12, non-overlapping parts are much more than the overlapping parts. However, dictionaries of today only are capable of only showing the overlapping parts.

Now let’s think whether there might be a way to be able to show both the overlapping and non-overlapping parts of L1 and L2. It is thought that an adaptation of Pavlenko’s MHM might be very helpful to be able to solve this problem.
In Figure 13, you see the adaptation of the Modified Hierarchical Model seen in Pavlenko (2009: 147), for the vocabulary learning through the bilingual dictionaries. As understood from the figure, it is thought that we actually make an equalization at the conceptual level first as seen in figure 13 and then we carry it to lexical level. In lexical level we just label the words. After the conceptual equalization process, every L2 word in our lexical memory is actually labelled as ‘L1 Equalized’, however not all the parts of the meaning areas. Because, as explained above and in figure 13, all the meaning areas of the two words do not overlap. So it is thought that we label the overlapping parts of an L2 word as ‘L1 Equalized’ and keep it in ‘L2 Words’ room, and equally we also label the non-overlapping parts of that L2 word, however we keep these parts separately in ‘L2 Specific’ room. On the other hand, L1 words that completed this equalization process successfully, are also labelled as overlapping and non-overlapping and only the overlapping parts are kept in ‘L2 Equalized’ room. Moreover, L2 specific words, such as culture related ones, are labelled and kept in this ‘L2 Specific’ room. And L1 specific words, non-equalized L1 words and non-overlapping parts of the equalized words of L1 are also all labelled and kept in ‘L1 Words’ room. We do not need to create a specific room for these kinds of L1 words, for we do not make a separation when we acquire our first language. So, creating only an ‘L2 Equalized’ room in L1 would be enough.

Accordingly, it is thought in this paper that the bilingual dictionaries should show not only the overlapping parts, but also the non-overlapping parts of a concept, when an equalization is made between L1 and L2, or L2 and L1. This is what we call in this paper as Polysemic Method.

4. Polysemic method and the genesis of a new kind of dictionary

In the previous section, we explained how the bilingual theory might be adapted to bilingual dictionaries. Now, we can show perceptibly what the Polysemic Method really is. In Figure 14, you see the genesis of a new kind of dictionary through the application of this Polysemic Method. This new kind of dictionary shows not only the overlapping parts but also the non-overlapping parts of L1 and L2. The definitions of ‘kirei’ are quoted wholly again from Japanese monolingual dictionary Digital Daijisen (2014) deliberately, to make the method more understandable.
As one will understand from the figures that we explained up to now, our new kind of bilingual dictionary is capable of showing the non-overlapping parts that we explained in figure 12, as well as it is able to explain the overlapping parts. Non-overlapping meaning areas of the word ‘kirei’ are shown at the right top of the figure 14 with a “×” mark at the beginning. Moreover, these meaning areas are also explained with the possible equivalents instead, in the “Note” part of the definitions.

On the other hand, you may think that such kinds of information occupy too much space in the dictionary, and giving just the examples or the real usage of the overlapping parts would be enough. However, if you think that a person can create countless sentences, you will soon realize that just giving the examples of the overlapping parts won’t make any use or prevent the learners using that word with the effect of their own mother language, as demonstrated in Kolukisa (2015b, and 2013). So, to be able to prevent the misuses we need both methods and technics together; giving examples of both overlapping and non-overlapping parts, and also showing not only the overlapping parts but also the non-overlapping parts together in a bilingual dictionary. If you are worried about the sufficient space in dictionaries, there might be also some other economic ways of applying this Polysemic Method to bilingual dictionaries, as seen in figure 15 below. However, when we use this Polysemic Method in a bilingual dictionary, we should not forget that we have to apply this method to every definition word. Thus, we can provide an all-encompassing bilingual dictionary.

![Figure 15](image)

In Figure 15, you see another way of applying the Polysemic Method to bilingual dictionaries. This method is much more economic however, it lacks of the detailed explanations. As understood from the figure, some of the definitions may not have any non-overlapping parts.

5. Conclusion

Bilingual dictionaries of today try to teach the equivalents of the words of a language in another language. By doing this, they make an equalization of concepts which belong to different cultures and two different languages. The fact that that these concepts are formed in different cultures, means they are not as exactly as the same. So while making an equalization of two different words which belong to two different cultures, it is not possible to make a one-to-one mapping all the time for polysemous words. Dictionaries of today should have realized this problem long before, because they have produced some solutions such as giving lots of examples, usages and the collocations of the headwords with the explanation in the definitions. However, as a result of the technics and the methods used in these bilingual dictionaries, a very big handicap has come into existence in the way of their treatment to polysemy in definitions.

This paper proposes a new solution method for the handicap seen in bilingual dictionaries. The new method is developed by adapting the bilingual theory and the model seen in Pavlenko (2009), and it is named as Polysemic Method in this study. As explained above, it is not very easy to find monosemous words in dictionaries. However the bilingual dictionaries of today treat every definition word as if they all are monosemous. As a result of this, also as stated in Kolukisa (2015b), conceptual deviations really cause the misuse and the mother tongue oriented usages of the headwords. Such kinds of misuses have been found in the sentences of the Japanese learning students, as explained and presented in Kolukisa (2013). So as a solution to this handicap the Polysemic Method is created.

By using the Polysemic Method explained in this paper, it becomes possible to show both the overlapping and non-overlapping of the words in L1 and L2. It is believed that using this method in bilingual dictionaries, will become a revolution and help the users to understand the concepts of a new language much better. Finally it will enable the users to have a good command of vocabulary and
correct usages of the words in the target language. Thus the errors made by the users of the bilingual dictionaries will be prevent and as a further outcome hopefully it may also be a great use in translating programs.

References


Dictionaries


Bilingual Dictionary and Encyclopedia – Where goes the line?
Cultural specificities as sources of dilemmas and solutions proposed

Mats-Peter Sundström
European parliament, Swedish translation unit
matspeter.sundstrom@europarl.europa.eu

Abstract
This paper aims to shed light on the border zone between dictionaries and encyclopaedias, with regard to bilingual dictionaries. More specifically, the intention is to discuss the occasional desirability and necessity to introduce encyclopaedic elements into bilingual dictionaries. Irrelevant though this issue may seem at first glance, given the obvious fact that in bilingual dictionaries the headwords are defined through the equivalents they are supplied with, there are, nonetheless situations where comments of an encyclopaedic nature are called for, viz. 1) cases where the headwords refer to objects, practices, institutions etc. specific to the source language culture and hence lacking target-language equivalents; 2) cases where the source language has words or set expressions referring to universally known phenomena which, however, lack exact target language equivalents (as an illustration reference may be made to the Finnish verb tarjeta which means 'be able to endure low temperatures without feeling uncomfortable') and 3) cases where comments of a pragmatic nature, relating to the use of a particular lexeme may be called for (like for instance, although the Swedish noun toalett finds an semantic equivalent in the English toilet, the latter is often best avoided in social conversation, where words like restroom or bathroom provide more appropriate substitutes. With the advances made in electronic dictionary editing, possibilities for such comments are well-nigh exhaustible and this paper intends to present a few suggestions regarding their implementation.

Keywords: lexical anisomorphism, pragmatic information boxes

1. Introduction

During the earlier stages of a budget period, Grandfathering means allowing the holder of a given right, concession etc. continue having this right or concession, whereas gold plating (apart from its literal sense) is an ironic way of referring to excessive or unnecessary regulation and greenwashing means making an activity seem environmentally sound just for appearance's sake. Then there are multiword items like sunset clause (a provision to the effect that a regulation, system etc. will automatically cease to be in effect after a given period of time) or plain vanilla (as in plain vanilla solution i.e. a conventional solution, conforming to traditional patterns), just to mention a few.

Since the overwhelmingly predominant language of the international scene is English, it is only natural that these lexemes should mainly be confined to the English language. In these cases the non-English dictionary user may need for receptive purposes a kind of half-way house between a dictionary article and an encyclopaedic presentation, to allow for a proper understanding of the lexemes concerned.

3. Emotion-triggering phrases

Some of the notoriously trickiest lexical items to deal with are stock phrases of a kind I have chosen to call emotion-triggering phrases. As the description selected suggests, they have as their main objective, not so much to communicate a given factual content but rather to transfer an emotional reaction to the receiver or trigger a similar reaction in him. Often enough, they are typical of one given language and lack equivalents in any other language. Here, the bilingual lexicographer finds himself at his wit's end, unless a concise description of their actual use can be added. And so, as McCreary pointed out, "foreign language learners
[...] are likely to search for them in a bilingual dictionary a few times before being discouraged at the generally spotty coverage” (McCreary 2007:87). In other words they amply illustrate the point made by Ariel about "cases where ‘what is said’ is arguably richer than the encoded semantic meaning” (Ariel 2008: 295).

Below, four examples of the above-mentioned tendency will be provided, drawn from four different languages. The first will be taken from the Swedish language:

*Surt, sa’ räven om rönnbären*

This expression would translate mechanically into English as

*[They are] sour said the fox about the rowanberries.*

Here, any attempt at finding an exact English translation is bound to fail. We simply have to examine on what occasions the expression is used and then arrive, if not at an equivalent, then at least at an explanation. Consider the following context.

Well, you didn’t get to go to Paris, did you, and now you keep telling everyone that there is, after all, nothing much to see in that city. *Surt, sa’ räven om rönnbären’, huh?*

In other words, this expression is used to indicate the realisation that someone is not being sincere in assuring that a certain experience he could not have was not worth very much, after all. It comes close to being a statement like *Stop pretending, we all know you were yearning to do this or that thing and now you are just trying to hide that not succeeding in doing so bothers you something awfully!*

Something similar is true of the Finnish

*Ei kannettu vesi kaivossa pysy*

Literally, *water that is carried [to the place from elsewhere] will not stay in the well.*

Here, similarly, the whole is something vastly different than the sum of its parts. Obviously, it is not possible to create a wellspring by digging a hole in the ground and filling it with water brought there from elsewhere, as this water will be absorbed by the surrounding layers of earth, but why should there be any need to say so explicitly? – In fact, this saying carries a message of resignation and an element of *Well, I told you so, didn’t I*, used as it is to indicate that changes or reforms cannot be imposed from the outside if there is no intrinsic support or mental preparedness for them. Cf below:

*Well, well, well … They encouraged legislative amendments, they initiated training programmes for professional law enforcers and much more in order to fight corruption in X country but nothing came of it all. But of course: ei kannettu vesi kaivossa pysy!*

Or yet another illustration, this time from Polish.

*Hulaj dusza, piekla nie ma!*

Literally *merry goes life, [my] soul: there is no hell.*

This basically denotes a kind of reckless way of acting or living as if there were no tomorrow. Again, an illustration
So the X family inherited one million euros and then what? Well: hulaj dusza, piekła nie ma! In less than a year they had run through all their wealth and racked up a mountain of debt so now the banks are threatening them with foreclosure.

And finally and evidently, there are a great many English expressions also that fall within the same category of expressions that cannot be readily rendered into other languages. Consider this Get a life! – While assuredly thousands of, for instance, Finnish or Swedish speaking teenagers may have feel sorely tempted to fire off this phrase into their parents' very faces, they have not been able to, owing to the lack of an even remotely similar standard phrase in their own native languages. – According to Wikipedia, the phrase is used to describe a person who "is devoting an inordinate amount of time to trivial or hopeless matters" or "people who are viewed as officious or meddling in the affairs of others".

– In every one of the four cases quoted above, it is hardly possible to fulfill the role of a bilingual dictionary for productive purposes. The lexicographer will have to confine himself/herself to telling the dictionary users that the phrase in question is so specific to a given language as to thwart any attempt at seeking equivalents. This goes a long way towards confirming the relevance of Adamska-Salaciak's statement: "Failure to allow for non-natural equivalence manifests itself in the unrealistic expectation that equivalents are always there, merely awaiting discovery" (Adamska-Salaciak 2010:399). All this bears out the point made by Moeimam and Steinhauer to the effect that "For the receptive dictionary this is not much of a problem, but for the user of the productive dictionary such a description is only useful when [the lexicographer] has to explain the meaning of an item to a [target-language speaker]. As a translation equivalent, however, the circumscription will rarely be adequate" (Moeimam and Steinhauer 2007:288).

For the purposes of a dictionary for receptive purposes, however, modern computer-based editing offers well-nigh limitless possibilities. Dictionary articles in which either focalising words or emotion-triggering phrases occur can be supplied with boxed-in miniature fact sheets, where a concise explanation of their use is provided, followed with a number of attestations from actual usage. In so many words: supplied with pragmatics boxes, to use the term coined in the introduction. This solution was in fact proposed by the present author at the previous Asialex conference (cf. Sundström 2013:22). At this point modern technology offers us all but inexhaustible possibilities, as lexicographers no longer are bound by the constraints imposed by the printed format. Additional information on headwords in bilingual dictionaries can be added basically at will, as was indicated by Yamada "pragmatic information in electronic dictionaries is unlimited" (Yamada 2014). – Such information boxes or frames are used, for instance, in the general dictionary of the Swedish Academy (Svenska Akademiens Ordbok). This is a monolingual dictionary of the Swedish language, with not only descriptive but also normative functions. Technically, these boxes are rectangular spaces in immediate connection with the lexemes they refer to. Printed on a grey-coloured background they discuss given lexemes and add comments and recommendations about their usage. Something similar could no doubt also be introduced on a larger scale in bilingual dictionaries, notably those in electronic format.

This only remaining problem is that of making dictionary users aware of this potential. As was quite tellingly pointed out by Frankenberg-Garcia "little of the spectacular improvements that have taken place over the past decades have had much influence on actual dictionary-use behaviour" (Frankenberg-Garcia 2011:98). In this context, the systematic integration of dictionary-use into foreign language teaching is more than desirable. All the more so, since phrases of the kind discussed above most probably are far more common when the source and target languages are structurally and culturally far from one another, as is the case for instance with European and Asian languages.
Pragmatics boxes could be seen as a kind of cross between pragmatics information notes and extralinguistic comments. Needless to say, then can serve a range of other purposes also than just clarifying by contextualisation lexical items of the kind referred to in the above two sections. To the present author, essentially three such purposes present themselves, viz frequency information, warnings of possible double entendres and, finally, clues about cultural connotations.

4. Usage frequency information

To probably anyone with some experience of translation it is a well-known fact that there are expressions in language A that in principle lend themselves perfectly well to word-for-word translation into language B, but even so, such a translation, while grammatically and semantically correct, would seem unidiomatic or even unnatural. Translations are often rife with examples, something which goes a long way towards proving Sumiyoshi's statement to the effect that "valency patterns are to be fully covered in dictionaries" (Sumiyoshi 2013:273). At its simplest expression we may take the lexeme selon in French or według in Polish, both for all practical purposes corresponding to the English according to. Now, as native English speakers are instinctively aware of, this prepositional expression does not normally occur in connection with the first person pronoun, at least not in the singular. In other words then, a construction such as *according to me would appear unidiomatic, whereas such one-to-one equivalents as selon moi in French or według mnie in Polish would be perfectly normal. In fact, as a Swedish-language translation assistant at the European Parliament pointed out to me, if in a discussion someone said enligt min åsikt, i.e. ‘according to my opinion’, the utterance would appear to her as a perfectly appropriate comment, whereas enligt mig, i.e. ‘according to me’ would make the speaker come across as an unbearable know-it-all! (Jalava 2014). To inform French or Polish users of this particular aspect of English (and, if need be, Swedish) pragmatics, recourse could be had to a device commented on by Chan: "Another issue concerning learner's understanding of word use from dictionary information [...] is the inclusion of common mistakes boxes. These boxes are a useful means of highlighting common mistakes made by target users of the dictionary" (Chan 2012:87).

At this stage it should be emphasized that the idea of common mistaken need not be taken as a reference to grammatical errors, as seen above. Another example of this aspect is the overuse of certain words. For example, in European Union texts translated into Swedish the English verb promote is habitually rendered into the target language as främja something which from a purely semantic point of view is perfectly impeccable. However, in texts spontaneously generated in Swedish, the idea contained in the English promote would probably rather be expressed by a different construction, namely verka(arbeta) för, both literally corresponding to the English work for. Thus the notion of promote international understanding

if originally produced in Swedish, would most likely come out as verka(arbeta) för internationell förståelse i.e. in English work for international understanding.

Somewhat similar is the case with the English verb secure. While in contexts such as secure adequate health care it would be perfectly possibly to translate the core verb into Swedish by means of equivalents such as säkerställa (derived from säker meaning 'secure') or se till att (literally 'see to it that'), a native user of Swedish wishing to express the underlying idea would primarily be most likely to use the verb ordna med (literally *arrange with), as in
Vi ska ordna med adekvat hälsovård

... which in turn would come across in English back translation *We shall arrange for there to be adequate health care.* Although formally correct, it would not be very likely to occur, at least not in texts of the kind seen at the European Parliament. – Conversely, in the words of a translator colleague of mine, a former Swedish prime minister was once reproached for using lexical items too typical of the Swedish-language translations at the European Union and thus not native-like! (cf Wolmerud 2015).

An objection may of cause be raised, in so far as, on the surface level, the English *secure* seems to be invested with a stronger element of assurance ("we shall not only arrange it, we shall make sure it is there") but the question is whether in fact this point is of relevance. After all it may be assumed that in contexts of the kind quoted the Swedish *ordna* implies the same element of assuring something as the English verb *secure* carries explicitly. This fact might well be included as an assurance to linguistically perhaps somewhat insecure dictionary users, again in a "mistakes box", which far more appropriately might be called a "hints for practical language use box". – However, caution must be exercised in an actual translation situation, since it may be very difficult to justify the use of a more native-like construction that does not, in purely semantic terms, quite correspond to the source language, although it is very likely to be used to communicate what is essentially the same message. For as was pointed out by Ishikawa: "Language is not merely a simple tool that represents the world It controls the way people feel about the world." (Ishikawa 2013:222).

However, as is well known, in certain fixed expressions used on particular occasions, the lexical items of a language A as compared to language B may vary considerably as far as lexical meaning is concerned and still be used to communicate the same message. A couple of illustrations from concrete real-life situations: At the end of a restaurant meal the waiter may ask the client in French *Ca etait?*, that is *Was it?* or in Finnish *Maistuiko*? i. e. *Did it taste?* – and both expressions convey the question *Was it good?* Likewise, English-speaking children bored during a long car journey will usually ask *Are we there yet?* whereas their Polish counterparts would say *Daleko jeszcze*? which in turn means *Far away still?* Or as a final illustration: the sign warning people to keep away electric installations will in French bear the text of *Danger de mort* i.e. *Mortal danger* and in German *Lebensgefahr*, approximately *Danger to life* whereas an English warning sign will confine itself to simply informing about *Danger*, without classifying the exact danger level. – Such pragmatics related aspects lend themselves ideally to inform about in pragmatics information boxes of the kind proposed.

5. Warnings for double entendres

Users of bilingual dictionaries wishing to express themselves in the language that is not their own (whether source or target language) may need warnings enabling them to avoid potential embarrassing situations that may arise involuntarily owing to the existence of hidden undermeanings. – The present author recalls a situation where a teacher admonished a class of upper-secondary ("senior high school") students in a Swedish-medium school in Finland for using an excessively colloquial, if not slangy, language in their essays written in preparation for their national matriculation exam and told them to use a language more fit for grown-up people when dealing with serious themes. Then, by way of explanation to an American foreign student, she translated her statement into English as a request for the students to use more *adult language* in their texts. The result was unmitigated merriment, owing to the specific connotations this phrase has acquired in connections such as *this movie contains adult language*. – Somewhat similar was the case with a non-native speaker of English wishing to write a letter to congratulate a female colleague on account of her getting married
and, among many others things wished her a happy change of life, which is an all but literal translation of the French un changement heureux dans ta vie ... blithely unaware of the fact that the English phrase change of life, is usually perceived as synonymous with menopause. Appropriate fact boxes in connection with the key lexemes adult and change respectively could have prevented these rather awkward situations. In this same category also fall examples such as the English noun toilet, which, although a perfectly polite technical word is often avoided in social conversation where nouns such as restroom or bathroom are used instead, a fact the dictionary user might find helpful to be informed about. In all these cases, a pragmatics information box has its potential role to play.

Sometimes, incidentally, it may be warranted to include information to the effect that a given lexeme is free from under-meanings. For such information a pragmatics information box is also well suited. To give just one example: Swedish learners of English may find it helpful to be told in a dictionary involving the two languages concerned that the English noun party is perfectly neutral, as opposed to the Swedish loanword of party which very often comes with overtones of a posh lifestyle of "life in the fast lane cum liquor galore", and thus, it is perfectly possible in English to use party in references to staid and very formal social functions. – Strikingly enough, something similar is true of the noun fest: while in the English language its use is essentially reserved for ironic and pejorative occasions (cf. drinkfest, gossipfest, sobfest etc.), in the Scandinavian languages, Swedish included, it is the normal word to designate a party. – As a side remark, against this background it is easy enough to understand the somewhat confused reaction, in English-language readers when confronted with the fact that a vintage car speed gala, or racing gala, was referred to in Swedish newspaper articles as a fartfest (fart being Swedish for speed) ... Again, pragmatics information boxes no doubt have a role to fulfil.

6. Clues about cultural connotations

At this point we come to straddle the bordering line between a dictionary and an encyclopaedia quite markedly. Obviously, the possibilities of adding remarks about cultural specificities that go along with various lexemes are well-nigh infinite. Consider for instance the names of the colours. In the English language, for instance, blue comes with connotations of sadness and melancholy, as in feel blue or the baby blues, the latter a colloquial expression for post-partum depression, whereas yellow, unlike in many other languages, sometimes has decidedly unpleasant connotations, as in the yellow press or turn yellow, in the sense of 'lose courage'. Something similar could be said of nouns denoting various animals. In English a bear connotes with someone grumpy or ill-tempered, whereas in both Finnish and many Slavic language, the corresponding noun brings to mind associations of someone strong and rather sympathetic.

Sometimes the cultural aspects relating to lexemes are clearly extralinguistic. Initially, mention could be made of such a seemingly uncomplicated English word as chicken soup. In many parts of the English-language culture, this particular food is traditionally regarded as a home remedy for the common cold. Adding a comment about this might facilitate the non-native English language user's understanding of why, for instance, a news item warning for the spread of the 'flu may end on a familiar note to the effect that chicken soup will be in high demand in the upcoming season, and also to grasp the allusion effect when a publishing company selling inspirational "feel-good" books advertises under the words of "Chicken soup for the soul"

A case in point to illustrate this interface between lexicon and extralinguistic reality is the Finnish lähiö sometimes translated into English as suburb. The problem with this choice of equivalent is that a Finnish lähiö looks vastly different from notably an American suburb. Instead of being compose of largish, one-family houses usually well spaced-out from each
other, the Finnish lähiö will most likely consist of a few dozen multi-storey blocks of flats grouped around a shopping mall of modest size. Arguably, suburb may not even function as a very good equivalent of lähiö. – Another example is offered by the Finnish noun vappu and its English equivalent May Day. While the denotational equivalence is indisputable, the respective associations evoked by these two words are for the most part quite different. Whereas the English May Day is likely to primarily evoke associations of the international labour movement, left-wing politics and demonstrations (something which indeed used to be true of vappu in Finland until a generation or so ago), in modern Finland vappu is a rather carnival-like family event.

7. Keeping the distance between dictionary and encyclopaedia

As far as the last example in the previous paragraph is concerned, we are clearly approaching the demarcating line between editing a bilingual dictionary as opposed to editing an encyclopaedia. It may well be called into question whether not actually the provision of information, such as data concerning the nature of vappu in Finland as opposed to May Day is irrelevant for the lexicographic purposes. This issue is certainly not straightforward. On the one hand, there is no exact line to be drawn between a dictionary and an encyclopaedia. On the one hand, there are cases where the overstepping of this boundary line, no matter how vague and blurred, is rather manifest. Consider this example from an otherwise very meritorious Polish-Swedish dictionary:

The headword Wisła, the name of a Polish river usually referred to as the Vistula in English, supplied with the Swedish equivalent Weichsel i.e. the German name of the same river, traditionally used also in Swedish, and then, in boldfaced fine print: Polens längsta flod, 1047 km, in other words Poland's longest river, 1047 km [long]. – Although perhaps a comment about the river being the longest in Poland may be warranted, giving exact information as to its length hardly falls within the lexicographer's remit any longer. Then, again, in a context like this, exact numerical information may be called for wiorsta verst ryskt längdmått, 1066,8 m

or in other words, telling the dictionary user not only that wiorsta is a Russian length measurement unit, but also giving its exact dimensions; information that could be well needed in a translation situation if the translator wishes to express a distance in a manner more familiar to the readers/listeners.

8. Conclusion

Throughout this paper, quite a forceful case has been made, at least implicitly, for "pragmatics boxes", as they were called in the introduction. However, we must be aware of their limits. For one thing, they may make the procedure of consulting a bilingual dictionary unnecessarily complicated by burdening the dictionary user with a multitude of information perhaps difficult to handle. Secondly, if taken to extremes, they will slow down the dictionary editing process. After all, a dictionary is normally edited according to a set timetable, leaving little or no time for the editors to conduct overly comprehensive searches, even in these days of the Internet with all its accompanying search engines. Care must also be taken to make sure the attentive dictionary users realises that the explanations provided in such a box are indicative and not exhaustive. Finally, but certainly not least importantly, we must not lose sight of the obvious fact that not even the most ambitiously structured dictionary, whether monolingual or bilingual, whether in electronic or in paper format, can ever be a substitute for actual language learning and for practical experiences of the language other than the user's mother tongue. They are an aid on the road to a goal – but neither the road nor the goal itself. Even so, by judiciously exploiting the possibilities offered by modern information technology and pairing them with lexicographic skills, chances are that future dictionary users will
approach the goal of language mastery (yes, approach it, not attain it, given that its full attainment is impossible!) along a multilane expressway rather than a winding country byway and this in itself is certainly no mean accomplishment.

Acknowledgements
My sincerest thanks are due to all my colleagues at the Swedish translation unit of European Parliament, who despite an often staggering workload willingly shared language related experiences with me, thus facilitating the compilation of this paper.

References
Roles That Labels in Dictionaries Are Expected to Play

Masaaki Ogura
The University of Tokyo
rrtrnrtnrtn@ogdict.com

Abstract
A dictionary is a source of knowledge or information by virtue of being a collection of references. In fact, its attitude toward language has often been discussed. The issue of prescriptivism versus descriptivism is one of the main issues in this regard. One of the major cases of this controversy emerged after *Webster’s Third New International Dictionary* (W3) came out in 1961. This dictionary, being regarded as “too permissive” from a prescriptivist point of view, caused a debate from a variety of perspectives. As Gold (1985) points out, several points such as under-labelling, pronunciation, capitalization, and omissions were criticized. This paper attempts to consider the role that “labels,” in particular, play and are expected to play in relation to the notion of prescriptiveness and descriptiveness. First, some aspects of labels in W3 are discussed in comparison with other dictionaries of the same vintage. Then, a discussion of the process of prescriptivism being generated is undertaken by considering the transitory properties of labels. The author seeks to explore where descriptivism turns into or meets prescriptivism and vice versa. Further focus will be on the continuity if any, between these two concepts from the perspectives of the user and the author.

Keywords: label, prescriptivism, descriptivism, Webster’s Third

1. Introduction

Much has been debated over the roles that dictionaries are supposed to play including the issue of prescriptivism versus descriptivism. To some extent, they act as guides and serve various functions. Simply put, prescriptive dictionaries incorporate “good” words and rule out “bad” words, or show an attitude of disapproval to “bad” words. In contrast, descriptive dictionaries include every usage, independent of their “goodness” or “badness” (Béjoint 2010: 80). Of course, these notions are not dichotomous, and the continuum existing between them has already been discussed (Beal 2004). Some scholars have already put forward some concepts that lie behind the continuum (Bergenholtz 2003; Balteiro 2011; Gally 2011). However, the exact meeting point of prescriptivism and descriptivism has not been meticulously discussed. This paper attempts to discuss the roles that labels in dictionaries are expected to play and to locate “labels” on the continuum. Béjoint (2010) defines labels as the following:

A usage label is a noun or an adjective indicating the kind of context in which the word is normally used: slang, literary, American, Medicine, etc. Usage labels are usually abbreviated (sl, lit, US, med, etc.). They are of different kinds, social, geographical, stylistic, etc., corresponding to different varieties of the language. (Béjoint 2010: 34)

This suggests that labels have multiple roles, which indicates the characteristics of words in dictionaries. At the same time, however, the connotations that labels could have varied from one dictionary to another. Such connotations are more or less independent and contextual, and it can be said that labels do not mean much and the details are often obscure” (Landau 2001: 453). But as already pointed out, labels more or less assume prescriptive meanings, and this is closely reflected in the controversy over *Webster’s Third New International Dictionary* (W3), which was judged to be highly descriptive. The dictionary was criticized for some of its aspects and under-labelling was one of them (Gold 1985). The fact that under-labelling came under criticism means that the labels showed themselves to be an indispensable component of dictionaries, which people hoped would guide them to the “right” use of the language.

2. Webster’s Third New International Dictionary and its controversy

44
W3, after its publication, became the target of debate regarding its descriptive attitude. One of the major cases related to the descriptive attitude was the inclusion of the word *ain’t* which was very much criticized for being listed in the dictionary. Some people seemed to have thought that it was the first dictionary ever to record the word, which, as Skinner (2013) points out, was factually mistaken. Putting this point aside, the tumultuous response to the word suggests that part of journalism expects dictionaries to choose the “right” words for inclusion. The attitude of Philip Gove, the editor of the dictionary, is well described in the following:

Gove knew that even the simplest words defied the neatness of definitions and categories and usage labels, all desperately trying to pinpoint when, where, and under what circumstances chair or apple or girl may be used. (Skinner 2013: 12)

From the above statement, we know that Gove did not believe that labels do not work in the expected manner. However, this, in turn, makes the labels existing in W3 worth discussion in terms of the fact that they were not excluded and of the roles they could play. In what follows, we are going to see briefly what kinds of labels are seen in W2 and W3. But before that, it is to be noted that words labeled rare in W2 were eliminated from W3. It implies that *ain’t* was relatively in wide use and hence Gove thought it to be entitled to inclusion. Hayakawa (2015) observes that the usages followed by W3 were from the middle of the 20th century. It follows that even slangs were chosen to be listed in the dictionary if they were thought to be in wide use at the time when he was compiling the dictionary.

3. Prescriptivism, descriptivism, and their continuum

3.1 Prescriptivism and descriptivism

It has already been pointed out that the conceptual border between prescriptivism and descriptivism is not clearly defined. Beal (2004), for example, is skeptical about “the binary opposition of prescriptive/descriptive” (Beal 2004: 122–123) in terms of prescriptive grammar. However, it is not specifically stated what kinds of notions or concepts could lie between the two. As for this, some authors have proposed several notions. One of them is “proscriptive,” explained by Bergenholtz (2003), by which lexicographers “tell the user that variant he should choose if there is more than one possibility” regarding language variations. In a similar manner, Balteiro (2011) argues for “descriptive prescriptivism” which focuses on whether a particular usage would be accepted or refused by a particular group of people. These are based on prescriptivism on the one hand and descriptivism on the other hand. However, Vorlat (1998) has a slightly different view in which categories of grammar are divided into descriptive, normative, and prescriptive modes. This categorization has been mentioned because normative and prescriptive modes, which are sometimes put together, are treated separately here: normative refers to being based on language use, favoring the usages of some groups, and often having pedagogical purposes; while prescriptive refers to being based on usages but on a set of logical (or other) criteria (Vorlat 1998: 485–486). In this sense, Vorlat sees prescriptivism in a narrower sense than conventionally argued. Of course, we should note the concept beyond this prescriptive-descriptive issue, which Gally (2011: 128) calls “preemptivism.” However, this is a side issue.

3.2 Where do prescriptivism and descriptivism meet?

As stated above, details of the prescriptive-descriptive continuum are yet to be discussed intensively. It seems the issue is still dependent on authors—what is their basis for the terms “prescriptive” or “descriptive.” Here, the author proposes a provisional framework for the continuum in terms of the extent to which the author’s “subjective” judgments are involved. They are: (A)“ideal, logical prescriptivism,” (B)“non-variational prescriptivism,” (C)“variational prescriptivism,” (D)“descriptivism with socio-linguistic prescriptivism,” and (E)“non-judgmental descriptivism.” (A) refers to extreme prescriptivism where all lexicographical judgments are based on the author’s subjectivity or linguistic preferences.
Next comes (B), where the author resorts to actual linguistic phenomena that are preferred by the author or a particular group, but does not accept linguistic variations. In (C) linguistic variations are accepted but still recommendations are made as to which variations should be chosen. (D) relates to the description of actual linguistic variations and explains to readers whether a particular language use is accepted by a certain group or not, with the author’s subjective judgment least involved. (E) can be defined as extreme descriptivism where all the language variations become the target of inclusion. Practically speaking, (A) and (E) are almost impossible. Thus, (B), (C), (D) are what should be noted in discussing the properties of dictionaries. Please note that (C) has close relations with (B) and (D). In variational prescriptivism, the types of language to be chosen must be based on some criteria. When the criteria are subject to the author’s linguistic view, it comes close to (B) while more emphasize on actual language shows the characteristics of (D).

4. Prescription in description

In W3 controversy, the opponents seemed to have considered that “bad” words are not entitled for inclusion in dictionaries and that if they are to be described, they need to be labeled in a prescriptive manner. This is why underlabelling bore the brunt of criticism. The inclusion of a particular word means that the lexicographer thought that it merited inclusion. Hence, it matters how the word is being dealt with or described, and labels play a pivotal part in this regard. It should be noted here that there are some stages in the process of inclusion of a particular word. The first stage involves the choice of inclusion while the second phase is concerned with how the word is described. After considering this point, we can see that the word ain’t was criticized in both terms. The criticism was aimed at the fact that ain’t and other “bad” words were chosen to be included in the dictionary and at under-labelling. Let us compare the description of ain’t in Fowler’s Modern English Usage and in W3. W3’s definition of ain’t is as follows:

Though disapproved by many and more common in less educated speech, used orally in most parts of the United States by many cultivated speakers especially in the phrase ain’t I.

On the other hand, Fowler’s definition is written as follows:

(a) A(i)n’t is merely a colloquial, and as used for isn’t is an uneducated blunder and serves no useful purpose. But it is a pity that a(i)n’t for am not, being a natural contraction and supplying a real want, should shock us as though tarred with the same brush.

Here, we can see that the definition of the word in Fowler is quite emotional and shows despise to the use of ain’t. This might corresponds to (B) in the continuum described in the previous section, though it is possible that Fowler made this definition in response to Gove’s “generous” attitudes. Turning to Gove’s definition, we can see that it is hasty to conclude that Gove was not prescriptive. He did not only describe that the word was used but also that the use was disapproved. This kind of explanation might be termed as “social prescription.” Looking at some other word descriptions, we can see that Gove was not prescriptive merely in that he did not put inherent value judgment on words themselves. Rather, he showed a prescriptive attitude in terms of responses caused by a particular word among people in a social context. Considering this would introduce us to “social prescription” in discussing the prescriptive-descriptive issues. Labels that show social prescription appear, for example, on the definitions of chink, Cajun, bugre, brummagem, and Asiatic. In W2, these words were merely labeled as slangs or even not labeled that way. These changes of description could show that W3 worked on to describe how a particular word could cause certain responses in a social context.

5. Conclusion
So far, as shown by some examples mentioned in this paper, some definitions in \( W3 \) describe social prescription, which means how a particular use of language would be accepted or interpreted in a certain society or community. It is true that \( W3 \) was criticized for its underlabeling—expressions such as “taken to be offensive” or “disapproved” should be considered to be a label in a broader sense. In addition, this paper has shown that labels should be divided into linguistic labels, which concern the rule of usage or grammars, and sociolinguistic labels, which relate to how a language is interpreted in a social context. The words discussed here are often deemed improper from prescriptivism, but \( W3 \), which is a seemingly descriptive dictionary, also imposes some prescriptions on these words. This shows that labels in dictionaries have a transitory nature or have room for diverse interpretations, which accommodates both prescriptivism and descriptivism. It means that the place of this transition would matter. It could be in the middle part of the continuum, which is around (C) and highly concerns social prescription. The meaning of this prescription is not only based on the judgment the author makes but also on the reader’s interpretations. To put it another way, whether a particular entry or part of the entry is descriptive or prescriptive is highly contextual. Therefore, it is expected that a more contextualized framework of interpretation is needed to discuss the attitudes adopted by dictionaries.

References
Strict Corpus-based Lexicography and the Development of Dictionaries in the South African Indigenous Languages

V.M. Mojela
University of Limpopo, South Africa
victor.mojela@ul.ac.za

Abstract:
In this research, strict corpus-based lexicography refers to the use of the corpus as sole or major requirement for the compilation of dictionaries for the South African indigenous languages. South Africa has eleven official languages of which nine were previously marginalized. Before 1994, the country had only two official languages, i.e. English and Afrikaans. In order to have all languages developed to fully-fledged official status, the 1996 Constitution of the Republic of South Africa established eleven National Lexicography Units, or NLU’s, to compile dictionaries for all official languages. Most NLU’s depend greatly on the use of established corpora, especially the University of Pretoria corpora, as source for the compilation of dictionaries. Northern Sotho, like all the indigenous languages, is still very much underdeveloped and the majority of its dialects have never been converted to written form. As such, the vocabularies and structures of most of its dialects do not form part of the official written orthography of the language. Under the influence of factors which are mostly related to purism, the status quo is maintained and, ultimately, the lexicographers find it difficult to differentiate between the compilation of comprehensive dictionaries and the compilation of the standard dictionaries. The main objective of this research is to critically analyse the effects of depending solely on the use of the established corpora in the compilation of dictionaries for the South African indigenous languages.

Keywords: corpus-based, lexicography, standardization, dialect, orthography

Introduction
The first written documents in all the South African indigenous languages were done by the European missionaries. As a result, the areas which were fortunate to be the first to have missionary stations were the first to have their dialects documented. These written documents were the first literature to be used in the churches and schools, which were built in these missionary stations to provide education for the communities. Ultimately, the written documents which were based on the dialects spoken in the vicinity of the missionary stations were used as education materials, thereby elevating the ‘vicinity’ dialects above the majority of the dialects which were not documented due to their remoteness from the missionaries. Consequently, these ‘vicinity’ dialects became the ‘languages’ associated with education, development and prestige. By virtue of being documented, these ‘vicinity’ dialects gradually became elevated to be regarded as standard ‘languages’ at the expense of the overwhelming majority of the dialects which were never converted to written forms due to their remoteness from the missionary stations. This is exactly what happened to the Northern Sotho language which has 30 dialects spoken basically in the Limpopo Province of the Republic of South Africa. The Northern Sotho standard language is dominated by few dialects like Sepedi, Sekone and the dialects spoken in the Mphahlele, Mokopane, Moleëti, Matlala and Polokwane areas. These are the areas which were first to have missionary activities during the early 19th and 20th Centuries, especially the missionaries of the Berlin Evangelical Missionary Societies who were the first to document the Northern Sotho language. These ‘prestige’ dialects are just a fraction of the Northern Sotho dialects spoken in the Sekhukhune district, the Southern part of the Capricorn district and the Eastern part of the Waterberg district of Limpopo Province, which excluded the entire majority of the Northern Sotho dialects spoken in the North and the Western part of the Waterberg district, the Northern part of the Capricorn district, the whole of the Mopani district and the whole of the Mapulaneng district in the Mpumalanga Province.
Most of these marginalized dialects are totally different from the standard language, which can practically be regarded as a foreign language to these communities. This is the situation which the lexicographers find themselves when compiling dictionaries for the South African indigenous languages, i.e. of having and using one-sided and unbalanced electronic corpora, which are solely based on standard languages which are practically not reliable representative of the languages they are meant for.

1. What is Corpus-Based lexicography?

1.1 Defining Corpus

The term corpus is defined by various linguistic and lexicographic scholars such as, inter alia, Watson (1976), Kennedy (1998), James (1994), Gouws and Prinsloo (2005), as a body of written or spoken text. For instance, Watson (1976: 243) describes corpus as: ‘a body of writings of a particular kind, or on a particular subject’.

James et al (1994:4), give an intensive description of the term, corpus, as follow:

‘In recent usage (the term corpus) has tended to refer to a comprehensively documented and structured collection of complete texts, or extracts from larger texts, whose components are generally separately accessible’

According to Kennedy (1998:1):

‘A corpus is a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description.’

The most direct and unambiguous description of the term, corpus, is given by Gouws and Prinsloo (2005:21) in their latest publication, *Principles and Practice of South African Lexicography*:

‘The collection of written and spoken material from the sources earmarked for the dictionary basis. Data is compiled and stored as a lexicographic data basis which should preferably be an electronic corpus’.

In accordance with this definition, corpora are supposed to be compiled from both written and oral materials. But, on the contrary, it is not always easy to compile corpora for the languages or dialects which have no orthographies or written forms. Gouws and Prinsloo (2005:21-22) say the following in this regard:

‘Unfortunately most corpora around the world lack sufficient data from spoken sources. The reason for this is that there are many logistical problems and ethical factors involved in the collection of spoken data. It is also much more expensive and time consuming to enlarge the corpus with spoken data compared to data available in electronic, printed or even handwritten format. Extending the corpus with data already in electronic format such as texts downloaded from the internet or texts already available on computer disk is relatively easy. Printed matter which is not available in electronic form can be relatively and easily computerized by means of Optical Character Recognition (OCR), commonly referred to as 'scanning’.

In the context of the South African indigenous languages, corpus based lexicography can be described as the compilation of dictionaries on the basis of a well-established corpus, especially a written standard corpus. The established corpora in the South African indigenous languages are usually electronic corpora. In accordance with this system, the lexicographers compile dictionaries using only the already published corpora which, as a matter of fact, were compiled only from written documents. The Sesotho sa Leboa National Lexicography Unit and the other National Lexicography Units for the previously marginalized indigenous languages, use an electronic corpus which is based on the standard language when compiling the Northern Sotho dictionaries, especially the so-called University of Pretoria Sepedi Corpus,
(today called Pretoria Sepedi Corpus or PSC). This is an electronic corpus which was, mostly, compiled by the non-mother-tongue speakers of Northern Sotho who relied entirely on written documents and the Northern Sotho published literary works as the main sources of information. The disadvantage of relying entirely on the PSC is because the documented materials in Northern Sotho are based solely on the Standard Language which has excluded the majority of the Northern Sotho dialects. Therefore, the use of the PSC as the only source of information for dictionary writing in Northern Sotho results in the production of one-sided dictionaries which cannot be regarded as ‘balanced and representative’ of the entire vocabulary of the Northern Sotho language.

1.2 What is electronic corpus?

De Schryver and Prinsloo (2000:90) define electronic corpus as a documented electronic archive which is incorporated into a machine-readable form:

‘Archive materials are incorporated into an electronic text corpus which in turn is processed in a database. It is precisely as a result of the ‘machine-readable form’ that current corpora are referred to as electronic corpora’ (De Schryver & Prinsloo, 2000:90)


‘An electronic corpus can be defined in an oversimplified way as a computerized collection of texts, such a collection of texts can, for example, consist of tape recordings of conversations and written texts which have been typed into the computer’

The majority of the corpora for the nine South African indigenous languages are electronic corpora, and most of the corpora are electronically recorded in the University of Pretoria Corpora database. The importance of these corpora is emphasized by most scholars like Moon (1998), De Schryver & Prinsloo (2000), etc. Moon (1998:47) stresses the growing importance of corpora-based lexicography as follows:

‘The crucial point here is that corpora are used, and are now widely accepted as valuable, arguably essential, resources in serious linguistic description of any kind’

The importance of corpus-based lexicography in the compilation of dictionaries for the African Languages is further described by De Schryver and Prinsloo (2000:90) in the following citation:

‘If Scholars of African languages are to take their rightful place in the new millennium, it is plain that the active compilation, querying and application of electronic corpora should become an absolute priority’

1.3 Electronic Corpora and standard languages in the South African context

Language standardization can be viewed upon as a continuous process whereby conventional forms of a language are established and maintained. In accordance with this definition, language standardization, especially the codification as well as status and corpus planning, is supposed to be a continuous process in order to facilitate language development. With regard to this continuous development in the South African indigenous languages, standardization and corpus development continue to favour the ‘privileged’ dialects, i.e. the documented dialects. Corpus planning and standardization were supposed to mean incorporation of the various vocabularies of the dialectal varieties existing in the languages to facilitate growth and developments of the corpora, but the fact that the existing electronic corpora are only based on the existing standard languages which are not balanced and representative means that the marginalized dialects continuously remain side-lined. For instance, the following
examples, by Mojela (2008), are standard lexical items with their ‘undocumented’ dialectal variants in the Northern Sotho language, but due to the fact that these dialectal variants were never accepted as part of the standard vocabularies, apparently because their dialects are not documented, these terms continue to be outside the standard language, and therefore, continue to be side-lined by the Northern Sotho corpus:

<table>
<thead>
<tr>
<th>Dialectal (NS)</th>
<th>Standard (NS)</th>
<th>(English)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khepepe (Khelobedu)</td>
<td>Bene</td>
<td>(‘small truck’)</td>
</tr>
<tr>
<td>Malobane (Sepulana)</td>
<td>maabane</td>
<td>(‘yesterday’)</td>
</tr>
<tr>
<td>Melothi (Seroka)</td>
<td>Sejagobe</td>
<td>(‘greedy’)</td>
</tr>
<tr>
<td>Pikiri (Sepulana)</td>
<td>komiki</td>
<td>(‘cup’)</td>
</tr>
<tr>
<td>Mmankhoṭo (Seroka)</td>
<td>Leribiši</td>
<td>(‘an owl’)</td>
</tr>
<tr>
<td>Mokhope (Khelobedu)</td>
<td>Morula</td>
<td>(‘Marula beer’)</td>
</tr>
<tr>
<td>Kotapeni (Sepulana)</td>
<td>Afokhatha</td>
<td>(‘avocado’)</td>
</tr>
<tr>
<td>Mmidi (Sepulana)</td>
<td>Lefela</td>
<td>(‘mealie’)</td>
</tr>
<tr>
<td>Tholi (Sepulana)</td>
<td>Bopi</td>
<td>(maize meal)</td>
</tr>
</tbody>
</table>

2. **Comprehensive dictionaries versus standard dictionaries**

The term comprehensive is defined by various dictionaries to mean ‘to be as broad or as inclusive as possible’. A comprehensive dictionary is, therefore, a dictionary which is adequately inclusive to include the entire vocabulary of the language. This type of dictionary is expected to have almost all the vocabulary of the language as against specialized dictionaries which contain specialized vocabulary on specific subjects and the standard dictionaries. The standard dictionaries are primarily regarded as dictionaries which contain the vocabularies of the standard languages. This implies that most dictionaries which are compiled on the basis of the established corpora in the indigenous languages today may be regarded as standard dictionaries because most of the electronic corpora are compiled from documented standard materials, leaving out the majority of the dialects which are still not documented. Gouws & Prinsloo (2005:50) describe a standard dictionary as follow:

‘The macrostructure primarily represents the standard variety of a treated language although a number of high usage frequency items from non-standard varieties, e.g. slang or special fields, may also be included This simply meaning that, basically, the lemmatized vocabulary in the macrostructure of a standard dictionary should have gone through standardization via the accreditation authorities to qualify for inclusion in the Standard Dictionaries’

3. **Purism as marginalizing factor in the standardization of languages**

In Northern Sotho, the missionary works formed a base for a one-sided standardization by converting only few dialects into written forms and creating an elite group within the communities they operated, who later continued to participate actively in the development and the standardization of the language. Unfortunately, this led to the rise of purism because these elite groups did not have an intention to develop the language by including the Northern and the Eastern dialects in the standard Northern Sotho. With regards to the South African indigenous languages this type of purism, or dialectal exclusions, is manifested through
intensive marginalization of other dialects by most standardizing authorities. In most cases, the standardizing officials tend to standardize their own mother dialects at the expense of those dialects which are not adequately represented in the standardizing bodies. Dialectal exclusions in Northern Sotho, for instance, resulted in an unbalanced one-sided Standard Language which is presently being disowned by most of its own dialectal communities, such as the Balobedu, Mapulana, Batokwa, etc. A huge gap existing between the standard Northern Sotho and many of its own dialects has created a situation where a standard language looks like a complete foreign language to most of the marginalized Northern Sotho communities. The following are examples of the marginalized dialectal terms which are spoken by the majority of the Northern Sotho communities in the Northern part of the Capricorn and the Lowveld areas of Limpopo and Mpumalanga (Mojela, 2013:290):

<table>
<thead>
<tr>
<th>Standard NS</th>
<th>Khelobedu dialect</th>
<th>(English)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mopani</td>
<td>moṱhanare</td>
<td>(mopani tree)</td>
</tr>
<tr>
<td>leribisi</td>
<td>mmankhọtọ</td>
<td>(owl)</td>
</tr>
<tr>
<td>mmankgagane</td>
<td>molema</td>
<td>(bat)</td>
</tr>
<tr>
<td>hlapi</td>
<td>khobe</td>
<td>(fish)</td>
</tr>
<tr>
<td>betha/itiya</td>
<td>moṱa/tya</td>
<td>(hit)</td>
</tr>
<tr>
<td>legotlo</td>
<td>lehoṱo/peba/manṱoro</td>
<td>(mouse)</td>
</tr>
<tr>
<td>legapu</td>
<td>lesalabu</td>
<td>(watermelon)</td>
</tr>
<tr>
<td>nona</td>
<td>kholophana</td>
<td>(be fat/gain weight)</td>
</tr>
<tr>
<td>mokgaditswana</td>
<td>mphekwa</td>
<td>(lizard)</td>
</tr>
</tbody>
</table>

4. The consequences of corpus-based lexicography in the SA indigenous languages

Corpus-based lexicography in languages like Northern Sotho leads to the publication of dictionaries which are not only unbalanced, but are also not representative of the language because most of its dialectal vocabularies will not form part of the lemmatized vocabulary in these dictionaries. This is due to the fact that most compilers of electronic corpora in the South African indigenous languages depend solely on the standard languages which are also not balanced and not representative of the majority of their dialects.

4.1 The issue of balanced and representative corpora

With regard to the importance of a ‘balanced’ and ‘representative’ corpus, Summers (1993:190) says the following:

The idea of representativeness has been central to our thinking about the structure of the corpus. We believe that unless the corpus is representative, it is ipso facto unreliable as means of acquiring lexical knowledge. Our answer to the question: “Representative of what?” would be “Representative of the standard language” in a general sense, not restricted to a regional variety-, or a narrow range of text types’

Emphasizing Kennedy’s (1998:20) argument regarding the importance of balanced and representative corpora, Mojela (2013) says the following in this regard:

‘A normal and appropriate general corpus for a language needs to be balanced and representative. According to Kennedy (1998: 20), ”a general corpus is typically designed to be balanced, by containing texts from different genres—including spoken and written”. Kennedy (1998:52) further emphasizes that ”for a corpus to be 'representative' there must be a clearly analysed and defined population to take the
sample from.’

Gouws and Prinsloo (2005:25) confirm the importance of a balanced and representative corpus for the South African indigenous languages as follows:

‘Important for lexicographic work in South Africa is that corpus compilers should be sensitive to all of these aspects, i.e. to build as far as possible, corpora that are big enough, well balanced and representative so that valid conclusions for lexicographic purposes can be drawn.’

Unfortunately, this is not the case with most indigenous languages, like Northern Sotho, whose standard language is solely compiled from the published documents which are based on the documented dialects. Indeed, the majority of the dialects which were never converted to written form are still marginalized. Consequently, the dictionaries which are compiled solely on the basis of the established electronic corpora, like the PSA, also exclude these marginalized vocabularies and therefore, also becoming ‘unbalanced’ and ‘unrepresentative’ dictionaries for the Northern Sotho communities.

4.2 Rise of nationalism and rejection of the standard language

The issue of the ‘unbalanced’ and ‘unrepresentative’ corpora, which resulted from a one-sided standardization, has led to the emergence of a nationalistic spirit among the marginalized communities. The emerging elite groups and the magoši (traditional leaders) from areas such as Bolobedu, Makhutšwe, Botlokwa, Senwabarwana, Mapulaneng, etc. have already started questioning the validity of incorporating their dialects in the Northern Sotho standard language, which is not only lexically and morphophonologically foreign to them, but in all practical respects, too different from their dialects.

5. Conclusion

This research demonstrates that corpus lexicography in the South African indigenous languages, like Northern Sotho, is not always possible, because there are still many dialects within these languages which are not included in both the official orthographies and the standard languages. Since the corpora are compiled mostly from the written languages and from published materials, the dialects which did not have written forms or published materials will not always be included in the corpora. Even though sometimes oral materials were collected and included into the corpora, very few of these published materials were recorded in languages like Northern Sotho. This is because, up to this moment, almost all the available Northern Sotho corpora do not have anything related to the marginalized dialects like Khelobedu, Sepulana, Seṱokwa, Sehananwa or Sephalaborwa. As a result, the available Northern Sotho corpora do not conform to the lexicographic principles of ‘balance’ and ‘representativeness’. The standard language is neither balanced nor representative because it reflects less than half of its dialects.

In conclusion, this research recommends further research and a thorough revision of the official orthography and the standard language to incorporate all the omitted dialects into the Northern Sotho language, before prescribing strict corpus-based lexicography. This will not only silence the emerging nationalistic spirit which threatens to divide the language, but will also double the size of the Northern Sotho corpus.
References


Comparing Japanese-English Dictionary Entries of English Loanwords with Corpus-Derived Lexical Profiles

Keith Barrs
Hiroshima Shudo University, Japan
barrs@shudo-u.ac.jp

Abstract
When loanword descriptions in Japanese-English dictionaries are compared with corpus-derived lexical profiles of their collocational and grammatical behaviour, important differences become apparent. In the face of the wealth of corpus data about the loanwords that can be generated rapidly and presented systematically by modern corpus software tools, lexicographic descriptions in dictionaries seem minimal and unstructured. In particular, corpus data on individual loanwords can be combined into a systematic analysis of groups of loanwords, in order to see if any patterns emerge in relation to their behaviour in the language. This can help to draw out information about the semantic and pragmatic properties of the loanwords, something that is largely absent from Japanese-English dictionaries. This study looks at the kind of analyses that can be made of English loanwords in Japanese using corpus data, compared with the available descriptions in popular Japanese-English dictionaries. It raises the questions of whether or not these dictionaries are fit-for-purpose in describing the loanwords, and discusses the implications for English-speaking learners of Japanese.

Keywords: English loanwords, Japanese language, dictionaries, corpus, word sketches

Introduction
The Japanese language is home to tens of thousands of English loanwords, many having become highly frequent in the lexicon (Stanlaw, 2004). Whilst some function primarily to fill lexical gaps created with the importation of new products and ideas, others are used for more stylistic purposes, such as providing ambiguity of expression and creating images of modernity. Because of the adaptations made to the words as they are borrowed from English and integrated into Japanese, it is widely acknowledged that a one-to-one meaning association between the borrowed word in Japanese and the source word in English cannot be automatically assumed (Irwin, 2011; Kay, 1995; Ringbom, 2007a). For whilst there are many examples of word pairs which are indeed highly similar in form and meaning, such as バナナ (banana) in Japanese borrowed from ‘banana’ in English, there are many others which have been adapted phonologically and semantically to an extent that obscures their English origins. Such an example is リストラ (risutora), shortened from English ‘restructuring’, which can be used euphemistically to describe a loss of jobs at a company due to its internal restructuring processes.

A great number of these loanwords have made their way into the core of the Japanese language. Daulton, for example, found that 1356 of the most frequent 3000 word families (45.2%) extracted from the British National Corpus had loanword forms in Japanese (2008, p. 83). As a result of their high frequency, many English loanwords are regularly encountered throughout Japanese society; in newspapers, in song lyrics, in government reports, on product packaging, on shop signs, and on clothing (Backhaus, 2007; Barrs, 2011). This then predictably impacts on language education, when Japanese-speaking learners of English (JLOE) and English-speaking learners of Japanese (ELOJ) may identify cross-linguistic lexical similarities when they bring their existing L1 knowledge to bear on the learning of the L2 versions of the words (Nation, 2008; Ringbom, 2007a, 2007b). In such learning situations, the fact that one-to-one meaning associations do not always occur between the loanwords and the original English words (Irwin, 2011; Kay, 1995) makes dictionaries an important reference tool. This article focuses specifically on ELOJ, looking at whether or not Japanese-English dictionaries are fit for purpose in assisting learners to identify similarities and differences between the English and Japanese versions of the words.
Coverage of English loanwords in Japanese-English dictionaries

Before investigating the adequacy of lexicographic descriptions of English loanwords in Japanese-English dictionaries, it is necessary to check how well these dictionaries account for the loanwords found in Japanese society. The web corpus, jpTenTen11, was chosen as a source of attested loanwords in Japanese society for the following three reasons. First, at over 10 million tokens, it is an ultra-large scale corpus covering an extensive range of genres found on the web, such as adverts, blogs, news stories, discussion forums, and comment threads. It is currently the largest available corpus of Japanese language data used in Japanese corpus linguistics research. Second, it is part of the TenTen family of 10-billion token web corpora, housed in the Sketch Engine corpus query system, which have all been designed, compiled, processed, and re-analysed by experts using the available software tools for each language. Thirdly, because the corpus is housed within a state-of-the-art corpus query system, it is possible to query it using a range of powerful techniques such as generating ‘sketches’ of the words’ collocational and grammatical behaviour in the corpus.

To extract frequent English loanwords from this corpus, the following process was followed. Loanwords in Japanese are predominantly written in katakana script, greatly assisting a computer’s ability to identify and retrieve them. However, there are cases where loanwords are more commonly found in other scripts (especially in the case of loan translations such as 空港 (kuukou) composed of the kanji characters representing ‘air’ and ‘port’ from English), and also where katakana is frequently used to write words of Japanese/Chinese origin, such as ネコ (neko) meaning ‘cat’. Accordingly, for the purposes of this research an ‘English loanword’ was defined in a three-step process. First a list of katakana strings was extracted from the most frequent 10,000 lemmas in the jpTenTen11 using a regular expression (a text string for describing a search pattern). The resulting list of katakana lemmas was then tagged for language origin using the unidic Japanese dictionary. This dictionary has been used extensively in Japanese corpus linguistic research, most notably in the compilation of the Balanced Corpus of Contemporary Written Japanese (BCCWJ). This filtered the list into katakana lemmas which could be classified as part of the ‘foreign loanwords’ lexical stratum (gairaigo in Japanese). The overwhelming majority of loanwords in Japanese are from English, and the third step involved manually discounting non-English lemmas from the list as and when they were identified, by cross-referencing them with a standard Japanese dictionary which tags loanwords with their language of origin. For example, the loanword テーマ (teema) is tagged as being derived from German ‘Theme’ rather than English ‘theme’, and was therefore discounted from the list.

To investigate the coverage of English loanwords by Japanese dictionaries, the top 100 and bottom 100 loanwords on the list created in the previous step were analysed. Each loanword was searched across a collection of six dictionaries: (1) 大辞林第3版 (daijirindaisanban) ‘Great Forest of Words, 3rd Edition’ (Japanese monolingual dictionary), (2) コンサイスカタカナ語辞典第4版 (konsaisukatakanaogodaiyonban) ‘Concise Katakana Words Dictionary, 4th edition’ (loanwords dictionary), (3) グランドコンサイス和英辞典 (gurandokonsaisuwaeijiten) ‘Grand Concise Japanese-English Dictionary’ (Japanese-English dictionary), (4) ワイズダム和英辞典第2版 (wizudamuwaeijitendainiban) ‘Wisdom Japanese-English Dictionary, 2nd Edition’ (Japanese-English dictionary), (5) コンサイスEX和英辞典 (konsaisuEXwaeijiten) ‘Concise Japanese-English Dictionary’ (Japanese-English dictionary), and (6) デイリーコンサイス和英辞典第5版 (deirikonsaisuwaeijitendaigoban) ‘Daily Concise Japanese-English Dictionary, 5th Edition’ (Japanese-English dictionary). It was possible to search across all these dictionaries at once, for each word, through the 三省堂 (Sanseido) Web Dictionary online interface. The monolingual ‘Great Forest of Words’ dictionary and specialised loanword ‘Katakana Words’ dictionaries were used as a base with which to compare the coverage in the four Japanese-English dictionaries. These two dictionaries are much larger than the Japanese-English dictionaries and were therefore expected to have a better coverage of the loanwords. However, they are not likely to be the first choice for ELOJ, especially beginner learners, because they are Japanese monolingual resources.

The dictionaries were checked to see if they had listings for each of the 100 loanwords at the top and bottom of the list. Many loanwords have several meanings which may be listed separately, depending on the methodology followed by each dictionary, and in some cases the same katakana orthographic form is used to represent different source words (such as バス (basu) used for English ‘bath’, ‘bus’ and ‘bass’). However for this stage of the analysis these differences were not considered important. Rather,
the dictionaries were checked only for an exact match with the katakana script representation of the loanword in the corpus. The results of the coverage analysis are shown in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>TOP 100</th>
<th>BOTTOM 100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Daijirin</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Katakana</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Grand Concise</td>
<td>98</td>
<td>2</td>
</tr>
<tr>
<td>Wisdom</td>
<td>97</td>
<td>3</td>
</tr>
<tr>
<td>EX</td>
<td>93</td>
<td>7</td>
</tr>
<tr>
<td>Daily Concise</td>
<td>86</td>
<td>14</td>
</tr>
<tr>
<td>TOTAL (%)</td>
<td>574 (95.7%)</td>
<td>26 (4.3%)</td>
</tr>
</tbody>
</table>

Table 1 The number of hits for the 200 loanword sample in six Japanese dictionaries.

The results show that, as expected, the monolingual dictionary and specialised loanword dictionary have the best coverage of the loanwords that appear in the jpTenTen11 corpus. The single loanword with no exact match in the Daijirin dictionary, ワオーター (woota), from English ‘water’, is because of the fact that this word has several possible orthographic representations. The word is indeed recorded in ‘Daijirin’ with the form, ウオーたー (woota) (the difference is with the size of the second character, オ or オ, which would alter the pronunciation from wo to wo). This in itself is interesting and may indicate a possible mistake with the orthographic listing of this loanword in the dictionary, revealed through reference to a corpus, because there are 128,436 hits (12.4 p/million) for ウオーター (small オ) in the jpTenTen11 compared to only 5,853 (0.6 p/million) for オオーター (big オ). Taking this analysis further, the corpus includes 2,758 hits (0.3 p/m) for オオータ (a short タ (ta) rather than long ター (taa)), but this orthographic form is not listed in any of the six dictionaries. The number of hits in the corpus may be just about large enough to suggest that this form should be listed as an alternative representation in the dictionary entry. The form オオータ (with big オ (o) and short タ (ta)) only has 12 hits (0.0 p/m) which could be fairly safely categorised as orthographic (i.e. spelling) errors. Accepting that ‘Daijirin’ records one variant orthographic loanword form of the English source word ‘water’, the combined coverage in the monolingual and specialised katakana dictionaries of the 200 loanword sample is 100%.

The Japanese-English dictionary with the most coverage of English loanwords is the ‘Grand Concise’ with 98%, and the one with the least is the ‘Daily Concise’ with 77.5%. This is a direct result of the size of the dictionaries, with the ‘Grand Concise’ having 210,000 entries compared to 70,000 in the ‘Daily Concise’. Taken together, the four Japanese-English dictionaries have 90.1% coverage of the 200 corpus-extracted loanwords, and the six dictionaries combined have coverage of 93.3%. Overall it is clear that the Japanese-English dictionaries have very good coverage of frequent English loanwords, with fluctuations in the extent of coverage by each dictionary being a result of their overall size. The next question to investigate is about the quality of the dictionary entries of the loanwords, when compared with corpus-derived lexical profiles.

Comparing loanword dictionary entries with corpus-derived lexical profiles

Advances in the size and quality of corpora, along with the software tools used to access them, have allowed significant advancements in the investigation of vocabulary. This is particularly evident in the development of ‘word sketches’ produced by the Sketch Engine corpus query software, which give an overview of how words behave in a corpus. Word sketches are defined as “one-page automatic, corpus-based summaries of a word’s grammatical and collocational behaviour” (Kilgarriff, Rychlý, Smrž, & Tugwell, 2004, p. 1). These word sketches can be produced for any lemma in any corpus uploaded or already housed within the Sketch Engine, and can therefore be produced for English loanwords in Japanese in the jpTenTen11 web corpus.

Word sketches were generated for the top 30 and bottom 30 loanwords on the list used in the preceding part of this article. The purpose was to use corpus methods to derive lexical profiles of the loanwords, and then to conduct an exploratory analysis into what these profiles can reveal about the behaviour of the words. The corpus-derived profiles would then be compared with the loanwords’ entries in
Japanese-English dictionaries. This was in order to assess the quality of the lexicographic descriptions in the dictionaries. The following discussion focuses on the English loanwords that exhibited a particularly interesting grammatical and collocational patterning in their word sketch, when compared to the other loanwords on the list.

The word sketch function of the Sketch Engine presents a one-page summary of the most salient grammatical relationships within which the loanwords participate in the corpus. The grammatical relationships (called ‘gramrels’) are presented in order of frequency, with the list of collocates in each gramrel ranked by a salience score (using a form of the LogDice coefficient). Each word sketch was then manually analysed, and the top ten most frequent gramrels were recorded for each loanword. These gramrel rankings were then analysed to investigate whether or not any patterns could be seen in how the loanwords behave in the corpus. Even in this small-scale analysis of 60 frequent English loanwords in Japanese, a large number of discoveries were made when examining their grammatical and collocational behaviour. The following discussion focuses only on the most frequent gramrel of each loanword, with further studies planned for investigating the gramrels in more depth.

Of the 60 loanwords investigated, 47 (78.3%) had ‘particle’ as their most frequent gramrel, making this the unmarked grammatical relationship of the loanwords. The 13 loanwords without ‘particle’ as the most common gramrel can be considered the marked examples, indicating that these loanwords are behaving somewhat differently from the others. 12 (20%) of them have ‘noun/noun’ (indicating a compound noun relationship) and one (1.7%) has ‘N_Adj’ (indicating the loanword is modified by another noun acting as an adjective). Therefore, two dominant patterns of loanword behaviour based on the most frequent gramrel of each loanword emerge from the word sketches.

For the 47 loanwords with ‘particle’ as the most frequent gramrel, this means that the loanword was immediately followed by any one of a large number of grammatical particles that show the relationship between the loanword and the rest of the sentence. Examples include case markers such as に (ni) indicating direction, を (wo) indicating the loanword is the subject of a following verb; and sentence ending particles such as か (ka) indicating a question. The fact that the loanwords are most frequently associated with a large range of grammatical particles shows that they have a wide range of functions in the language. This indicates that the loanwords have been fully integrated into the syntax of the Japanese language.

It seems that when ‘particle’ is the most frequent gramrel, the loanword is being used primarily in relation to its semantic properties; in other words it is being denotatively used to refer to an object, concept or idea expressed through the loanword. Examples of such loanwords are include ブログ (burogu, ‘blog’), which is ‘read’, ‘uploaded’ and has ‘articles posted to it’; デザイン (dezain, ‘design’) which is ‘created’, and used to ‘make decisions’ and ターミナル (taaminaru, ‘terminal’) which is ‘visited by travellers’, and ‘networked with computers’. Looking at one example in more detail, the three most common particles of the loanword サービス (saabisu, ‘service’) are を (wo), など (nado), and や (ya). Going into some of the concordance lines for each of the particles (of which there are often tens of thousands) shows that  wo indicates the loanword is playing the role of subject of a verb (e.g. サービスを利用する, saabisu wo riyou suru, ‘to use a service’),  nado indicates that the loanword is being put forward as a prototypical example, and ya indicates that it is part of a list (e.g. サービスや商品, saabisu ya shouhin, ‘service and products’).

The other dominant pattern, that of ‘noun/noun’ as the most frequent gramrel, shows a different grammatical behaviour of some of the loanwords. Here, the loanword is joining with another noun to form a new compound noun. As such, the loanword is acting as an adjectival noun. This means that rather than acting as a standalone thing in its own right, such as a blog, design, or terminal, the loanword is used more for its ability to modify a different noun. Examples of this group are ソフト (sototo, ‘soft’), ベイ (bei, ‘bay’), ミステリー (misuterii, ‘mystery’) and インターナショナル (intaanashonaru, ‘international’) which collocate with other words to make compounds such as ソフトドリンク (sototo dorinku, ‘soft drink’) and インターナショナルスクール (intaanashonaru sakuuru, ‘international school’).

The behaviour of this group of loanwords is particularly interesting when looking at the types of words which appear as the frequent collocates in the ‘noun/noun’ gramrel. For some of the loanwords, they
seem to only collocate with other loanwords. When this happens, it tends to be the case that many of the loanword collocates have viable alternative expressions in the native language; often termed semantic near-equivalents (Onysko & Winter-Froemel, 2011). In these cases it therefore appears that the loanword is fulfilling a more pragmatic than semantic function; in other words, the loanword is being selected instead over a viable native term in order to imbue some stylistic effect in the particular instance of language use. For example, in the word sketch for the loanword べイ (bei, ‘bay’), it is shown to collocate with 22 other loanwords. Many of these, such as シティー, (shitei, ‘city’) and サイド (saido, ‘side’), have regular alternative expressions in the native Japanese vocabulary. The remaining collocates are one English initialism (FM) and two kanji character compounds (proper noun place names). When it collocates with the other loanwords it seems to be adding a modern/international flavour. For example, the compound べイタウン (bei taun, ‘bay town’) is used in the corpus to describe a living area in Chiba, Japan designed on a foreign city planning model and which includes many foreign food stores, coffee shops and international restaurants. In this example, bei could feasibly be replaced with the native word 湾 (wan, ‘bay’) and taun could be replaced with 町 (machi, ‘town’), but this compound is not attested in the corpus.

It is clear that a word sketch of a loanword provides an abundance of interesting data concerning how the word behaves in the corpus. Firstly, we can get a clear idea of the frequency of the loanword in the corpus. Then we are able to see which grammrels the loanword participates in, and these are ranked in a frequency order. Then within each grammrel we can see a list of collocates, this time ranked by a salience score, which allows the disambiguation of different senses of the word and an overall idea of how the word is being used. Focusing in on particular features of the word sketch, such as the type of collocates in each grammrel, allows a deeper analysis of issues such as the semantic and pragmatic properties of the loanwords.

When these lexical profiles are compared with dictionary entries for the loanwords, it is clear that language learners are limited in the quality of the dictionaries’ lexicographic descriptions. To begin with, none of the Japanese-English dictionaries used for this research record any frequency data for the loanwords. This means that the learner has little idea whether a loanword they have come across is a frequent part of Japanese and worth acquiring. Further, the dictionaries record very little systematic grammatical data about the loanwords. Whilst basic details are given of the part of speech of the loanwords (most commonly, noun), there is no indication of any frequent grammatical patterns. In terms of collocates, the dictionaries list a limited number of the more common collocational patterns, but by alphabetical order rather than frequency or salience. This means that learners are unable to deduce any significant patterns in how the loanwords collocate with other words in the language. Indeed, it was difficult to see any particular systematic handling of the descriptions of the loanwords in the Japanese-English dictionaries. Not only did they differ between themselves in the extent of the lexicographic descriptions, but they also differed within themselves with it appearing as if there was no particular rule on what to include in the description or how to present it.

**Conclusion**

The main result of this research, that the lexicographic descriptions of loanwords in Japanese-English dictionaries are extremely minimal when compared with corpus-derived lexical profiles, is for the most part an expected and obvious finding. Corpus software allows the rapid and reliable summarisation of a massive number of concordance lines, something that is far beyond the manual labour that continues to go into Japanese dictionaries (Kilgarriff et al., 2014). Rather, the value of this research should be seen in the investigation of what corpus-derived data can reveal about how the loanwords are behaving in the Japanese language. After generating individual word sketches, and then comparing them, patterns begin to emerge about the grammatical and collocational behaviour of the loanwords, which can be investigated further by delving into the concordance lines. This kind of intricate picture of the loanwords in a language can be of great value to learners who may already be familiar with the words as they exist in their L1, and who need to know how the L2 loanword forms may differ in semantic and/or pragmatic functions.
References


The possibility of building a dependency-based Japanese-English construction dictionary

Masanori Oya
Mejiro University, Japan
m.oya@mejiro.ac.jp

Abstract
This study explores the possibility of building a dependency-based Japanese-English construction dictionary. Dependency Grammar (Tesnière 1959) has recently been attracting attention in the field of theoretical linguistics and natural language processing. It assumes that the dependency relationship between words in a sentence is the essential feature of syntactic analysis, and the syntactic structure in the Dependency-Grammar framework is simpler than phrase-structure trees in that it does not contain syntactic categories such as NP or VP. This paper is a preliminary proposal to build a dependency-based Japanese-English construction dictionary, in which Japanese constructions (Goldberg 1995) are represented as dependency trees, and possible English translations are provided, which are also represented as dependency trees. The paper is structured as follows. First, we examine the advantage of Dependency Grammar, mainly in terms of its familiarity with computational processing of linguistic data. Second, we discuss the significance of dependency-based construction dictionaries, comparing them with traditional word-based dictionaries. Third, we examine the results of a parallel-corpus-based study that was conducted for this paper. In this research, the sentences that share a Japanese construction noun-to nanteiru were selected from a Japanese-English parallel corpus, and the constructions that their English translations can take are provided as evidence that a Japanese construction corresponds to more than one English construction. The results remind us of the simple, yet often ignored, fact that we need to consider something other than word-to-word correspondence when learning foreign languages. Finally, directions for future research are discussed.

Keywords: Dependency Grammar, Construction, Japanese-English Parallel Corpus, One-to-Many Correspondence

1. Introduction

This study explores the possibility of building a dependency-based Japanese-English construction dictionary. Dependency Grammar (Tesnière 1959) has recently been attracting attention in the field of theoretical linguistics and natural language processing. Because it assumes syntactic trees for sentences that are simpler than phrase-structure trees, it has been applied to a number of research fields, such as information retrieval or machine translation. This research is an attempt to extend the research field of Dependency Grammar into lexicography, with respect to constructions (in Goldberg’s (1995) sense) where there is no one-to-one correspondence between the words in a source sentence and the words in a target sentence.

The paper is structured as follows. First, we examine the advantages of Dependency Grammar, mainly in terms of its familiarity with computational processing of linguistic data. Second, we discuss the significance of dependency-based construction dictionaries, comparing them with traditional word-based dictionaries, with respect to the syntactic mismatches in Mel’čuk and Wanner (2006). Third, we examine the results of a parallel-corpus-based study that was conducted for this paper. In this study, sentences that share a Japanese construction noun-to nanteiru are selected from a Japanese-English parallel corpus, and the constructions that their English translations can take are provided as evidence that a Japanese construction corresponds to more than one English construction. The results remind us of the simple, yet often ignored, fact that we need to consider factors other than word-to-word correspondence when learning or using foreign languages. Finally, directions for future research are discussed.

2. Dependency Grammar
Dependency grammar focuses on the dependency relationships among words in a sentence. Since it was first proposed by Tesnière (1959), a number of frameworks of dependency grammar have been developed; for example, extensible dependency grammar (Debusmann 2003), word grammar (Hudson 2010), and Stanford Dependency (de Marneffe and Manning 2015). What is common among these different dependency representations of sentences is that they can represent the syntactic structures of a sentence in a simpler manner than phrase-structure representation does, because typed-dependency representation does not contain syntactic categories such as NP (noun phrase) or VP (verbal phrase). In this research, Stanford Dependency is chosen as the framework to represent the dependency relationships among the words in a sentence, because it has been implemented into the Stanford Parser allowing us to obtain typed-dependency representations for a fairly large number of sentences. In addition, it has been applied to a variety of fields (mentioned later).

Stanford Dependency posits a typed-dependency representation for a sentence, in which the words in a sentence of a language (or language in general) constitute a network of dependency in which a given word depends on another word. The dependency relations are categorized into different types. Figure 1 is a typed-dependency tree for a sentence.

![Figure 1 The typed-dependency tree for “I am studying construction grammar.”](image)

Stanford Dependency has been applied to a variety of fields, such as text mining, sentiment analysis, information retrieval, and textual entailment extraction. Detailed references of studies that use Stanford Dependency are found in de Marneffe and Manning (2015).

2.1 **Syntactic mismatches between languages in the Dependency-Grammar framework**

Dependency grammar has also been applied to the field of machine translation, and its simplicity leads to the idea that the syntactic discrepancy between a source sentence and its target counterpart can be represented in a simple, principled, and computable manner. Mel’čuk and Wanner (2006) was one such attempt. This idea can also be applied for the purpose of creating a construction dictionary, for reasons that will be discussed later. In this section, let us summarize Mel’čuk and Wanner’s idea through a number of example Japanese-English translation pairs.

Mel’čuk and Wanner (2006) defined the types of difference between a sentence in a source language and its translation in a target language as follows:

The major types of syntactic mismatches (Mel’čuk and Wanner 2006: 82-83)
1. **Conversion**
2. **Head switching**
3. **Lexical fission /fusion**
4. **Part-of-speech change**
5. **Function-word introduction**
6. **Branch transposition**

2.1.1 **Conversion**

In the syntactic mismatch conversion, the grammatical function of the source language is converted into another function in the target language. For example, in the sentence below, the Japanese noun *hon* with the nominative particle *-ga* functions as the subject, and corresponds to the English ‘book’ as the object. In addition, the Japanese noun *watashi* with the topic marker *-wa* functions as the topic of the sentence, and corresponds to the English ‘I’ as the subject.

(1)

```
watashi-wa kono hon-ga kiniitteiru
```

*I-top this book-nom like

‘I like this book.’

The typed-dependency trees for the source and target sentences are shown below: the converted grammatical functions are emphasized in bold italics. In this study, the typed-dependency representation of English is based on Stanford Dependency (de Marneffe and Manning 2009), and that of Japanese is based on Oya (2014b).

![Figure 2](image1)

**Figure 2** The typed-dependency trees for (1).

2.1.2 **Head switching**

In the syntactic mismatch head switching, the direction of a dependency in the source language is reversed in the target language. For example, in the sentence below, the Japanese word *ryori* ‘cooking’ depends on *jouzu* ‘good,’ and this dependency relationship is reversed in English in which the word ‘good’ depends on ‘cook.’

(2)

```
watashi-no ani-wa ryori-ga jouzu-da
```

*I-gen brother-topic cooking-nom good-present

‘My brother is a good cook.’

![Figure 3](image2)

**Figure 3** The typed-dependency trees for (2).
2.1.3 Lexical fission
In the syntactic mismatch *lexical fission*, one word in the source language corresponds to more than one word in the target language. For example, in the sentence below, the Japanese verb *wakatta* corresponds to the English phrasal verb ‘turned out.’

(3)
Taro-no iu toori-datta-to wakatta.
Taro-part say right-was-part turned.out
‘It turned out that Taro was right.’

![Figure 4. The typed-dependency trees for (3).](image)

2.1.4 Lexical fusion
In *lexical fusion*, more than one word in the source language corresponds to one word in the target language. In the example below, the Japanese expression *nakereba naranai* corresponds to the English auxiliary ‘must.’

(4)
Taro-wa motto benkyo-shi-nakereba naranai.”
Taro-topic more study-do-if.not neg
‘Taro must study harder.’

![Figure 5. The typed-dependency trees for (4).](image)

2.1.5 Part-of-speech change
In the syntactic mismatch *part-of-speech change*, a source word of a certain part of speech corresponds to a target word of a different part of speech. In the example below, the Japanese noun *ame* corresponds to the English verb ‘rain.’

(5)
ame-ga futeiru.
2.1.6 Function-word introduction
In the syntactic mismatch *function-word introduction*, a function word is added in the target language. For example, the preposition ‘at’ and the article ‘the’ are added in the target language.

(6)
```
kokuban-wo mite kudasai
```
blackboard-acc look please
‘Please look at the blackboard.’

2.1.7 Branch transposition
In the syntactic mismatch *branch transposition*, a certain word in the source language changes its dependency head in the target language. For example, the Japanese word *san-dai* depends on the verb *motteiru*, while the English word corresponding to *san-dai* does not depend on the verb ‘has,’ which corresponds to the source verb *motteiru*, but on the noun ‘cars.’

(7)
```
watashi-no chichi-wa kuruma-wo san-dai motteiru.
```
I-part father-topic car-acc three-class have
‘My father has three cars.’
2.2 Syntactic mismatches between the Japanese construction -to natteiru and its English translations

What these syntactic mismatches tell us can be summarized as follows: one-to-one correspondence between a source word and a target word cannot capture the discrepancy between a source sentence and a target sentence. If we intend to compile a dictionary that informs us of the correspondence between each construction of a source language and its possible translations in a target language, just listing them is not helpful, because it does not take the discrepancy into consideration. For example, the Japanese construction X-to natteiru (literally, ‘has become X’) has a variety of possibilities to be translated into English sentences, and the source verb natteiru does not necessarily corresponds to ‘has become.’ In the example Japanese-English translation pairs (which are based on the sentences taken from the corpus data mentioned below), the Japanese verb natteiru corresponds not only to ‘become,’ but also to other verbs.

(8) genzai-wa kono iken-ga teisetsu-to natteiru.
     today-topic this opinion-part accepted.opinion-part become-present
     ‘This opinion is accepted today.’

(9) kugatsu-ni okonaw-areru-no-ga tsuurei-to natteiru
     September-part hold-passive-part-part usual-part become-present
     ‘It is usually held in September.’

(10) shiryo-toshite kichou-na mono-to natteiru
     documents-part valuable-part thing-part become-present
     ‘They serve as valuable documents.’

(11) Sono zou-wa monocoque-ni chikai kouzou-butsu-to natteiru.
     the statue-topic monocoque-part close structure-thing-part become-present
     ‘The structure of the statue is close to monocoque.’

(12) juyo-bunka-zai-to natteiru.
     important-culture-property-part become-present
     ‘It has been designated as an important cultural property.’

It is not helpful to list the corresponding verbs as entries for the Japanese verb natteiru. Hence, if we are to compile a construction dictionary that is really helpful for users, it is desirable to
include the information about these possible constructions, along with the syntactic mismatches in each of them. The next sections focus on the construction -to natteiru in a corpus, and discuss the implications of the findings.

3. Corpus Analysis

3.1 Corpus Data
The corpus used in this study was the Japanese–English Bilingual Corpus of Wikipedia’s Kyoto Articles, v. 2.01 (National Institute of Information and Communications Technology 2011). Approximately 500,000 manually translated Japanese–English sentence pairs were taken from Japanese Wikipedia articles on things or people that are related to Kyoto. These articles were translated manually into English text by several translators, and native speakers of English checked their quality (English articles were not used). This corpus contains 16 subcorpora according to topic (such as religion, famous people, and famous buildings). Oya (2015) used 5,000 sentence pairs randomly chosen from the subcorpus on Buddhism (26,890 English-Japanese pairs in total) for a different purpose; the present study covered the Japanese sentences containing -to natteiru in the same subcorpus.

3.2 Extracting sentences containing -to natteiru
Both the English and Japanese sentences in the parallel corpus were parsed by two parsers: the Stanford Parser (de Marneffe and Manning 2006; 2008; 2015) for the English sentences, and Juman/KNP (Kurohashi and Nagao 1992; 1994; 1998; Kawahara and Kurohashi 2007) for the Japanese sentences. Using Juman, a morphological analyzer for Japanese, an input sentence was divided into morphemes. Juman’s output was then imported to KNP, a dependency parser for Japanese, which outputs the dependency among words in the input sentence. Then, Japanese sentences containing -to natteiru and their English translations were extracted from the parser output by a simple Excel query. As a result, 2,325 sentences were found to contain the -to natteiru construction.

3.3 Analyzing representative examples in terms of their syntactic mismatches
Because it is impossible to present all the Japanese-English pairs containing the construction -to natteiru in this paper alone, let us consider some of their representative instances. This is a first step to constructing a more comprehensive list of the correspondence patterns of the construction -to natteiru in Japanese and its English counterparts.

Each instance is analyzed in terms of which of the types of syntactic mismatch (introduced above) are found in their typed-dependency representations, because the list is also intended to function as a linguistically principled reference. The assumption is that we can categorize the Japanese-English pairs in terms of the number of syntactic mismatches and this information can be used for the benefit of language learners. In other words, we can align Japanese-English pairs in the order of the structural discrepancies between them in terms of the number of syntactic mismatches (provided that the weight of these mismatches are the same), and it is advisable to show these learners Japanese-English pairs that contain fewer syntactic mismatches first, followed by pairs that contain more syntactic mismatches.

Caveat: It is self-evident that all the English definite and indefinite determiners are instances of function-word introduction, because Japanese does not have any. It can be said that the same is true for English pronouns, because it is often the case that Japanese sentences elide pronouns. In addition to this, the dependency types used in Japanese and those in English are not the same, and this means that there are many instances of conversion; hence, these syntactic mismatches are not mentioned later in this analysis, unless they are worth special attention.

First, let us consider the pair in (13) (= (8)) below:
Figure 9 The typed-dependency trees for (13).

Notice that the English counterpart does not contain the verb ‘become,’ the most basic counterpart for the Japanese verb *natteiru*. Rather, the main verb of the English counterpart is ‘is.’ The majority of sentence pairs found in the subcorpus contain this type of correspondence. This reflects the linguistic insight that Japanese language prefers *naru/natteiru* ‘become/be becoming’ where English does not use ‘become/be becoming.’ (Ikegami 1981).

The dependency relationship between *genzai-wa* ‘today-topic’ and *natteiru* is not retained in the English counterpart; the word ‘today’ depends on the past participle ‘accepted,’ as an instance of *Branch transposition*.

The dependency relationship between *teisetsu-to* and *natteiru* corresponds to that between ‘is’ and ‘accepted’; the noun *teisetsu-to* ‘accepted.opinion’ corresponds to a past participle ‘accepted.’ This instance is a kind of *lexical fission*, where one word in the source language is divided, yet only part of it corresponds to the word in the target language, and the rest is missing (or elided because of repetition; e.g., ‘opinion’ in the example above).

Next, let us consider the following example (14)(=9):

(14)

*kugatsu-ni okonaw-areru-no-ga tsuurei-to natteiru*

September-part hold-passive-part-part usual-part become-present

‘It is usually held in September.’
The dependency tree of the target sentence in (13) above preserves the structural setting of the source sentence, except for the target subject ‘It’ and the preposition ‘in.’ Each word in the target sentence semantically corresponds to that of the source sentence, yet their parts of speech are different. First, the source noun phrase *tuurei-to* ‘usual-part’ corresponds to the target adverb ‘usual.’ Second, the source noun phrase *okonaw-areru-no-ga* ‘hold-passive-part-part’ corresponds to the target past participle ‘held.’ In other words, there are part-of-speech changes along with functional-word introduction.

Next, let us consider (15)(=(10)) below; this example includes the syntactic mismatch *head switching* and *lexical fusion*:

(15)

```
shiryo-toshite kichou-na mono-to natteiru
documents-part valuable-part thing-part become-present
‘They serve as valuable documents.’
```

*Head switching* occurs between the source and target sentence; the source dependency relationship between *shiryo-toshite* ‘documents-as’ and *kichou-na* ‘valuable’ is reversed in the target dependency (the word ‘valuable’ depends on ‘documents’). In other words, the predicative adjective *kichou-na* in the source sentence is converted into the attributive adjective ‘valuable.’ In addition, the source word *mono-to* ‘thing-part’ does not correspond to any word in the target sentence. This is an instance of *lexical fusion*, where the word *moto-to* fused with the target word ‘documents.’

Next, look at the pair (16)(=(11)) below. This example includes the syntactic mismatch *Branch transposition*, along with others:

(16)

```
Sono zou-wa monocoque-ni chikai kouzou-butsu-to natteiru.
the statue-topic monocoque-part close structure-thing-part become-present
‘The structure of the statue is close to monocoque.’
```
The dependency relationship between zou-wa and natteiru is not retained in the English counterpart; the word ‘statue’ depends on the preposition ‘of,’ which is introduced into the English dependency tree (function-word introduction), and this preposition depends on the word ‘structure.’ This is one instance of Branch transposition. The dependency relationship between chikai ‘close’ and kouzou-butsu-to ‘structure-thing-particle’ is not retained in the English counterpart; the word ‘close’ depends on the root verb ‘is.’ This is another instance of Branch transposition found in this sentence.

Lastly, let us consider the pair (17)(=(12)) below:

(17) juyo-bunka-zai-to natteiru.
important-culture-property-part become-present
‘It has been designated as an important cultural property.’

Along with the lexical fission of juyo-bunkazai-to into ‘important cultural property,’ this example contains a syntactic mismatch of interest. The word ‘designated’ in the target sentence does not correspond to any word in the source sentence; this is an instance of content-word introduction, which is not included in Mel’čuk and Wanner (2006). This introduction seems to be driven by the semantics of the word juyo-bunkazai ‘an important cultural property.’ As a matter of fact, things do not become cultural properties; they are designated as such by certain institutions such as local or national governments. The Japanese
word *natteiru* itself can cover the meaning ‘has been designated’ when it takes as its complement something that should be designated. When translating such sentences, then the word ‘designated’ must be introduced. The same scenario applies to the introduction of the word ‘registered.’

The types of syntactic mismatch found in the example pairs above are shown in the following table. The sentence pairs above are aligned from (13) to (17) according to the number of syntactic mismatches:

<table>
<thead>
<tr>
<th>Conversion</th>
<th>(13)</th>
<th>(14)</th>
<th>(15)</th>
<th>(16)</th>
<th>(17)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head switching</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexical fusion</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexical fission</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part-of-speech change</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Function-word introduction</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Content-word introduction</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Branch transposition</td>
<td>1</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

4. Discussion

All other things being equal, it can be said that we can align Japanese-English pairs in the order of the structural discrepancies between them. As far as the example pairs in the previous section are concerned, (17) shows larger discrepancy than (13) does. The validity of this claim remains to be proven through future research, and the scenario will be as follows: First, Japanese leaners of English are provided with Japanese-English translation pairs of different structural discrepancies in terms of the number of syntactic mismatches. Second, they are asked to answer the question concerning which pairs are more difficult than others (how to construct such questionnaires should be discussed in future research). If their answers match the syntactic discrepancies in terms of the number of syntactic mismatches, then it will be safe to conclude that the claim is not rejected.

Through such a scenario mentioned above, it may be ascertained that the syntactic mismatches are not equal; in other words, learners may consider some syntactic mismatches more difficult than others. For example, learners find Japanese-English pairs that contain head switching more difficult than those that only contain function-word introduction, since the latter preserves the structural setting of the source sentence. We need to conduct further research to say for sure that this is the case, or which is more difficult than which at present.

The frequency of each syntactic mismatch with respect to *noun-to natteiru* in the parallel corpus is not investigated in this research, because such a simple frequency is less informative than the information that can be obtained through the scenario mentioned above. For example, the fact that Function-word introduction is more frequent than Head switching does not necessarily mean that the former is more difficult for learners to learn than the latter, or vice versa.

Putting aside the alignment of Japanese-English pairs in terms of the degree of syntactic discrepancy, the information about each of the constructions in the source language and its target-language constructions corresponding to the source constructions will serve as a valuable reference for the learners of the target language. The future direction with respect to this aspect of the present research is to explore more of the other constructions in the source language (Japanese in this case) and its one-to-many correspondence with a target language (in this case English), or with target languages. The results will be compiled into a
construction dictionary that contains linguistically and pedagogically significant information. Again, we need further investigation on the constructions contained in source languages and their corresponding construction in target languages.

5. Conclusion

This study explored the possibility of building a dependency-based Japanese-English construction dictionary, with reference to a Japanese construction noun-to natteiru where there is no one-to-one correspondence between the words in a source sentence and the words in a target sentence. Mel’čuk and Wanner’s (2006) idea on syntactic mismatches between a source language and a target languages is applied into Japanese sentences containing the construction noun-to natteiru and their English translations in a Japanese-English parallel corpus, and the result suggests that these syntactic mismatches can be part of important information to be included into a Japanese-English construction dictionary. Finally, observing only a tiny fraction of sentence pairs has reminded us of the simple truth that one-to-one correspondence between source words and target words is an illusion. In order for us to emancipate ourselves from this illusion, and to compile future dictionaries of real significance, we need to observe the linguistic facts concerning constructional aspects of, and syntactic mismatches between, source and target languages. I hope this article serves as one of the first attempts to explore them.

Acknowledgments
This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant Number 26370500.

References


The Crowdsourced Compilation of the *China English Dictionary*

**Qin Xiaohui**  
University of Science and Technology Beijing, China  
qinxiaohui@ustb.edu.cn

**Abstract**  
From an Anglo-Saxon dialect to a world language, English has displayed incredible tolerance, extensiveness and variance in the past 1500 years. Currently, in the context of English globalization and localization, increasing linguistic and sociolinguistic attention has been paid to the structure and use of non-native English varieties. Starting from the 1980s, the academic world saw the legitimacy of China English as a new English variety, to be established and consolidated in the sustaining discussion and debates by domestic and overseas researchers. It is widely agreed that lexical features are the most prominent evidence of an autonomous variety. The world has witnessed the compilation of English variety dictionaries in inner circle and outer circle countries since the 19th century. China English words have been melting into English for thousands of years and have been recorded in authoritative Monolingual English dictionaries and Chinese English dictionaries for hundreds of years. This paper introduces the lexicographical principles and practice of the online *China English Dictionary* (bilingual English-Chinese) in a crowdsourcing mode (a distributed problem-solving and production model in which solutions are produced from voluntary Internet users, also known as the crowd), with a focus on its genesis, macro-structure, micro-structure and mass contribution mechanism.

**Keywords:** China English Dictionary, crowdsourcing, China English variety, lexis

1. **Introduction**

For more than two millennia, successive languages have formed a chain of lingua francas: Sanskrit, Greek, Latin, Arabic, and French are examples at various periods and in different parts of the world. However, there has never before been a single language which spread so widely as English has done in the past century. One thing is clear in the trend: the more regions English is extending to, the less its development is determined by the usage of its native speakers. In other words, the control of the English language is passing to non-native speakers of English-using communities in various continents, and the Englishes they speak and use are nativized varieties each with its own flavor and characteristics. This paper is directing at one of these New English Varieties, namely, China English, and hopes to reveal some of its recognizable features.

2. **Genesis of the project**

Thirty years ago, China English was identified by Karchu (1982, 1983, 2005) as one of the performance varieties in the expanding circle. However, it is in these past three decades and in the expanding circle that the most extensive spread of English in terms of numbers of speakers has occurred. Of all the exciting new stars, China is undoubtedly the most shining one. Wei and Su (2012) have reported 390 million people who have learned English, and Bolton and Graddol (2012) stated 400 million English learners in China, approximately a third of its population. Starting from the 1980s, the academic world saw the legitimacy of China English as a new English variety be established and consolidated in the sustaining discussion and debates by domestic and overseas researchers. (Ge 1980; Cheng 1982; Cannon 1988; Sun 1989; Wang 1991; Li 1993, 2012; Jiang 1995; Du & Jiang 2001; Bolton 2003; Jenkins 2014).
In terms of linguistic features at various levels (phonetics, phonology, grammar, semantics, lexis and discourse), it is generally believed that lexical features are the most relevant rationale for the autonomy of an English variety. Thus, the dimension of lexicography and dictionary compilation is of singular importance in the codification and recognition of World Englishes. Since the 19th century, inner circle countries (US, Canada, Australia, New Zealand etc.) have compiled their English dictionaries and outer circle countries (Singapore, South Africa, Jamaica etc.) have followed their steps (Béjoint 2010, Bolton 2003). Chinese loanwords have been melting into the English language for hundreds of years (OED 1933), and have been included sporadically in classic English dictionaries and authoritative bilingual (Chinese-English, especially) lexicons since the appearance of A Dictionary of the Chinese Language (Morrison 1815-1823). So far there has not been a comprehensive review of China English vocabulary. It is in this context that the present project is initiated. Specifically, it is aiming to build a dynamic China English word inventory in a crowdsourcing mode.

3. The crowdsourcing mode

The term “crowdsourcing” is a portmanteau of “crowd” and “outsourcing”. It is a distributed problem-solving and production model in which solutions are produced from voluntary Internet users, also known as the crowd. This practice was originally applied in software and service industries and currently expanded to various fields of society (Howe 2006, 2008; Cove 2007; Brabham 2008; Callison-Burch & Dredze 2010; Callison-Burch & Zaidan 2011).

Crowdsourcing combines the efforts of self-identified volunteers or part-time workers, where each contributor adds a small portion to the greater result. The marriage of crowdsourcing and lexicography is an old one. The earliest example, the Oxford English Dictionary (OED, 1857-1928) is compiled on the basis of 6 million quotations solicited from the public. The modern representative, the Wiktionary, is written collaboratively by volunteers. Its wiki software, MediaWiki, allows almost anyone with access to the website to create and edit entries. It is in this fashion that the China English Dictionary is intended and constructed.

4. The China English Dictionary

4.1 Technological framework

The China English Dictionary (which can be accessed at www.chinese.org) is based on MediaWiki 1.23.2 with a framework of Nginx+PHP+Mysql. It is built upon Ubuntu 14.04 (kernel version: 3.13) and deployed on the ECS service of Ali Cloud for Internet Service. The overall system architecture is:
4.2 Macro-structure
As an online dictionary, the nature of digital media has made irrelevant all details of the framing structure of a traditional paper dictionary, what the sections are, or in what sequence. The China English Dictionary is freed from the print space limit which has haunted paper lexic on editors for hundreds of years. Therefore, it sets no limitation to the number of entries, and absorbs new words as soon as they appear. In other words, it is never a static volume but always an in-progress project. The dictionary is intended for Chinese-English bilingual speakers in the world and English learners, users and researchers in China. Any words of Chinese origin which have entered the English language through transliteration, translation or coinage are potential lemmas waiting to be sought, edited, reviewed and expanded. The data sources are mainly published materials reflecting Chinese-unique objects and concepts written by proficient bilingual speakers, and include public or official English newspapers in China, Chinese-English dictionaries, English monographs, literature written (translated) by Chinese, translated classic works, English broadcasting and TV texts in China. In terms of lemmatization, the user-friendliness principle is at work, that is, all words, compounds and phrases will be treated as independent headwords.

4.3 Micro-structure
The internal structure of each entry in the China English Dictionary provides a maximum model which may not be fully needed in every single case. The outline of the model is presented in the following: spelling, pronunciation, part of speech, other forms, origin Chinese, definition, examples, derivatives and phrases, among which part of speech, origin Chinese, examples are mandatory information. The “examples” consists of quotation, its publishing year, the author, title, source and page number. This is in the fashion of the Oxford English Dictionary for the same purpose of documenting the development of language through solid quotation evidence.
Figure 2 A sample editing page the entry “besiege Wei to rescue Zhao”

4.4 The collaborative mechanism
All the words are contributed, edited and revised by registered volunteers. A letter entitled “China English Needs You” on the website’s main page illustrates generally what China English words are, where we can find China English words and how we edit a China English word entry. Subsequent training manuals and rules will be released for editors as soon as common problems are detected and new technical codes developed. In China English Dictionary, all edits are tracked and listed chronologically in the history tab. This ensures that good articles are stored and readily accessible and bad edits are quickly revised. We assume most participants are good willed and responsible. At this stage, whenever anything new is added to the dictionary, a team of reviewers (or anyone who is interested) will check it and make sure it conforms to our guidelines and policies. In case that vandalism does occur culpable parties will be blocked by administrators (elected volunteers that have demonstrated reliable editing qualities and steady performance levels).
5. Concluding remarks and future work

Variation in language structure and use is one of the prime characteristics of human language. The present research, variation-oriented, aimed to examine the lexical features of the new major China English variety by constructing a crowdsourcing-based dynamic platform to document the development of China English words. The trial operation points to three urgent issues in need of address: (1) comprehensive lexicographical policies including the status and verification of a China English word entry; (2) technological advances such as audio pronunciation, automated detecting programs, and community codes such as voting rules and a hierarchy of editor positions; and (3) a more efficient quality control mechanism needs to be specified which is critical to maintain the integrity and sustainability of dictionary.

Acknowledgements
This research is supported by the National Fund of Philosophy and Social Sciences and Fundamental Research Funds for the Central Universities.

References


A Dictionary of Hong Kong English vs The GloWbe Corpus?:
Triangulating sources of lexicographic evidence

Vincent B Y Ooi
National University of Singapore
vinceooi@nus.edu.sg

Abstract
This paper addresses the general question of reconciling differing and competing sources of evidence for ensuring that both frequency of occurrence and ‘socio-cultural salience’ are achieved in codifying a particular variety of English. In so doing, it compares the efforts of Cummings and Wolf (2011) and Davies and Fuchs (2015) with regard to Hong Kong English. Cummings and Wolf (2011) look at various sources of evidence for their selection of entries in the dictionary, and lexical items are also given a frequency rating based on their ‘brief Google.hk surveys’ (the highest two tiers having more than 1000 occurrences on Google.hk). In this paper, ten selected lexical items from Cummings and Wolf (2011) that have these high frequencies are compared with their respective (non-) occurrences in the Hong Kong component of the GloWbe corpus that contains 1.9 billion words from 20 different countries (Davies and Fuchs 2015:2). The texts in GloWbe are taken from various web pages and blogs that presumably represent both currency and socio-cultural salience. The selected lexical items for discussion include the following: Mid-levels, dim sum, jetso, iron rice bowl, yum cha, wushu, kowtow, qigong, laisee and Legco.

Keywords: lexicographic evidence, corpus, frequencies, GloWbe Corpus, Hong Kong English, triangulation

1 Introduction

Meyer and Nelson (2006) detail the types of linguistic data – introspection, experimentation, and corpus – and some methodological issues that these types of linguistic evidence bring about in support of linguistic claims. Such types of linguistic data may similarly count as lexicographic evidence for the codification of lexemes in the dictionary. However, the proportion of these types of lexicographic evidence still varies among lexicographers, and there is no necessary uniform practice for the corroboration of such varying sources of lexicographic evidence in order to achieve the valued social sciences notion of ‘triangulation’.

In this paper, some of these triangulation issues are discussed in terms of the codification of 10 lexemes in The Dictionary of Hong Kong English (Cummings and Wolf 2011) and their corresponding (non-) occurrences in the Hong Kong component of a recently completed mega corpus, The GloWbe Corpus (Davies and Fuchs 2015). While it is obvious that one is a dictionary and the other a corpus, this paper hopes to show that both enterprises may benefit from and complement each other in the quality of their linguistic claims.

2 The Dictionary of Hong Kong English (DHKE)

Why single out The Dictionary of Hong Kong English (henceforth ‘DHKE’) for our discussion on triangulation? Reasons include its methodology and various reviews by well-known linguists and lexicographers.

In its introduction, the DHKE lists itself as the first reference dictionary of Hong Kong English and that “only words and word senses that are particular to HKE (Hong Kong English) or have a specific reference to Hong Kong are listed.” (p. xii) However, ‘particularity’ does not mean ‘exclusivity’, and Cummings and Wolf acknowledge that these
terms may be similarly used in another (Asian) variety of English. The selection of entries is made “from a variety of sources”, including English-language newspapers based in Hong Kong, cartoons and literary works by local authors, government information, local student essays, spoken language heard on TV, and internet websites. The authors mention having consulted various literature on Hong Kong English, including “short lists” of Hong Kong English words (encapsulated in Bolton 2003). Also, while the authors mention the use of frequency information from the Hong Kong component of the International Corpus of English (Greenbaum and Nelson 1996), they seem to have mainly used “brief Google.hk surveys…at different times in the year 2010” (p. xxvi) The top band includes “more than 10,000 occurrences on Google.hk” and the second next most frequent band includes between 1000 and less than 10,000 occurrences on Google.hk. While giving readers an idea of the frequency of the items, Cummings and Wolf (2011) also acknowledge the limitations of such an approach: the data from Google.hk are not exclusively curated for Hong Kong usage alone, and the salience of items may vary over time in the media. Nonetheless, readers should still be able to “fathom the importance of widespreadness” of the items included in the dictionary.

In other words, what has been employed is a mix of introspection, citational evidence (casual occurrences in various print and spoken sources), a curated dataset (the ICE corpus), and a non-curated dataset (“brief Google.hk” surveys). While the sources of citational evidence are listed in DHKE, the proportions of these sources of evidence are not detailed. At the end of the day, it would very much depend on the lexicographers’ overall experience and judgment on the inclusion and treatment of these entries in the dictionary. In this regard, Salazar (2014: 103) cites an article of mine (Ooi 2001) for necessitating both etic (outsider) and emic (insider) perspectives in lexicography. Salazar concurs with me that such perspectives are necessary for “the accurate description of the lexicon of world Englishes, including those of the Southeast Asian region.” The distinction between ‘etic’ (from ‘phonetic’) and ‘emic’ (from ‘phonemic’), first introduced by the late linguist Kenneth Pike, is to offer a methodological solution to the problem of objectivity in the social sciences. Applying this distinction to lexicography, knowing the lexicographer’s etic and emic perspectives, i.e. where s/he is ‘coming from’, would provide greater understanding regarding the inclusion, exclusion, and treatment of entries in the dictionary. Applying this distinction to Cummings, it would seem that the etic experience is somewhat complemented by a somewhat emic one of having been in Hong Kong for 10 years; for Wolf, who is introduced as living in Germany, it would very much seem to be mainly an etic experience of Hong Kong. And, in my case, it is one of knowing Malaysian Cantonese natively, learning Malay (and English) in school in Malaysia and Singapore, being exposed to both Mandarin and Cantonese Chinese in Singapore for more than half my life, being a frequent visitor to Hong Kong during the past 10 years or so, and interacting with cousins in Hong Kong.

So, in reading through the approximately 460 entries selected for DHKE, one can safely say that a number of them are certainly shared by both Malaysian and Singaporean English. Such non-unique entries in DHKE include iron rice bowl (a term quite popular in Singapore at one time), pajamas (U.S. spelling) or pyjamas (British spelling), bird’s nest, Hakka, cheongsam, and bubble tea (which, anyway, is a term borrowed from Taiwan English). And, while DHKE asserts that ketchup originates from either Hong Kong or Chinese English, both The Merriam-Webster Dictionary and Dictionary.com cite Malay (‘kechap’ – fish sauce) as the origin.

Notwithstanding such friendly ‘ownership’ disputes, the 10 lexemes for examination in this paper will be based on the top two frequency bands used by the DHKE. This approach is meant to offer a more up-to-date basis for comparison with the Hong Kong component of the GloWbe corpus which was completed in 2013.
The Global Web-based English Corpus (henceforth, ‘GloWbe’) contains approximately “1.9 billion words of text from 20 different countries, which includes six Inner Circle and 14 Outer Circle countries” (Davies and Fuchs 2015:2), based on Braj Kachru’s well-known distinction between countries in which English functions as a Native Language (hence, ‘Inner Circle’ countries) and those that have English as a Second Language (hence ‘Outer Circle’ countries). These 20 countries include the United States, United Kingdom, Canada, Australia, Ireland, New Zealand, India, Sri Lanka, Pakistan, Bangladesh, Singapore, Malaysia, the Philippines, Hong Kong, South Africa, Nigeria, Ghana, Kenya, Tanzania, and Jamaica.

In order to achieve the current size of the corpus, the only possible source is web pages (see also Kilgarriff and Grefenstette 2003). Davies and Fuchs (2015) also explain that 60 per cent of the corpus comes from blogs, so that informal language is represented “fairly well”. The remaining 40 per cent consists of “more formal genres and text types.” Also, the curation of this corpus goes beyond what has been achieved in Cummings and Wolf’s brief surveys of google.com.hk. Davies and Fuchs (2015: 4) say that they use Google’s Advanced Search and “Limiting by Region” features in order to ensure that the web pages are correctly associated for the respective countries. They also remove “boilerplate” material (recurring headers, footers, sidebars etc.), tag the entire corpus, and then import the texts into a database such that these texts from the various countries can be compared side-by-side in a normalized manner i.e. frequency per million words.

4 Discussion of 10 lexemes from DHKE

As mentioned in Section 2, lexemes that will be discussed in this section are first filtered from the top two frequency bands (reflected at the bottom right of each DHKE figure in this section). In addition, the terms are selected based on what is felt about their distinctiveness relative to other Asian varieties of English, notably both Malaysian and Singapore(an) English that I have first-hand knowledge of. This ‘feeling’ will then be gauged more objectively by means of the GloWbe corpus.

4.1 Mid-levels

The DHKE lists the term as follows (see Figure 1):

Figure 1 Mid-levels (DHKE)

Let us compare the corresponding evidence from GloWbe (see Figures 2 and 3):

Figure 2 Mid-levels (GloWbe)
GloWbe’s Mid-levels (ranked the highest in C&W) occurs 240 times for Hong Kong (out of the 271 occurrences world-wide) – hence, there is a close fit between GloWbe and google.hk. In other words, we can say that this term is very much particular/distinctive to Hong Kong English. In addition, the concordance listing indicates that the term indicates richer associations of ‘premium areas’, ‘preferred living quarters for expats’, ‘prime locations’ and even ‘mums’ – beyond the mere collocate of ‘Western’ in DHKE.

### 4.2 Dim sum

In the GloWbe corpus, another popular term dim sum occurs only 529 times (for Hong Kong) out of 1584 times world-wide, showing that the term has spread to other countries such as the UK, U.S., Canada, Malaysia and Singapore (see Figure 4).

DHKE has a section entitled “Hong Kong English words now used internationally”, and the term dim sum is not in this section; the evidence from GloWbe shows otherwise. The term is also found in both Dictionary.com and The Merriam-Webster Dictionary, with the origin listed as “Chinese (Guangdong)”. The Guangdong part is corroborated in the corresponding DHKE entry (see Figure 5), in which the source language is indeed Cantonese:

```
dim sum

Source language: Cantonese (広東話).
Definition: a variety of different traditional Cantonese foods usually of a size to be picked up by chopsticks. Examples include har gow and shumai.

Text example:
They ordered a tea, and a steamer basket of dough, the only dim sum the kitchen had left.

Note: The spelling tim sum is obscure in HKE, but more common in Singapore.
```

DHKE further claims that the alternative spelling tim sum is “more common in Singapore.” On checking GloWbe, there are only 2 and 4 instances of this term from Singapore and Malaysia respectively. Instead, the preferred term in Mandarin-speaking Singapore is dian xin (which occurs 5 times in Singapore, and 3 times in Hong Kong respectively).

### 4.3 jetso

At the other end of the spectrum, there is an apparent conflict between google.hk and GloWbe, with the term jetso (‘a bargain or special discount’, Cummings and Wolf 2011: 86) ranked the highest but not represented at all in the GloWbe corpus (see Figure 6):
It would seem that Hong Kong bloggers (at least on the surface of it) do not tend to use the term *jetso* to talk about bargains or special discounts. However, the term *discount* in GloWbe shows that Hong Kongers do love their discounts, comparing favourably with a number of other countries, i.e. Hong Kong (1528), Singapore (1367), Malaysia (1308), the Philippines (1110) and South Africa (762).

### 4.4 iron rice bowl

DHKE codifies this term as follows (see Figure 7):

In Figure 7, the given example is not quite salient, as it is contained within quotation marks. Beyond this, there is another entry section entitled “underlying conceptualization (generic)”. As a reference dictionary for the lay reader, it would certainly be strange to include Lakoffian conceptual metaphors for which the reader could be confused regarding the ways in which the underlying meaning is to be decomposed. In another entry (not included here because of its low frequency), the term *mafu* (essentially a pimp) is said to be related to *mafoo* (a stable boy) and so the underlying conceptualization “sex is riding” makes one do a double take on the way in which this sentence is to be read and understood. Therefore, one must agree with Béjoint’s (2011:478) comment that such notes are expressed “in a language that is too formulaic and abstract for the users of such a dictionary, even though the information to be conveyed is relatively simple: basically, they say that ‘X is like Y’.”

Turning to GloWbe, my belief that the term is found in Singapore English is borne out (see Figure 8):

More interestingly, the term occurs 9 times in the UK component (see Figure 9):
From this, one can glean that the term is not unknown to readers of *The Guardian* broadsheet.

### 4.5 yum cha

The lexeme *yum cha* (literally ‘drink tea’) is codified in the DHKE as follows (see Figure 10):

> **yum cha**
> 
> *Yum cha.*
> 
> **Source language:** Cantonese (広東話).
> 
> **Definition:** to have teatime in a Chinese style; 2 to drink tea.
> 
> **Textexample:**
> 1. “Although *yum cha* and *dim sum* are used synonymously in Hong Kong. There is, properly, a distinction—when you go to a restaurant to *yum cha*, you eat *dim sum* or, when you *yum cha* you drink tea and eat *dim sum*.”
> 2. “*Yum cha*, a term in Cantonese, literally meaning ‘drinking tea,’ refers to the custom of eating small servings of different foods while sipping Chinese tea…”

![Figure 10 yum cha (DHKE)](image)

Before commenting on Figure 10, let us also take a look at the corresponding frequencies in GloWBE (see Figure 11):

<table>
<thead>
<tr>
<th>CONTENT</th>
<th>ALL</th>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>IE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>UK</th>
<th>IE</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NL</th>
<th>GR</th>
<th>KZ</th>
<th>TZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHA</td>
<td>107</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>74</td>
<td>15</td>
<td>11</td>
<td>16</td>
<td>10</td>
<td>49</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

![Figure 11 yum cha (GloWBE)](image)

The combination of both Figures 10 and 11 indicates the term *yum cha* is indeed a well-loved activity in Hong Kong. Among Singaporean and Malaysian Chinese, this activity is no less loved but expressed in terms of going to the tea house and eating *dim sum* or *dian xin*. The accompaniment of hot tea is usually a given, hence the apparent distinction between *yum cha* and *dim sum* in the DHKE textual examples do not quite apply in Southeast Asia. More surprising, though, the term *yum cha* occurs 74 times in Australia compared with 49 times in Hong Kong; it is certainly the preferred term to *dim sum* in the Antipodean country. An explanation for this is not only the high numbers of Asian migrants (from China, Hong Kong, Taiwan, Singapore and Malaysia) to Australia but the activity being more blogged about there (as a cursory concordance inspection of the 74 occurrences there shows).

### 4.6 wushu

The term *wushu* is generally understood as the Mandarin equivalent of *kung fu*, traditionally its more codified counterpart (whose origin is Cantonese) in dictionaries. In Ooi (2001), I noticed that the term *kung fu* was codified in a number of English learner dictionaries then, compared with *wushu* which was not. Nowadays, both *The Merriam-Webster Dictionary* (for U.S. English) and *Dictionary.com* (a popular dictionary found also in smart phones) indicate that both terms are embraced in such native English contexts. In the DHKE, however, the term *kung fu* is relegated to the list of “Hong Kong English words now used internationally” (which, actually, is also conflated with “Chinese English” words) while the term *wushu* is found in the main text (see Figure 12):
wushu

Source language: Mandarin (щит).

Definition: a martial art similar to kung fu.

Text example:

"If you wushu medals at one level authenticate him, there are purists who question whether wushu (designed for performance, not combat) is a ‘proper’ martial art.

Figure 12 wushu (DHKE)

While this definition somewhat differs from the one found in the other two dictionaries (with “wushu” meaning Chinese martial arts collectively), the spelling as two words in the textual example is out of sync with the lexical entry which indicates its preferred spelling as one word (not two). Also, turning to GloWbe (Figure 13), we can see that the term wushu is by no means unique to Hong Kong English:

GloWbe therefore indicates that the term is known in countries as far as Sri Lanka (abbreviated as LK) and New Zealand (abbreviated as NZ).

4.7 kowtow

In checking out the entry for kowtow and its accompanying illustration in DHKE (see Figure 14), it is surprising to learn that the meaning also includes the first sense, i.e. the idea of ‘knocking [one’s] bent index and middle fingers on the table to express gratitude to the person who served the tea.”

kowtow

Source language: Cantonese ( nominate) or Mandarin.

Definition: 1 a finger-based method of saying “thank you” to the person serving tea; 2 a form of Chinese bowing where the forehead is knocked on the floor to (a) God or an altar; (b) a political leader (archaic); 3 a metaphor for political submission.

Text example:

1 “After a person’s cup is filled, that person may knock his bent index and middle fingers (or some similar variety of finger tapping) on the table to express gratitude to the person who served the tea.

2 The attendance of the elders and the gentry was compulsory, while those over sixty were invited as guests of honour. The kowtow and the three prostrations were in the order of the government officials first, then the gentry, then the elders, then whoever happened to be there.”

3 Relations between Seoul and Washington have been tense at times under Roh, a progressive elected on a wave of anti-Americanism in December 2002 after he vowed never to kowtow to the United States.”

Figure 14 kowtow (DHKE)

This practice of tapping one’s index and middle fingers on the table as an expression of polite gratitude to the person refilling one’s teacup is commonly practised among Chinese Malaysians too. The practice is said to have begun with the Qianlong Emperor who used to travel incognito and, when he poured tea for his retinue, would receive their tacit gratitude through the use of this gesture.

A detailed examination of the entire Hong Kong concordance in GloWbe fails to turn up this first textual meaning (see Figure 15):
Instead, its metaphorical usage (in the sense of “political submission”) seems to be used more in a number of countries, including the United States (Figure 16) for which *kowtow* occurs 7 times more than Hong Kong:

<table>
<thead>
<tr>
<th>CONTEXT</th>
<th>ALL</th>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>DE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>FR</th>
<th>ES</th>
<th>IN</th>
<th>PK</th>
<th>ID</th>
<th>SG</th>
<th>MY</th>
<th>PH</th>
<th>HK</th>
<th>ZA</th>
<th>NL</th>
<th>GR</th>
<th>SE</th>
<th>BE</th>
<th>TR</th>
<th>IN</th>
</tr>
</thead>
<tbody>
<tr>
<td>KOWTOW</td>
<td>476</td>
<td>120</td>
<td>11</td>
<td>79</td>
<td>10</td>
<td>37</td>
<td>49</td>
<td>13</td>
<td>18</td>
<td>8</td>
<td>6</td>
<td>24</td>
<td>15</td>
<td>17</td>
<td>5</td>
<td>7</td>
<td>25</td>
<td>6</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 16 *kowtow* (GloWbe)

With its well-known usage in Inner Circle countries, the term *kowtow* is well-codified in various native English dictionaries and so (like *typhoon*) is no longer a mere Hong Kong English feature. Following the DHKE structure, it should have been placed in the section entitled “Hong Kong English words now used internationally.” Indeed, this section is unnecessary – the words there should have been merged with the ones in the main text, with the indication that such words have now become standard English ones.

4.8 qigong

The preceding remarks regarding the standardness of *kowtow* also applies to *qigong*, which is listed in DHKE as follows (see Figure 17):

<table>
<thead>
<tr>
<th>qigong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source language: Mandarin (淸,功), 微信 (功).</td>
</tr>
<tr>
<td>Definition: <em>set of breathing and movement exercises</em>.</td>
</tr>
<tr>
<td>Text example: <em>First you indulge in a Qigong session (the ancient Chinese art of energy work, $800) followed by a spot of Past Life Therapy ($5,000), in which your therapist will take you back to a previous life and help you complete any unfinished business.</em></td>
</tr>
</tbody>
</table>

Figure 17 *qigong* (DHKE)

The standardness of this term is well-attested in GloWbe, with an almost equal number of entries (to the Hong Kong ones) in the North American and British contexts (see Figure 18):

| CONTEXT | US | CA | GB | DE | AU | NZ | IN | FR | ES | IN | PK | ID | SG | MY | PH | HK | ZA | NL | GR | SE | BE | TR | IN |
|---------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| QIGONG  | 666 | 135 | 195 | 97 | 6  | 16 | 38 | 4  | 3  | 14 | 23 | 108 | 4  | 114 | 2  | 3  | 7  | 5  |  |

Figure 18 *qigong* (GloWbe)

4.9 laissez/ red packet

This Cantonese term for red packet (a cash gift placed in a red or pink envelope, during auspicious occasions such as Chinese New Year) is particularly Hong Kong English, whose main substrate influence is Cantonese. One can predict its occurrence in other Asian contexts which are heavily influenced by Cantonese, e.g. the Cantonese-speaking cities of Ipoh and Kuala Lumpur. In Singapore, which is influenced more by Mandarin and Hokkien Chinese,
the preferred terms are hong bao and ang pow respectively. Both Figures 19 and 20 show the entries for lai see and red packet respectively in DHKE:

**lai see**

*Source language: Cantonese (を見せ)
*Definition: 1 money given as a gift in a red envelope; 2 a gift of money intended as a bribe.
*Text example:
  1: "Children and young adults will find more money in their lai see packets this Lunar New Year, according to a survey."
  2: "Staff must not solicit ‘Lai See’ from any external business associates in any circumstances."
*Underlying conceptualization: A *BRIBE* IS A GIFT [TARGET DOMAIN Δ CORRUPTION] [SOURCE DOMAIN Δ GIFT].

Figure 19 lai see (DHKE)

**red packet/pocket (money)**

*Fixed expression:*
*Definition: see lai see.
*Text example:
  1: "Students are given some red packets (only one red packet contains a 10-dollar bank note)."
  2: "Anybody is [sic] still single loves the Chinese New Year because they are entitled to receive the red packet money from the married people."

Figure 20 red packet (DHKE)

While red packet lists lai see as its definition, its textual examples are wholly positive. However, lai see is also seen to have a negative nuance by being associated with its underlying conceptualization of bribery and corruption (“A bribe is a gift”).

Turning to the GloWbe corpus, the distribution of lai see is as follows (see Figure 21):

```
<table>
<thead>
<tr>
<th>CONTENT</th>
<th>ALL</th>
<th>US</th>
<th>CA</th>
<th>GB</th>
<th>DE</th>
<th>AU</th>
<th>NZ</th>
<th>IN</th>
<th>UK</th>
<th>NL</th>
<th>DK</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>lai see</td>
<td>96</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>11</td>
<td>4</td>
<td>3</td>
<td>11</td>
<td>8</td>
<td>26</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
```

Figure 21 lai see (GloWbe)

Without reproducing the attendant concordances here, the negative sense is indeed found in the Hong Kong context (e.g. “in corruption and ‘lai see’ taking”); in Singapore and Malaysia, though, there is no evidence of the bribe/corruption sense but instead there is the association with a token sum of money (e.g. “He did not take any salary, not even lai see”) or “lucky money” (“It is customary for married couples to give lai see”).

**4.10 Legco**

For me, this term – referring to the Legislative Council of Hong Kong -- is quintessentially Hong Kong English. Evidence that this term has not found its way to standard English is that it is not found in either Dictionary.com or *The Merriam-Webster Dictionary* (although one can encounter it in freedictionary.com). Strangely, though, the term is not found in DHKE at all, despite occurring overwhelmingly 602 times for Hong Kong English in the GloWbe corpus (see Figure 22):
Without again reproducing the attendant concordances from the various contexts, one can see that the media in various countries which use the term Legco mainly report the occurrences in Hong Kong (e.g. for India, there is the report on what “the President of the Legislative Council (Legco), Mr Jasper Tsang Yok-Sing, said.”) Of course, there are local uses in other countries such as Ghana, e.g. in the sentence “By 1960, LEGCO had an African majority.”

Focusing on blogs and web pages in the GloWbe corpus would mean that, in addition to local news and events, global matters are also written about.

5 Conclusion

In this paper, we have shown that triangulation of lexicographic evidence exercised for a project such as the DHKE can be more rigorous in the application of both emic and etic perspectives – and the recourse to a curated dataset, i.e. the GloWbe corpus. However, in its defence, the GloWbe corpus was not available in 2010 when the dictionary was being compiled. And, in itself, the GloWbe corpus does have its limitations (see *English World-Wide* 36). Davies and Fuchs (2015:26) acknowledge that the corpus does not include spoken material, and that “there may be speakers from other countries who may have posted to [the website of a particular country].” GloWbe is best used as one of the relevant tools for the discovery, comparison and codification of the various Inner and Outer circle varieties that it purports to represent. Finally, the DHKE should be congratulated as a fine – though first – systematic effort at dictionary-making that showcases Hong English to the rest of the world.

References


Davies, Mark, and Robert Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English corpus (GloWbe). *English World-Wide* 36:1, 1-29.


Accessed 20 May 2015


Proverbs in A Comprehensive Indonesian-English Dictionary

Dora Amalia
The Language Agency of Indonesia
yayak_amalia@yahoo.co.id

Abstract
Proverbs are highly culture-bound expressions and have specific lexicographical aspects. The main problem dealing with proverbs in bilingual lexicography is finding the right, one-to-one equivalents. The aspects are mainly related to the culture of SL and TL. This paper discusses how the proverbs are handled in an Indonesian-English dictionary, what kind of proverbs are selected, which methods are used to give cultural information to the user and some conclusion and suggestions relating to the case.

In Malay and western Indonesian, proverb has been very dominant in communication at least until 1970s. Young speakers of Indonesian are not so familiar with it and consider it as an old-fashioned way in today’s communication. It becomes very compelling to find proverbs in a comprehensive Indonesian-English Dictionary compiled by two American lexicographers published in 2004. Although the exact number of proverbs in the dictionary is not clearly stated, with a brief review there are some that are not found in Indonesian Big Monolingual Dictionary (KBBI).

This critical evaluation of the dictionary investigates the function of proverbs to give enough cultural information about Indonesian. The research is done by collecting proverbs from three headwords, i.e. air ‘water’, batu ‘stone’ and bunga ‘flower’. There 52 proverb samples from the lemmas, and only 16 of the proverbs have one-to-one equivalents, whereas the others have only explanation and literal translation.

Keywords: proverb, culture-bound praseology, bilingual lexicography

1. Introduction

Proverb is generally considered as a brief saying in common use that conveys a moral. It couches conventional wisdom in clever form and imagery, thereby making it memorable and easy to pass on from one generation to another. In certain parts of the world, however, the equivalents of the word ‘proverb’ do not refer only to brief, witty saying, but also to extended allusions like the parable, anecdote, or any series of allusive statements cited to demonstrate a lesson in discourse.

In Indonesian culture, proverbs are especially used in predominantly illiterae societies. They live especially in traditional occasion featuring use of indirectness, of which proverbs are a typical example. Indirectness constitutes a polite way of saying or asking for something. Although it is diminishing nowadays with the progressive acquisition of western education, for some Indonesian native speakers, especially the elder speakers, telling something directly, straight to the point, is still considered rude. That is why, proverbs are very familiar among them. Young speakers of Indonesian rarely use this kind of idiomatic expressions in their daily communication.

2. Proverb as cultural truth, moral percept and persuasion

Proverbs state cultural truisms (Yankah 1994). The truism may be in the form of an empirically valid statement, or an existing superstition or social norm. It may even have a questionable logic, or make an unverifiable claim. In any case, the proverb’s cultural validity is hardly disputed.
The proverb may be either prescriptive or descriptive. It may advise a course of action by drawing attention to the moral or ethical benefits that accrue when that suggestion is taken, or allude to the negative consequences inevitable if a line of behavior is ignored. Rather than explicitly prescribing behaviour, the proverb may be merely descriptive, highlighting a common irony or tendency in life.

Generally, the element of education in the use of proverbs can be subsumed under the rhetorical function: the use of proverbs to persuade. The proverb user seeks to alter or reinforce the listener’s conviction or attitude by referring him to lessons from parallel events in the proverb world. By getting the addressee to agree with the moral precept in the proverb used, the speaker thereby hopes to win him over to his viewpoint. Thus in a situation where someone has a big difficulties, it is said *air diminum rasa duri, nasi dimakan rasa sekam* ‘drinking water like drinking thorn, eating rice like eating husk’ to emphasis that everything become tasteless when someone in a serious trouble. The proverbs is used to describe the situation. The message is keep away from trouble otherwise you will get into do not do something wrong otherwise

As stated earlier that proverbs are highly culture-bound expressions. From the viewpoint of linguistics, proverbs are a central type of phrasemes. Despite the extensive literature on proverbs, a generally acknowledged definition has not yet been arrived at. Many proverbs are figurative and have far-reaching cultural significance. They reveal all of the types of cultural phenomena. They are not only aspects of material culture, but many proverbs are also directly interrelated with other culturally relevant texts. However, proverbs are most significantly connected with aspects of culture-based social interaction. Proverbs are general statements that are believed to express a universal truth, i.e. they refer to allegedly shared knowledge about rules governing social behaviour. Of course they do not allow the drawing of conclusions about attitudes and values of an entire language community but only of special groups at a special times. Besides, proverbs can have the illocutionary force of ‘recommendation/recommending’. They can provide moral support for an argument or action by referring to a generalised proposition and thus give advice on how to behave in certain situations. Proverbs quote socially approved ideas that can be used instead of an argumentation; they can reveal traces of social concepts (of a special groups and/or former times) and hand them down to future generations. An example of such model is the proverb type.

3. Proverbs in bilingual dictionary

It is admitted by Atkins and Rundell (2008: 168) that “other phrasal idioms (proverb included) are the most difficult MWEs to handle in lexicography. In the absence of hard and fast criteria, it is well nigh impossible to be wholly consistent”. In many cases, such obstacles are presented by alternative components, parallel idioms, loose canonical forms, syntactic restrictions on one hand, and morpho-syntactic flexibility on the other.

In A Comprehensive Indonesian-English Dictionary (henceforth CIED), the author, Stevens and Schmidgall-Tellings, put many proverbs in the entry. They use several methods to explain the meaning of the proverb that can be subsumed into three categories: 1) to choose the right or one-to-one equivalents in target language, 2) to translate the proverbs literally (word by word) and provide the explanation afterward, and 3) to explain the main point of the proverbs.

The research begins with collecting all the sample proverbs from CIED and list them parallel to their equivalent or explanation. From 52 proverbs there are only 16 having one-to-one equivalents. The other have translation and explanation (6) and only explanation (30). It can be seen that the authors prefer to explain the meaning of the proverbs to provide the equivalents. Besides, translating the whole sentence word by word is apparently not a preferred strategy. It seem that the authors want to convey the main points of the proverbs.
Bilingual dictionary is all about seeking the right equivalents and that is the biggest problem if we talk about proverb that is very highly culture-bounded. Here, the lexicographical aspects of proverbs are interesting to discuss. Proverbs are based on a variety of cultural experiences that require exercise of individual discretion in various moments. Even in the same culture, like Malay and Indonesian, it is uncommon to find proverbs stating apparently opposing principles. Above all, if we talk about two different culture, Indonesia (Asia) and English (Europe).

As stated above, the authors of CIED use different methods dealing with different types of proverbs. Providing explanation is the most commonly used because it is hard to find the right equivalent.

**One-to-one equivalents**

Below a sample of five proverbs that have one-to-one equivalents in SL dan TL and the discussion as follows.

- **Air kali tidak selamanya banjir** ≈ life’s not all beer and skittles

  The Indonesian proverb literally means ‘river is not always full of water’ which implies that life is not always full of happiness and prosperity. Water is a symbol of prosperity, especially when it is in dry season and there is water shortage. Meanwhile, to say the same meaning, English proverb use beer and skittles. The words are shorthand for a life of indulgence spent in the pub. Skittles, also known as ninepins, which was the precursor to ten-pin bowling has been a popular English pub game since the 17th century. The pins are set up in a square pattern and players attempt to knock them down with a ball. It is still played but not so much as previously. ([www.phrases.org.uk/index.htm](http://www.phrases.org.uk/index.htm)). The proverb is actually an advice or indirect invitation not to be on spree because life is not always going well.

- **Mengungkit batu di bencah** ≈ to carry out a Sisyphean task

  The Indonesian proverb literally means ‘to lift stones to the muddy ground’ which implies something that is extremely difficult to do. To move your body in the muddy ground is so difficult to do, let alone to lift stones. To say something very difficult or even impossible to do the authors use word Sisyphean which comes from Sisyphus in nominal form. Sisyphean (of a task) impossible to complete From the Greek myth in which Sisyphus was punished for the bad things he had done in his life with the never-ending task of rolling a large stone to the top of a hill, from which it always rolled down again. There is always an explanation behind the proverb that can reveal way of life or cultural life of a society.

- **Menepuk air di dulang, tepercik muka sendiri** ≈ it is an ill bird that fouls its own nest

  The Indonesian proverb literally means ‘if you clap water in the tray the water will splash onto your face’. This proverb describes someone who did a bad thing, then the consequences will be of him/her. The English proverb use a bird as the doer and its own nest as beneficiary. In other words, only a foolish or dishonorable person would bring dishonor to his/her self or his/her surroundings ([www.oxfordreference.com](http://www.oxfordreference.com)). This is a condemnation of a person who vilifies his own family, country, etc.

- **Sambil menyelam minum air** ≈ to kill two birds with one stone

  The proverb has a meaning to accomplish two aims by a single action ([www.thefreedictionary.com](http://www.thefreedictionary.com)). This expression is quite widely used in daily communication,
even today. Young Indonesian speakers are quite familiar with this one. The literally translation of Indonesian proverb is ‘while diving drinking water’ which describe things that can be done at the same time. Meanwhile, in English equivalent, the same notion is expressed by killing two birds with one stone. It would be remarkable indeed if someone sling a stone at a bird got one bird, let alone two.

- **Membuang bunga ke jirat** ≈ to cast pearls before swine

The Indonesian proverb is literally translated into English as ‘to throw away flowers to tomb’. It means to give something of value of someone who won’t appreciate it; to waste something good on someone who doesn’t care about it. Flowers are something of great value as pearls in equivalent English proverb. Things which describe someone who doesn’t know or care about the valuable things are tomb and swine.

The other methods employed in CIED in handling proverbs are by giving explanation. This is especially done if SL lexemes lack exact equivalents in the target language or lack of one-to-one equivalents by way of explanatory metalanguage. The SL proverbs of this type is as follows.

- **air dicindang tiada (tidak akan pernah) putus** ≈ said of relatives who make up again after a short estrangement
- **bagai air di daun keladi/talas** ≈ inconstant, unstable, fluctuating, variable
- **bermain air basah, bermain api letur(p)** ≈ every effort has its consequences
- **batu hitam tak bersanding** ≈ to be deceptive in appearance

In this method of explaining, the lexicographer is just to inform TL pragmatics without giving alternative expression. That indicates lack of lexicographer knowledge about TL or it’s simply because there’s no appropriate equivalent.

On the other hand, there are also several proverbs that is merely translated literally into TL without giving any information. It can be seen in the proverb **air tenang menghanyutkan** ≈ still waters run deep.

**Arcaic forms**

Because Indonesian proverbs live in illiterate society, no wonder that many arcaic form found in the proverbs. Words like tuba ‘poison to catch fish’, beriak ‘to wave’, bersibak ‘to wave’, pelimbahan, dicindang, berkumbah ‘to shower’, membandarkan ‘to lift’, dulang ‘food tray’, sekam ‘husk’, sauk ‘ladle’, lubuk ‘deep pool in a stream’, bendah, dan jirat ‘tomb’. The words are hardly used in daily communication and considered as arcaic and considered as foreign words.

It is quite strange to find more proverbs in this dictionary than the biggest monolingual Indonesian dictionary, KBBI. If we look closer to the preface of this dictionary, it was stated that the data was collected from various sources, including old scripts that using the proverbs. Among others are 1885 Dictionary of Islam by T.P. Hughes, 1950 Nuttige Planten van Indonesie, the classic text Javaans-Nederlands Handwoordenboek and 1959 Malay-English Dictionary by Wilkinson.

4. Conclusion and suggestion

Bilingual lexicography is all about finding right equivalents. This simple statement need further strategies and methods when dealing with various proverb situation. Like Atkins and Rundell already said it is “the most difficult MWEs to handle in lexicography so it seems impossible to be wholly consistent. It would be very nice if the lexicographer has a wide knowledge about the culture, not only the language, of SL and TL. It will facilitate his/her
work to find the right equivalent of the proverb in TL. But the problem is not always about
the lexicographer’s capacity, but also the cultural aspect itself which is different in every
society.
It is suggested to use illustrative methods to add information about one language. One-to-one
equivalent is not enough. To give additional information, lexicographer should use
explanatory metalanguage too.

References
# Appendix

Sample proverbs in CIED

<table>
<thead>
<tr>
<th>No.</th>
<th>Proverb in Malay</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Air kali tidak selamanya banjir</td>
<td>Life is not always beer and skittles</td>
</tr>
<tr>
<td>2.</td>
<td>Air laut siapa yang garamin</td>
<td>Self-praises is no recommendation</td>
</tr>
<tr>
<td>3.</td>
<td>Air susu dibalas dengan air tuba</td>
<td>To return good for evil</td>
</tr>
<tr>
<td>4.</td>
<td>Air beriak tanda tak dalam</td>
<td>The empty vessel makes the greatest sound</td>
</tr>
<tr>
<td>5.</td>
<td>Air besar batu bersibak</td>
<td>Relationships fall apart after conflict</td>
</tr>
<tr>
<td>6.</td>
<td>Air cucuran atap jatuhnya ke pelimbahan juga</td>
<td>Like father, like son</td>
</tr>
<tr>
<td>7.</td>
<td>Air dicindang tiada (tidak akan pernah) putus</td>
<td>Said of relatives who make up again after a short estrangement</td>
</tr>
<tr>
<td>8.</td>
<td>Air dalam berenang/berkumbah, air dengkat bercebek</td>
<td>To live within one’s means</td>
</tr>
<tr>
<td>9.</td>
<td>Air tenang menghanyutkan</td>
<td>Still waters run deep</td>
</tr>
<tr>
<td>10.</td>
<td>Bagai air di daun keladi/talas</td>
<td>Inconstant, unstable, fluctuating, variable</td>
</tr>
<tr>
<td>11.</td>
<td>Bagai air titik ke batu</td>
<td>a) Constant dripping will wear away a stone; b) unteachable</td>
</tr>
<tr>
<td>12.</td>
<td>Bagai membandarkan air ke bukit</td>
<td>To perform Sisyphean labor</td>
</tr>
<tr>
<td>13.</td>
<td>Bermain air basah, bermain api letur(p)</td>
<td>Every effort has its consequences</td>
</tr>
<tr>
<td>14.</td>
<td>Menepuk air di dulang, tepercik muka sendiri</td>
<td>It is an ill bird that fouls its own nest</td>
</tr>
<tr>
<td>15.</td>
<td>Air diminum rasa duri, nasi dimakan rasa sekam</td>
<td>Said of s.o. having difficulties so that everything become tasteless</td>
</tr>
<tr>
<td>16.</td>
<td>Tak air hujan ditampung, tak air peluh diurut/tak air palang dipancung</td>
<td>The end justifies the means</td>
</tr>
<tr>
<td>17.</td>
<td>Pandai berminyak air</td>
<td>a) Toady; b) to flatter, cajole, wheedle</td>
</tr>
<tr>
<td>18.</td>
<td>Bagai mencencang air</td>
<td>To carry coals to Newcastle</td>
</tr>
<tr>
<td>19.</td>
<td>Tambah air tambah sagu</td>
<td>Increased work means increased income</td>
</tr>
<tr>
<td>20.</td>
<td>Sekali air besar, sekali tepian beranjak/beralih</td>
<td>(so) many hands/men, (so) many minds</td>
</tr>
<tr>
<td>21.</td>
<td>Ada air ada ikan</td>
<td>One can find the means of subsistence in every country</td>
</tr>
<tr>
<td>22.</td>
<td>Bagai air jatuh ke pasir</td>
<td>It is like pouring water into a sieve</td>
</tr>
<tr>
<td>23.</td>
<td>Sambil menyelam minum air</td>
<td>To kill two birds with one stone</td>
</tr>
<tr>
<td>24.</td>
<td>Sauk air mandikan diri</td>
<td>To stand on one’s own two feet</td>
</tr>
<tr>
<td>25.</td>
<td>Seperti air basuh</td>
<td>Obtainable in abundance</td>
</tr>
<tr>
<td>26.</td>
<td>Seperti air di atas bulu bebek</td>
<td>Like water off a duck’s back</td>
</tr>
<tr>
<td>27.</td>
<td>Batu bulat tak bersanding</td>
<td>A brave man fears nobody</td>
</tr>
<tr>
<td>28.</td>
<td>Bak batu masuk lubang</td>
<td>To disappear</td>
</tr>
<tr>
<td>29.</td>
<td>Hujan emas di negeri orang, hujan batu di negeri sendiri</td>
<td>East, west, home’s best; be it so humble there’s no place like home</td>
</tr>
<tr>
<td>30.</td>
<td>Batu hitam tak bersanding</td>
<td>To be deceptive in appearance</td>
</tr>
<tr>
<td>31.</td>
<td>Seperti batu jatuh ke lubuk</td>
<td>Vanished into thin air; spirited away; refers to s.o. who has left his kampung and never comes back again</td>
</tr>
<tr>
<td>32.</td>
<td>Lempar(kan) batu, sembunyi(kan) tangan</td>
<td>Not willing to admit that one has committed an act (or, a crime)</td>
</tr>
<tr>
<td>33.</td>
<td>Mencampakkan batu ke luar</td>
<td>All lay goods on a willing horse</td>
</tr>
<tr>
<td>34.</td>
<td>Mengungkit batu di bencah</td>
<td>To carry out a Sisyphean task</td>
</tr>
<tr>
<td>35.</td>
<td>Patah batu hatinya</td>
<td>His desire to finish the job etc. That he has undertaken has completely disappeared</td>
</tr>
<tr>
<td>36.</td>
<td>Menunggu angin lalu</td>
<td>To wait in vain</td>
</tr>
<tr>
<td>37.</td>
<td>Angin berputar, ombak bersabung</td>
<td>Most difficult (of case)</td>
</tr>
<tr>
<td>38.</td>
<td>Tahu di angin berkisar</td>
<td>To know the capricious humor of people</td>
</tr>
<tr>
<td>39.</td>
<td>Tahu di angin turun naik</td>
<td>To see it coming</td>
</tr>
<tr>
<td>40.</td>
<td>Ke mana angin yang deras, ke situ condongnya</td>
<td>He will turn with every wind that blows</td>
</tr>
<tr>
<td>No.</td>
<td>Proverb in Indonesian</td>
<td>English Translation</td>
</tr>
<tr>
<td>-----</td>
<td>-----------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>41.</td>
<td>Berjalan peliharaan kaki, berkata peliharaan lidah</td>
<td>Look before you leap</td>
</tr>
<tr>
<td>42.</td>
<td>Kaki naik kepala turun</td>
<td>To bustle about</td>
</tr>
<tr>
<td>43.</td>
<td>Kaki untuk dipakaikan gelang</td>
<td>It’s like a blacksmith with a white silk apron i.e. inappropriate</td>
</tr>
<tr>
<td>44.</td>
<td>Bunga bukan sekaki</td>
<td>“there’s more than one flower”</td>
</tr>
<tr>
<td>45.</td>
<td>Tangan kanan jangan percaya akan tangan kiri</td>
<td>Don’t trust anybody</td>
</tr>
<tr>
<td>46.</td>
<td>Tangan mencencang bahu memikul</td>
<td>You have to be responsible for your actions</td>
</tr>
<tr>
<td>47.</td>
<td>Lempar bunga dibalas lempar tahi</td>
<td>Repay good with evil; return evil for good</td>
</tr>
<tr>
<td>48.</td>
<td>Membuang bunga ke jirat</td>
<td>To cast pearls before swine</td>
</tr>
<tr>
<td>49.</td>
<td>Jauh bau bunga, dekat bau tahi</td>
<td>Familiarity breeds contempt</td>
</tr>
<tr>
<td>50.</td>
<td>Bunga dipetik perdu ditendang</td>
<td>Aiming at making profits, being out for profit making, wanting to live in clover</td>
</tr>
<tr>
<td>51.</td>
<td>Bunganya dipersunting, pangkal bunga diberaki</td>
<td>Aiming at making profits, being out for profit making, wanting to live in clover</td>
</tr>
<tr>
<td>52.</td>
<td>Bunga gugur petik pun gugur</td>
<td>Nobody dies before his time</td>
</tr>
</tbody>
</table>
Cuyonon and Tausug Cultural Dictionaries: Databank for determining mutual intelligibility of two indigenous Philippine languages

Rosario B. Dizon (rbd_fil@yahoo.com)
Mary Ann S. Sandoval (maryzan@yahoo.com)
Myrna Halim (darling 1223 @ yahoo.com)
Lydia L. Mata (mata_lydia@yahoo.com)
Ser Robert David B. Dizon (srdbd@yahoo.co.uk)

1MSU-Iligan Institute Technology, the Philippines
2Siliman University, the Philippines

Abstract
This paper examines and compares two southern Philippine languages namely Cuyonon and Tausug through their cultural dictionaries which were the output of a previous research project entitled “Positive Expressions of Southern Philippines: Towards Compiling Cultural Dictionaries. The 200 Swadesh word-list formulated by Morris Zorc (1955) was used as comparable instrument to determine the mutual intelligibility between these languages. The technique used in detecting genetic relationship was done by way of lexico-statistics, a glottochronology approach that can determine even the historical period when the languages under study were able to contact with each other. Results show that there are only thirty (30) or 15 % probable cognates and one hundred seventy (170) or 85% non probable cognates. This means that Tausug and Cuyonon languages are not mutually intelligible. Statistical data revealed that the probable cognates are attributed on basic human activity, parts of the body, nature and atmospheric elements. Numerals or count nouns belong to non probable cognates. Findings show that it was not during trading empire when Jolo was the “center of trade” in the 18th century that the possible link or contact between the two languages occurred. It may have happened as early as the 10th century A.D. when Tausugs from Borneo crossed the land bridges via the islands connecting Palawan that contact with the Cuyonons happened.

Keywords: mutual intelligibility, probable cognates, lexico-statistics, glottochronology

1. Introduction

Language is a vital part of man’s existence for it is his way of communicating or expressing his thoughts. It is a basic ingredient of his life for by nature he is gregarious and he makes use of language to interact with people. However, the interaction can only be made possible if the communicators are using languages that are mutually intelligible.

This endeavor takes into consideration two language varieties: Tausug and Cuyonon. Tausug is the dominant language of Jolo in Sulu archipelago, an island belonging to Southern Philippines which is nearer to Sabah and Malaysia than to Manila. On the other hand, Cuyunon is the language spoken in the Cuyo island and most parts of Puerto Prinsesa, capital of the province of Palawan that stretches from Mindoro in the northeast to Borneo in the southwest. It lies between the South China Sea and the Sulu Sea.

This is a descriptive research involving two languages examined from 4 speech communities - 2 from Jolo, Sulu and 2 from Puerto Prinsesa, Palawan. The comparable data consisted of 200 Swadesh word-list from Zorc (1977) was used as instrument to determine the mutual intelligibility between Tausug and Cuyunon.

Using glottochronology, the study sought to detect genetic relationship between two languages through computational linguistics in an approach called lexico-statistics. This involves quantitative comparison of common lexical cognates. With this technique, one can determine the historical period when the languages under study were able to contact with each
other and separated based on the analysis of their mutual intelligibility and statistical comparison of shared word listing.

Other support techniques used were inspection method (Gudschinsky, 1956) and rapid appraisal. Both are sociolinguistic research approaches (Bergman 1991, Stalder 1996) which utilized focus group and individual interview questionnaire. These approaches provided a general sociolinguistic profile or situation of speech varieties being studied. Informant responses revealed patterns of contact and degree of lexical relationship. Researchers relied on speakers’ utterances in determining the high lexical similarity and intelligibility of lexical cases. An understanding of language vitality and viability are important to determine the environment of the speakers, thereby researchers can predict development or extinction of relationship between speech groups. Survival of the language has a direct bearing on the active or passive language maintenance. These valuable pieces of information are needed in confirming lexical relationships.

1.1. Objectives of the study

This study sought to determine the degree of mutual intelligibility of two indigenous languages namely; Tausug and Cuyunon through lexico-statistics thereby reconstructing to some extent the historical links between the said languages. Specifically, it desired to:

1. determine the lexical similarities/dissimilarities between Tausug and Cuyunon;
2. identify the percentage of cognates and non-cognates; and
3. trace back historical, socio-economic and political factors that influence similarities/dissimilarities between Tausug and Cuyunon.

1.2. Significance of the study

- Results of this study shed light on the lexical relationship between Tausug and Cuyunon, the active languages spoken in the southern part of the Philippines.
- The ethno-historical findings regarding these two indigenous languages are beneficial to linguists considering that there is no existing glottochronological study yet similar to the present thesis.
- The codified dictionaries of these two languages with English and Filipino equivalent can aid the non-speakers of Tausug and Cuyunon as well as the local and foreign tourists who would want to visit the country.
- Finally, the collected vocabulary of both Tausug and Cuyunon languages can be used as inputs for the modules and instructional materials’ development for K+12 curriculum, literacy programs both for the out-of-school youth and adults. Results and findings of this research can be used as data for language studies of linguists and cultural anthropologists.

1.3. Definition of Terms

**Cognate.** Word in one language which is similar in form and meaning to a word in another language because both languages are related (Richards: 1992). It is also refers to a word that has been inherited by two or more languages from parent language. In this study, it is the name for similarity in sound and spelling between Tausug and Cuyunon.

**Dialect.** The word refers to a variety of a language, spoken in one part of a country (regional dialect), or by people belonging to a particular class (Social dialect or sociolect), which is different in some word grammar, and/or pronunciation from other forms of the same language. A dialect gains status and becomes the standard variety of a language (Richards: 1992).

**Intelligibility.** Degree to which a message can be understood. Studies of speech perception point out that the intelligibility of speech is due to various factors including accent and intonation, the listener’s ability to predict parts of the message, the location of pauses in the utterance, the grammatical complexity of sentences, and the speed at which utterances are
produced \textit{(Richards:1992)}. In this paper, it is the existing common understanding between spoken and written Tausug and Cuyunon languages.

2. Review of Related Studies and Literature

In their research about the Papuan languages of Alor-Pantar, Holton, Klamer and Kratochvil (2005) assessed 17 Eastern Indonesian languages and had proven lexicostatistics a very useful tool for genetic classification. The first attempt to examine internal sub-grouping of these languages was made by Stokhof (1975) based on lexicostatistical analysis of 117 item Swadesh lists. Stokhof identified a number of lexical features. The data further helped him resolve migration route question.

Ayotte, Michael and Charlene (2002) in their language surveys of Atong, Ambele and Menka, Congo examined lexical similarities that existed between these languages and Ngwo. They also examined the level of contact among the speakers, and the extent of understanding that would benefit Ngwo language development. Findings showed that the word list from the three languages shared many similarities. Probable cognates came from pidgin or trade terms, church and intermarriages, as well as education terms. At home, with age-mates, in the field and with traditional ceremonies, wider variety of mother tongue was spoken.

Other studies directly related to the present paper are that of Donohue, M. (2007) that traced the phonological history of the non- Austronesian languages of Southern Indonesia. In the said research, investigators also discovered the migration route of trade and commerce and how pidgin terms were creolized by the early peoples of this place. Pawley, A. (2001) investigated the Proto TransNew Guinea obstruents: arguments from top-down reconstruction. This is a study in Melanesian linguistics as compiled by the Pacific Linguistics Research School of Asian Studies of Australian National University. The researchers realized that in some areas, language families have not been classified properly and that there exists no classification of the world languages done by a single, consistent method as the lexicostatistics.

Gudschinsky, Sarah (2009) stated basic assumption in lexicostatistical investigation which can easily be noted. She said that some basic core vocabulary of any language, on empirical evidence, is much less to change than any parts. This concept is similar to Sapir’s idea of a basic nucleus of morphological structure. Terms for material culture, on the other hand, are frequently borrowed along with other cultural items. This is the contrast between core vocabulary and general vocabulary. There is also a presumption that the rate of loss of basic vocabulary is approximately the same in all languages.

Baguio, Flores and Dizon (2010) study entitled Trilingual Dictionary revealed that the language of Higaonons or Binukid / Higaonons have undergone changes from classic to active. There is confusion in spelling and terms. Translation made by foreign missionaries and the influx of people from coastal areas influenced the language change of the natives. Such translation resulted in phonological changes. However, in the case of the Tausug language, changes are not noticed in the language the past many years. Perhaps, it is because Tausug aside from being the lingua franca of the archipelago is also a dominant language.

In this study, the lexical similarities between the Tausug and Cuyunon were compared and analyzed on the basis of their probable cognates and probable non–cognate categories. In this aspect, mutual intelligibility, Brown (1998) emphasized that “intelligibility is mutual and inherent”.
2.1 Theoretical Foundation

Sapir, Edward (1921) founded the theory on linguistic relativism wherein he believed that the methods of comparative linguistics are equally valid when applied to indigenous languages. He said “We see and hear and otherwise experience very largely as we do because the language habits of our community predispose certain choices of interpretation,” (Sapir, p. 69). This was supported later by Benjamin Lee-Whorf when he said that “we dissect nature along lines laid down by our native languages.” Furthermore, by simple inspection and comparison, closely related native languages can be discovered. Sapir (1921) suggested that marked similarities in the basic morphological structure of languages may indicate common origin. The use of this principle increases the number of languages that can be postulated as belonging to a given linguistic grouping, but it cannot tell however, when the relationship began to diverge from one another.

This study was anchored on the principle of linguistic relativism believing that every indigenous language has its own speech community that can develop its varieties through contact and time and later on be found probable cognates with other indigenous tongues like Cuyunon to Tausug and vice-versa.

Schematic Diagram

![Diagram](image1)

Figure 1 Diagram

3. Methodology

This study used the combination of qualitative-quantitative methods of research. It also made use of rapid appraisal method in gathering data (Stalder, 1996) utilizing group and individual interview to provide the idea of the general sociolinguistic situation of the speech varieties being studied. Informant responses revealed the pattern of contact and degree of multilingualism. Researchers relied on the speaker’s perception in conjunction with high lexical similarities to determine levels of mutual intelligibilities in both languages. An understanding of language vitality was important in predicting active language maintenance. Maintenance of the language has a direct bearing in the reconstruction of ethno-historical
building-up theory. Synchronic and diachronic analysis were also used in examining the historical past of contact between the two languages.

One essential step in making lexicostatistical comparison of two or more languages was the collection of comparable word list that would provide a clear sure way of estimating linguistic relationships even in the basis of relatively small data. A convenient list for this purpose is the Swadesh 200-word list. Using this tool formulated and tested by Zorc in 1954 has several advantages: It is made up of chosen core vocabulary and has been tested for percentage of retention in languages with written historical records; the list has been used in a number of comparison so that results of comparison between Tausug and Cuyunon may be compared with studies already made in other languages.

Determining the probable cognates and non-probable cognates using the 200-Swadesh word list was the next step done in order to ascertain the number of pairs of words which were mutually intelligible or not.

Flow Chart of Activities. The following steps were taken in this study:

![Flow Chart of Activities](image)

3.1. Key Informants
The informants for the Tausug language are Hadji Anang Muabbad and Hadji Sakkim Muabbad. They hailed from Bus-bus, Jolo. They were also the translators for the Swadesh 200-word basic vocabulary and originally categorized 245 SIL vocabulary. The translations were properly validated by Ahmad Sapii and Abdurajik Sapii. The same informants also did the evaluation and validation of the recorded texts.

The informants for the Cuyonon language are Mrs. Maryjane Parcon, a high school teacher and Vicenta Ensigne, an administrative employee from Palawan State University. Mr. Romy Daizon, his wife Mrs. Ellen Medes-Daizon also from Palawan State University were the validators. The coordinator was Prof. Teresita Tajolosa, PhD Linguistics whom we owe deep gratitude for organizing a team to take care of putting everything in order including the courier service.
The criteria of choosing the informants were their exposure to other languages age, status in the community, place of residence and the nature of their work. The informants were all native speakers of their respective languages. The researchers investigated the informants - their background, past experience, standing in their respective communities, cultural background, interactions with the older members of the community, role of their work, their relationships with other business associates, and active participation with the members of their religious congregation.

3.2. Research Locale
The study covered two locales: Jolo, Sulu for the Tausug Language and Puerto Prinsesa, Palawan for Cuyunon language.

Jolo is the main center of Tausug language. It is mainly inhabited by the Tausugs who are considered to be the original settlers of Jolo. The name “Tausug” is a combination of two native terms, “tau” which means people and “sug” which means current. “Sug” is also used in southern Sulu to mean the island of Sulu. The term ‘Tausug’ also refers to the language spoken by the natives in the place. Hassan, 1975 (Gallman, 1977) states that in Sulu, “Tausug” is the mother tongue of a quarter of more than a million and is the trade language, as well as the lingua franca spoken throughout the archipelago.

There are 8 barangays in mainland Jolo. They are Alat, Asturias, Busbus, Chinese Pier, San Raymundo, Takut-takut, Tulay, and Walled City.

Palawan, on the other hand, is a group of islands located in the southwest portion of the Philippine archipelago. It is bounded on the north by Mindoro, the China Sea on the west, the Sulu Sea on the east, and Borneo on the south with land area of 14,896.3 square kilometres and a length of 425b km from tip to tip. (Ocampo, 1985).

There are 28 principal rivers of which the Babuyan River in central Palawan is the longest measuring around 54 kilometers. These mountains and rivers have been one of the major sources of subsistence for the inhabitants of the island for thousands of years. Different fresh water fishes and shells abound on the rivers and provide the protein, calcium and natural iodine needs of the people.

The climate of Palawan is affected by the annual cycles of northeasterly wind and southwest monsoon. During the months of January to April, the northeasterly wind brings cooler temperature and less rain. The rainy season occurs in the months of May to December. The west coast experiences more rain than the eastern portion which is shielded by the mountains during the southwest monsoon.

The first Spanish account of Palawan appeared in 1521 when the surviving ships of Magellan stopped for provisions at an island called Pulaoan.

In the middle of 16th century, due to political superiority, the Muslims of Borneo and Mindanao exercised authority over the inhabitants of Palawan. Thus, the people gave tribute to the Muslims of Borneo, a custom that persisted until 1588.

The province is a melting pot of 81 different ethno-linguistic groups; Mcfarland (1980) listed six Palawan languages on the basis of genetic classification: Kalamianon, Agutaynon, Batak, Tagbanwa, Palawan and Molbog. The first two form a northern grouping, while the other four are parts of the southern grouping. There are 52 dialects in the province, with Tagalog
being spken by 28 percent of the people. Other major dialects are Cuyunin (26.27 %), Pinalwan (11.08 %), and Hiligaynon (9.6%) (www.palawan.gov.ph 2005).

Figure 3 Map of Jolo, Sulu

3.3. Instrumentation
Translation to Tausug and Cuyonon of the 200 Swadesh lexical items was the initial task done by the researchers. To have common understanding of meaning, they used Filipino language in their translations.

The next step was the translation analysis for parrellism of equivalents from Tausug and Cuyunon languages. Pairs of lexical items were identified whether probable cognates or non-probable cognates. The determination of the cognates was done using Gudchinsky’s (1956) “inspection method”.

To determine the percentage of probable cognates, add all the pairs of words in Tausug and Cuyonon that are perfectly similar and multiply the total by 100; and then divide the product by 200. The result is the percentage of the probable cognates. Meanwhile, get the frequency of non-probable cognates by counting the dissimilar words from pairs of Tausug and Cuyonon and then, multiply the product again by 200.

Thus,

\[
\text{Probable cognate} = \text{no. of similar words} \times \frac{100}{200}
\]

\[
\text{Probable cognate} = \text{no. of dissimilar words} \times \frac{100}{200}
\]
Or,

\[
\text{lexical similarities} = \frac{\text{no. of probable cognates}}{\text{total no. of pairs of words}}
\]

\[
\text{lexical differences} = \frac{\text{no. of probable non cognates}}{\text{total no. of pair of words}}
\]

Simply, the following is the formula in identifying the cognates:

\[
\text{Mutual Intelligibility (MI) Percentage} = \frac{\text{specific frequency}}{\text{total frequency}} \times 100
\]

4. Presentation and Analysis of Data

To show comparative illustration of two languages being studied, the researchers used the 200-Swadesh list of Zorc (1954) as presented in Table 1.

Table 1. 200 Core Vocabulary of the Swadesh List for Tausug and Cuyunon Languages

<table>
<thead>
<tr>
<th>English</th>
<th>Tausug</th>
<th>Cuyunon</th>
<th>Filipino</th>
</tr>
</thead>
<tbody>
<tr>
<td>above (adv.)</td>
<td>hataas’ npc</td>
<td>kabui</td>
<td>sa itaas</td>
</tr>
<tr>
<td>alive (adj.)</td>
<td>buhi’ npc</td>
<td>ubong</td>
<td>buhay</td>
</tr>
<tr>
<td>all (adj.)</td>
<td>katan’ npc</td>
<td>tanan</td>
<td>lahat</td>
</tr>
<tr>
<td>and (conj.)</td>
<td>iban’ npc</td>
<td></td>
<td>at</td>
</tr>
<tr>
<td>at (prep.)</td>
<td>in’ npc</td>
<td>sa</td>
<td>sa</td>
</tr>
<tr>
<td>back (n)</td>
<td>liko’d pc</td>
<td>likod</td>
<td>likod</td>
</tr>
<tr>
<td>bad (adj.)</td>
<td>mangih’ npc</td>
<td>marimo/ beken y mayad</td>
<td>masama</td>
</tr>
<tr>
<td>because(conj.)</td>
<td>amu man yan npc</td>
<td>samoro</td>
<td>dahil</td>
</tr>
<tr>
<td>belly (human body)</td>
<td>tiyan’ npc</td>
<td>tian</td>
<td>tiyan</td>
</tr>
<tr>
<td>below (adv.)</td>
<td>hababa’ npc</td>
<td>ibaba</td>
<td>sa ibaba</td>
</tr>
<tr>
<td>big(adj.)</td>
<td>malag’go npc</td>
<td>mabael</td>
<td>malaki</td>
</tr>
<tr>
<td>bird (animal)</td>
<td>manuk-manuk npc</td>
<td>lamgam</td>
<td>ibon</td>
</tr>
<tr>
<td>bite(verb)</td>
<td>kutkut’ npc</td>
<td>kagat</td>
<td>kagat</td>
</tr>
<tr>
<td>black (adj.)</td>
<td>hom’ npc</td>
<td>mangitit</td>
<td>itim</td>
</tr>
<tr>
<td>blood (human body)</td>
<td>dugu’ pc</td>
<td>dug’o</td>
<td>dugo</td>
</tr>
<tr>
<td>blow (verb)</td>
<td>huyop’ pc</td>
<td>huyóp</td>
<td>ihipan</td>
</tr>
<tr>
<td>bone (human body)</td>
<td>bukog’ npc</td>
<td>tulan</td>
<td>buto</td>
</tr>
<tr>
<td>branch (noun)</td>
<td>sanga pc</td>
<td>sanga</td>
<td>sanga</td>
</tr>
<tr>
<td>breast(human body)</td>
<td>daghal’ npc</td>
<td>dib dib</td>
<td></td>
</tr>
<tr>
<td>breathe (breathe)</td>
<td>napas, hagok npc</td>
<td>agginawa</td>
<td>huminga</td>
</tr>
<tr>
<td>burn (verb)</td>
<td>sunog’ pc</td>
<td>sunóg</td>
<td>paso</td>
</tr>
<tr>
<td>buy (buy)</td>
<td>bi’ni npc</td>
<td>agbacal/bacal</td>
<td>bumili</td>
</tr>
<tr>
<td>chew (verb)</td>
<td>siopa/supao’ npc</td>
<td></td>
<td>guya</td>
</tr>
<tr>
<td>child (noun)</td>
<td>bata’ pc</td>
<td>bata</td>
<td>bata</td>
</tr>
<tr>
<td>choose (verb)</td>
<td>pinih’ npc</td>
<td>pilien/pili</td>
<td>pilin</td>
</tr>
<tr>
<td>climb (verb)</td>
<td>dimag’ npc</td>
<td>saka</td>
<td>umakyat</td>
</tr>
<tr>
<td>cloud(noun)</td>
<td>panganod pc</td>
<td>panganod</td>
<td>ulap</td>
</tr>
<tr>
<td>cold(noun)</td>
<td>mahagpay’/ mahag’got npc</td>
<td>maramig/ ramig</td>
<td>malamig</td>
</tr>
<tr>
<td>come(verb)</td>
<td>mari’ npc</td>
<td>pakon</td>
<td>halika</td>
</tr>
<tr>
<td>cook (verb)</td>
<td>lutoon npc</td>
<td>loto/paglutó</td>
<td>magluto</td>
</tr>
<tr>
<td>count (verb)</td>
<td>mag-itong’ npc</td>
<td>bilang/numiro</td>
<td>bilang</td>
</tr>
<tr>
<td>cry(verb)</td>
<td>tumangis, npc</td>
<td>tangis</td>
<td>umiyak</td>
</tr>
<tr>
<td>cut/chop (verb)</td>
<td>iyutod npc</td>
<td>girit</td>
<td>putulin</td>
</tr>
<tr>
<td>day (noun)</td>
<td>adlaw’ pc</td>
<td>adlaw</td>
<td>araw</td>
</tr>
<tr>
<td>English Word</td>
<td>Tagalog Word</td>
<td>English Meaning</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>--------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>die (verb)</td>
<td>patay’ npc</td>
<td>patay</td>
<td></td>
</tr>
<tr>
<td>dig (verb)</td>
<td>kalot’ npc</td>
<td>makali</td>
<td></td>
</tr>
<tr>
<td>dirty (adjective)</td>
<td>malum’-mih npc</td>
<td>maburing</td>
<td></td>
</tr>
<tr>
<td>dog (animal)</td>
<td>iruh’ npc</td>
<td>t’o aso</td>
<td></td>
</tr>
<tr>
<td>dream (verb)</td>
<td>tagainop’ npc</td>
<td>taginep</td>
<td></td>
</tr>
<tr>
<td>drink (verb)</td>
<td>ino’m pc</td>
<td>inom/inem inom</td>
<td></td>
</tr>
<tr>
<td>dry (verb)</td>
<td>tahay’ npc</td>
<td>iga tuyo</td>
<td></td>
</tr>
<tr>
<td>dry up (verb)</td>
<td>piatahay’ npc</td>
<td>patuyuin</td>
<td></td>
</tr>
<tr>
<td>dull (adjective)</td>
<td>pol’ npc</td>
<td>maburing marumi</td>
<td></td>
</tr>
<tr>
<td>earth (noun)</td>
<td>lupa ‘ npc</td>
<td>lugtâ / kalibotan</td>
<td></td>
</tr>
<tr>
<td>eat (verb)</td>
<td>kaon’ npc</td>
<td>kaen/kain kumain</td>
<td></td>
</tr>
<tr>
<td>egg (noun)</td>
<td>itlog’ pc</td>
<td>itlog</td>
<td></td>
</tr>
<tr>
<td>eye (human body)</td>
<td>mata’ pc</td>
<td>mata</td>
<td></td>
</tr>
<tr>
<td>far (adjective)</td>
<td>malayo’ npc</td>
<td>maraye cao malayo</td>
<td></td>
</tr>
<tr>
<td>fat (adjective)</td>
<td>tambok’ npc</td>
<td>matambek mataba</td>
<td></td>
</tr>
<tr>
<td>father (noun)</td>
<td>amah’ npc</td>
<td>tatay ama</td>
<td></td>
</tr>
<tr>
<td>fear (verb)</td>
<td>mabugah’ npc</td>
<td>adlik takot</td>
<td></td>
</tr>
<tr>
<td>feather (noun)</td>
<td>bahibu’ balahibo npc</td>
<td>balahibo</td>
<td></td>
</tr>
<tr>
<td>fire (noun)</td>
<td>ka’yo npc</td>
<td>calayô apoy</td>
<td></td>
</tr>
<tr>
<td>fish (noun)</td>
<td>isiah’ npc</td>
<td>canen-canen isda</td>
<td></td>
</tr>
<tr>
<td>flow (verb)</td>
<td>labi’ npc</td>
<td>ilig v. i. agos</td>
<td></td>
</tr>
<tr>
<td>flower (noun)</td>
<td>sumping’ npc</td>
<td>borak bulaklak</td>
<td></td>
</tr>
<tr>
<td>fly (verb)</td>
<td>lupad’ npc</td>
<td>lepad lipad</td>
<td></td>
</tr>
<tr>
<td>foot/leg (human body)</td>
<td>siki/bihitiis pc</td>
<td>sikê/batis paa</td>
<td></td>
</tr>
<tr>
<td>forest (noun)</td>
<td>kakahuyan’ npc</td>
<td>cagueban gubat</td>
<td></td>
</tr>
<tr>
<td>four (count noun)</td>
<td>upat’ npc</td>
<td>apat apat</td>
<td></td>
</tr>
<tr>
<td>fruit (noun)</td>
<td>bunga’ npc</td>
<td>protas prutas</td>
<td></td>
</tr>
<tr>
<td>good (adjective)</td>
<td>marayaw’ npc</td>
<td>mayad mabuti</td>
<td></td>
</tr>
<tr>
<td>grass (noun)</td>
<td>sagbut’ npc</td>
<td>elamonon Damo</td>
<td></td>
</tr>
<tr>
<td>grow (verb)</td>
<td>tubo’ npc</td>
<td>kolay Tubo</td>
<td></td>
</tr>
<tr>
<td>guts (noun)</td>
<td>isog’ npc</td>
<td>lakas-loob</td>
<td></td>
</tr>
<tr>
<td>hand (human body)</td>
<td>lima’ npc</td>
<td>alimâ Kamay</td>
<td></td>
</tr>
<tr>
<td>he/she (pron)</td>
<td>siya’ npc</td>
<td>lalaki lalake/babae</td>
<td></td>
</tr>
<tr>
<td>head (human body)</td>
<td>o’u’u npc</td>
<td>olo/barot Ulo</td>
<td></td>
</tr>
<tr>
<td>head/hair (human body)</td>
<td>buhok’ npc</td>
<td>bok Buhok</td>
<td></td>
</tr>
<tr>
<td>hear (verb)</td>
<td>diongog’ npc</td>
<td>pamati Dinggin</td>
<td></td>
</tr>
<tr>
<td>heavy (adjective)</td>
<td>bug-at’ npc</td>
<td>mabegat Mahigat</td>
<td></td>
</tr>
<tr>
<td>hide (verb)</td>
<td>tiapok’/tiao’ npc</td>
<td>palipid Itago</td>
<td></td>
</tr>
<tr>
<td>hit (verb)</td>
<td>tiodjo/kiogdan npc</td>
<td>Pagsuntok</td>
<td></td>
</tr>
<tr>
<td>hold (verb)</td>
<td>kapotan’/kapot npc</td>
<td>awid Hinawakan</td>
<td></td>
</tr>
<tr>
<td>hot (adjective)</td>
<td>mapaso’ npc</td>
<td>mainit Maanghang</td>
<td></td>
</tr>
<tr>
<td>house (noun)</td>
<td>bay’ npc</td>
<td>balay Bahay</td>
<td></td>
</tr>
<tr>
<td>how (adv.)</td>
<td>unohon npc</td>
<td>marasano Paano</td>
<td></td>
</tr>
<tr>
<td>hunt (verb)</td>
<td>piangita’ npc</td>
<td>pangayâm Pangaso</td>
<td></td>
</tr>
<tr>
<td>husband (noun)</td>
<td>bana’ npc</td>
<td>asawang lalaki Bana</td>
<td></td>
</tr>
<tr>
<td>I (proper noun)</td>
<td>ako’ npc</td>
<td>aco Ako</td>
<td></td>
</tr>
<tr>
<td>if (conjunction)</td>
<td>bang’ npc</td>
<td>Kung</td>
<td></td>
</tr>
<tr>
<td>inside (noun)</td>
<td>halaum’/sud npc</td>
<td>sa siled Saloob</td>
<td></td>
</tr>
<tr>
<td>kill (verb)</td>
<td>piatay’ npc</td>
<td>patay Patayin</td>
<td></td>
</tr>
<tr>
<td>knowledge (noun)</td>
<td>kaingatan’ npc</td>
<td>kinaelaman Karunungan</td>
<td></td>
</tr>
<tr>
<td>lake (body of water)</td>
<td>lanaw’ npc</td>
<td>leek Lawa</td>
<td></td>
</tr>
<tr>
<td>laugh (verb)</td>
<td>katawa’ npc</td>
<td>magkadlaw Tawa</td>
<td></td>
</tr>
<tr>
<td>leaf (noun)</td>
<td>dahon’ npc</td>
<td>daon Dahon</td>
<td></td>
</tr>
<tr>
<td>left (adjective)</td>
<td>lawa’ npc</td>
<td>walâ</td>
<td>Kaliwa</td>
</tr>
<tr>
<td>lie down (verb)</td>
<td>kumulang’ npc</td>
<td>kelat/leni</td>
<td>humiga</td>
</tr>
<tr>
<td>lightning (noun)</td>
<td>kilat’ pc</td>
<td>kelat/lenti</td>
<td>kidlat</td>
</tr>
<tr>
<td>liver (human body)</td>
<td>atay npc</td>
<td>atay</td>
<td>atay</td>
</tr>
<tr>
<td>long (adjectives)</td>
<td>mataas’ npc</td>
<td>malabeg</td>
<td>mahaba</td>
</tr>
<tr>
<td>louse (insect)</td>
<td>lisa’ npc</td>
<td>kutô/kaymed</td>
<td>lisa</td>
</tr>
<tr>
<td>man (noun)</td>
<td>isogo’ npc</td>
<td>lalaki</td>
<td>tao</td>
</tr>
<tr>
<td>meat/flesh (noun)</td>
<td>unod’ npc</td>
<td>karne</td>
<td>karne</td>
</tr>
<tr>
<td>moon (heavenly body)</td>
<td>bulan’ pc</td>
<td>bulan</td>
<td>buwan</td>
</tr>
<tr>
<td>mosquito (insect)</td>
<td>hilam’ npc</td>
<td>namok</td>
<td>lamok</td>
</tr>
<tr>
<td>mother (noun)</td>
<td>inah’ npc</td>
<td>nanay</td>
<td>ina</td>
</tr>
<tr>
<td>mouth (human body)</td>
<td>simud’ npc</td>
<td>babâ/sungad</td>
<td>bibig</td>
</tr>
<tr>
<td>name (noun)</td>
<td>ngan’ npc</td>
<td>aran</td>
<td>pangalan</td>
</tr>
<tr>
<td>narrow (adjective)</td>
<td>masigpit’ npc</td>
<td>makitid</td>
<td>makitid</td>
</tr>
<tr>
<td>near (adjective)</td>
<td>masuok’/daig npc</td>
<td>marapit</td>
<td>malapit</td>
</tr>
<tr>
<td>neck (human body)</td>
<td>liog’ npc</td>
<td>lieg</td>
<td>leeg</td>
</tr>
<tr>
<td>needle (non)</td>
<td>jaum’ npc</td>
<td>tingway</td>
<td>karayom</td>
</tr>
<tr>
<td>new (adjective)</td>
<td>ha’go pc</td>
<td>bago</td>
<td>bago</td>
</tr>
<tr>
<td>night (noun)</td>
<td>dom’ npc</td>
<td>gabi</td>
<td>gabi</td>
</tr>
<tr>
<td>no/not (adverb)</td>
<td>bukun’di’ npc</td>
<td>hindi</td>
<td>hindi</td>
</tr>
<tr>
<td>nose (human body)</td>
<td>ho’ng npc</td>
<td>irong</td>
<td>ilong</td>
</tr>
<tr>
<td>old (adjective)</td>
<td>ma’as npc</td>
<td>malam/gurang</td>
<td>matanda</td>
</tr>
<tr>
<td>one (count term)</td>
<td>isa’ npc</td>
<td>isarâ</td>
<td>isa</td>
</tr>
<tr>
<td>open (verb)</td>
<td>ukab’ npc</td>
<td>abri</td>
<td>bukas</td>
</tr>
<tr>
<td>other (adjective)</td>
<td>kaibanan’ npc</td>
<td>iba</td>
<td>iba</td>
</tr>
<tr>
<td>person (noun)</td>
<td>ho’ng npc</td>
<td>tao</td>
<td>tao</td>
</tr>
<tr>
<td>plant (plant)</td>
<td>tanom’ npc</td>
<td>tanim</td>
<td>tanim/halaman</td>
</tr>
<tr>
<td>pound/beat (verb)</td>
<td>dok-dok’ npc</td>
<td>mabáyô</td>
<td>dikdik</td>
</tr>
<tr>
<td>rain (noun)</td>
<td>ulan’ npc</td>
<td>oran</td>
<td>ulan</td>
</tr>
<tr>
<td>rat (animal)</td>
<td>ambaw’ npc</td>
<td>ambe</td>
<td>daga</td>
</tr>
<tr>
<td>red (adjective)</td>
<td>pula’ pc</td>
<td>pulà</td>
<td>pula</td>
</tr>
<tr>
<td>right (noun)</td>
<td>to-o’ npc</td>
<td>kanan</td>
<td>kanan</td>
</tr>
<tr>
<td>road (noun)</td>
<td>da’-an npc</td>
<td>daan</td>
<td>daan</td>
</tr>
<tr>
<td>roof (house part)</td>
<td>atup’ npc</td>
<td>bubong</td>
<td>bubong</td>
</tr>
<tr>
<td>root (noun)</td>
<td>gamot’ npc</td>
<td>ugam</td>
<td>ugam</td>
</tr>
<tr>
<td>rope (noun)</td>
<td>lubid’ pc</td>
<td>lubid</td>
<td>lubid</td>
</tr>
<tr>
<td>rotten (adjective)</td>
<td>haloh’ npc</td>
<td>pans</td>
<td>pans</td>
</tr>
<tr>
<td>salt (noun)</td>
<td>asin’ npc</td>
<td>asin</td>
<td>asin</td>
</tr>
<tr>
<td>sand (noun)</td>
<td>buhangin’ npc</td>
<td>baras</td>
<td>buhangin</td>
</tr>
<tr>
<td>say (verb)</td>
<td>bayahah’/laong npc</td>
<td>makôn</td>
<td>sabi</td>
</tr>
<tr>
<td>scratch (verb)</td>
<td>kiato’/kia’nias npc</td>
<td>kalot</td>
<td>kamot</td>
</tr>
<tr>
<td>sea (body water)</td>
<td>dagat’ pc</td>
<td>dagat</td>
<td>dagat</td>
</tr>
<tr>
<td>see (verb)</td>
<td>ki’ta npc</td>
<td>kitô</td>
<td>tingnan</td>
</tr>
<tr>
<td>sew (verb)</td>
<td>tahi’ npc</td>
<td>bidbid</td>
<td>tahi</td>
</tr>
<tr>
<td>sharp (adjective)</td>
<td>maha-it’ npc</td>
<td>matarem/tarem</td>
<td>matalas</td>
</tr>
<tr>
<td>shoot (verb)</td>
<td>timbakon npc</td>
<td>barilin</td>
<td>barilin</td>
</tr>
<tr>
<td>short (adjective)</td>
<td>mahawpoh’ npc</td>
<td>maikli</td>
<td>maikli</td>
</tr>
<tr>
<td>shoulder (human body)</td>
<td>abaga’ pc</td>
<td>abagâ</td>
<td>balikat</td>
</tr>
<tr>
<td>shy/ashamed (adjective)</td>
<td>sipug npc</td>
<td>mahiyain</td>
<td>mahiyain</td>
</tr>
<tr>
<td>sick/painful (adjective)</td>
<td>saki’t npc</td>
<td>maysakit</td>
<td>maysakit</td>
</tr>
<tr>
<td>sit (verb)</td>
<td>lingkod’ npc</td>
<td>magurâng</td>
<td>upo</td>
</tr>
<tr>
<td>skin (human body)</td>
<td>pais’ npc</td>
<td>kulit</td>
<td>balat</td>
</tr>
<tr>
<td>sky (heavenly body)</td>
<td>langit’ npc</td>
<td>maeneyaen</td>
<td>himpapawid</td>
</tr>
<tr>
<td>sleep (verb)</td>
<td>to’-og npc</td>
<td>maturug</td>
<td>maturug</td>
</tr>
<tr>
<td>small (adjective)</td>
<td>anak-anak’ npc</td>
<td>maitê</td>
<td>maitê</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>Tagalog</td>
<td>English</td>
<td>Tagalog</td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>smoke (verb)</td>
<td>asu’ pc</td>
<td>asô</td>
<td>usok</td>
</tr>
<tr>
<td>snake (animal)</td>
<td>has’ npc</td>
<td>ahas</td>
<td>ahas</td>
</tr>
<tr>
<td>sniff (verb)</td>
<td>simhot npc</td>
<td>palangamô</td>
<td>Singhot</td>
</tr>
<tr>
<td>spider (insect)</td>
<td>lawa npc</td>
<td>bukaw-bukaw</td>
<td>Gagamba</td>
</tr>
<tr>
<td>spit (verb)</td>
<td>lumurah’ npc</td>
<td>durâ</td>
<td>Luwa</td>
</tr>
<tr>
<td>split (verb)</td>
<td>sipakon’ npc</td>
<td>Biyakin</td>
<td>Biyakin</td>
</tr>
<tr>
<td>squeeze (verb)</td>
<td>piogah’ npc</td>
<td>palangamô</td>
<td>S</td>
</tr>
<tr>
<td>stab (verb)</td>
<td>diogsoan’/bio’noh npc</td>
<td>bunô</td>
<td>Saksak</td>
</tr>
<tr>
<td>stand (verb)</td>
<td>tindog’ npc</td>
<td>tindeg</td>
<td>Tayo</td>
</tr>
<tr>
<td>star (heavenly body)</td>
<td>bito-on’ npc</td>
<td>biton</td>
<td>Bituin</td>
</tr>
<tr>
<td>steal (verb)</td>
<td>tiakaw’ npc</td>
<td>panákaw</td>
<td>Magnakaw</td>
</tr>
<tr>
<td>stone (noun)</td>
<td>bato’ pc</td>
<td>batô</td>
<td>Bato</td>
</tr>
<tr>
<td>suck (verb)</td>
<td>sup’ npc</td>
<td>sipsip</td>
<td>Sopsop</td>
</tr>
<tr>
<td>swim (verb)</td>
<td>langoy’ pc</td>
<td>langoy/eseb</td>
<td>Lumangoy</td>
</tr>
<tr>
<td>tail (noun)</td>
<td>ikog’ npc</td>
<td>Buntot</td>
<td>Buntot</td>
</tr>
<tr>
<td>that (2nd p) (pronoun)</td>
<td>iyan’ npc</td>
<td>daya</td>
<td>Iyan</td>
</tr>
<tr>
<td>that (3rd p) (pronoun)</td>
<td>gadto’ npc</td>
<td>dato</td>
<td>Lyon</td>
</tr>
<tr>
<td>thick (objects)</td>
<td>marakmol’ npc</td>
<td>madamel</td>
<td>Makapal</td>
</tr>
<tr>
<td>thin (objects)</td>
<td>manipis’ pc</td>
<td>manipis</td>
<td>Manipis</td>
</tr>
<tr>
<td>think (verb)</td>
<td>huna-hunaon, pikilan npc</td>
<td>Isipin</td>
<td>Isipin</td>
</tr>
<tr>
<td>this (pronoun)</td>
<td>ini’ npc</td>
<td>dia/dague</td>
<td>Ito</td>
</tr>
<tr>
<td>thou (pronoun)</td>
<td>ikaw’ npc</td>
<td>darwa</td>
<td>Ikaw</td>
</tr>
<tr>
<td>three (noun)</td>
<td>to-o npc</td>
<td>tatlô/tres</td>
<td>Tatlo</td>
</tr>
<tr>
<td>throw (verb)</td>
<td>ti’luh/liabay npc</td>
<td>pilak</td>
<td>Ihagis</td>
</tr>
<tr>
<td>thunder (noun)</td>
<td>daog-daog’ npc</td>
<td>daledeg</td>
<td>Kulog</td>
</tr>
<tr>
<td>tie (verb)</td>
<td>hukut’ npc</td>
<td>igot</td>
<td>Tali</td>
</tr>
<tr>
<td>tongue (human boy)</td>
<td>dila pc</td>
<td>dila</td>
<td>Dila</td>
</tr>
<tr>
<td>tooth (human body)</td>
<td>ipun npc’</td>
<td>ipen</td>
<td>Ngipin</td>
</tr>
<tr>
<td>true/correct (adjective)</td>
<td>amo’ npc</td>
<td>matod</td>
<td>totoo/tumpak</td>
</tr>
<tr>
<td>turn (verb)</td>
<td>lingi’ npc</td>
<td>biring</td>
<td>Lumington</td>
</tr>
<tr>
<td>two (noun)</td>
<td>dua’ npc</td>
<td>darwâ/dos</td>
<td>Dalawa</td>
</tr>
<tr>
<td>vomit (verb)</td>
<td>suka’ pc</td>
<td>sukâ</td>
<td>Sumuka</td>
</tr>
<tr>
<td>walk (verb)</td>
<td>manaw’ npc</td>
<td>panaw</td>
<td>Lumakad</td>
</tr>
<tr>
<td>water (body water)</td>
<td>tubig’ pc</td>
<td>tobig</td>
<td>Tugit</td>
</tr>
<tr>
<td>we (excl.) (pronoun)</td>
<td>kamo’/kato npc</td>
<td>Tayo</td>
<td>Tayo</td>
</tr>
<tr>
<td>we (incl.) (pronoun)</td>
<td>kami’ pc</td>
<td>kami</td>
<td>Tayo</td>
</tr>
<tr>
<td>wet (noun)</td>
<td>basah’ npc</td>
<td>mabasa/botod adj.</td>
<td>Basa</td>
</tr>
<tr>
<td>what (pronoun)</td>
<td>ono’ npc</td>
<td>anono</td>
<td>Ano</td>
</tr>
<tr>
<td>when (pronoun)</td>
<td>koh’-no npc</td>
<td>inoro</td>
<td>Kalian</td>
</tr>
<tr>
<td>where (pronoun)</td>
<td>hanono/haritiin npc</td>
<td>sadin</td>
<td>Saan</td>
</tr>
<tr>
<td>white (adjective)</td>
<td>puti’ pc</td>
<td>puti</td>
<td>Puti</td>
</tr>
<tr>
<td>who (pronoun)</td>
<td>hisiyo’ npc</td>
<td>sino</td>
<td>Sino</td>
</tr>
<tr>
<td>wide (adjective)</td>
<td>maluag’ npc</td>
<td>mawayang</td>
<td>Malapad</td>
</tr>
<tr>
<td>wife (noun)</td>
<td>asawa’ npc</td>
<td>Maybahay</td>
<td>Maybahay</td>
</tr>
<tr>
<td>wind (noun)</td>
<td>hangin’ npc</td>
<td>angen/eyep-eyep</td>
<td>Hangin</td>
</tr>
<tr>
<td>wing (noun)</td>
<td>pikpik’ npc</td>
<td>Pakpak</td>
<td>Pakpak</td>
</tr>
<tr>
<td>woman (noun)</td>
<td>babae’ npc</td>
<td>babai</td>
<td>Binibini</td>
</tr>
<tr>
<td>wood (noun)</td>
<td>kahoy’ npc</td>
<td>caoy</td>
<td>Kahoy</td>
</tr>
<tr>
<td>work (verb)</td>
<td>hinang’ npc</td>
<td>buwát/ binuatan</td>
<td>Trabaho</td>
</tr>
<tr>
<td>yawn (verb)</td>
<td>manghuy’-ab npc</td>
<td></td>
<td>Hikab</td>
</tr>
<tr>
<td>year (noun)</td>
<td>tahon’ npc</td>
<td>dagun</td>
<td>Taon</td>
</tr>
</tbody>
</table>
### Table 2 Probable Cognates

<table>
<thead>
<tr>
<th>Back (noun)</th>
<th>Taykod/likod</th>
<th>Likod</th>
<th>Likod</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood (human body)</td>
<td>Dug'</td>
<td>Dugo</td>
<td></td>
</tr>
<tr>
<td>Blow (verb)</td>
<td>Liopad/huyop</td>
<td>Huyóp</td>
<td>Ihipan</td>
</tr>
<tr>
<td>Branch (noun)</td>
<td>Sanga</td>
<td>Sanga</td>
<td></td>
</tr>
<tr>
<td>Burn (verb)</td>
<td>Sunog'</td>
<td>Sunóg</td>
<td>Paso</td>
</tr>
<tr>
<td>Child (noun)</td>
<td>Bata</td>
<td>Bata</td>
<td></td>
</tr>
<tr>
<td>Cloud (noun)</td>
<td>Panganod/gabun</td>
<td>Panganod</td>
<td>Ulap</td>
</tr>
<tr>
<td>Day (noun)</td>
<td>Adlaw'</td>
<td>Adlaw</td>
<td></td>
</tr>
<tr>
<td>Drink (verb)</td>
<td>Inom/inem</td>
<td>Inom</td>
<td></td>
</tr>
<tr>
<td>Egg (noun)</td>
<td>Itlog'</td>
<td>Itlog</td>
<td></td>
</tr>
<tr>
<td>Eye (human body)</td>
<td>Mata'</td>
<td>Mata/tao-tao</td>
<td></td>
</tr>
<tr>
<td>Foot/leg (human body)</td>
<td>Siki/bihtiis</td>
<td>Sike/batis</td>
<td>Paa</td>
</tr>
<tr>
<td>Lightning (noun)</td>
<td>Kilat'</td>
<td>Kelat/lenti</td>
<td></td>
</tr>
<tr>
<td>Moon (heavenly body)</td>
<td>Bulan'</td>
<td>Bulan</td>
<td></td>
</tr>
<tr>
<td>New (adjective)</td>
<td>Ba'go</td>
<td>Bago</td>
<td></td>
</tr>
<tr>
<td>Red (adjective)</td>
<td>Pula'</td>
<td>Pulá</td>
<td></td>
</tr>
<tr>
<td>Rope (noun)</td>
<td>Lubid'</td>
<td>Lubid</td>
<td></td>
</tr>
<tr>
<td>Sea (body water)</td>
<td>Dagat'</td>
<td>Dagat/laud</td>
<td></td>
</tr>
<tr>
<td>Shoulder (human body)</td>
<td>Abaga'</td>
<td>Abagá</td>
<td></td>
</tr>
<tr>
<td>Smoke (verb)</td>
<td>Asu'</td>
<td>Asó</td>
<td>Usok</td>
</tr>
<tr>
<td>stone (noun)</td>
<td>bato’ pc</td>
<td>bató</td>
<td>Bato</td>
</tr>
<tr>
<td>swim (verb)</td>
<td>langoy’ pc</td>
<td>langoy/eseb</td>
<td>lumangoy</td>
</tr>
<tr>
<td>thin (objects)</td>
<td>manipis’ pc</td>
<td>manipis</td>
<td>manipis</td>
</tr>
<tr>
<td>tongue (human boy)</td>
<td>dila pc</td>
<td>dila</td>
<td>Dila</td>
</tr>
<tr>
<td>vomit (verb)</td>
<td>suka’ pc</td>
<td>sukà</td>
<td>sumuka</td>
</tr>
<tr>
<td>water (body_water)</td>
<td>tubig’ pc</td>
<td>tubig</td>
<td>Tubig</td>
</tr>
<tr>
<td>we (inc. ) (pronoun)</td>
<td>kami’ pc</td>
<td>kami</td>
<td>Tayo</td>
</tr>
<tr>
<td>white (adjective)</td>
<td>puti’ pc</td>
<td>putú</td>
<td>Puti</td>
</tr>
<tr>
<td>woman (noun)</td>
<td>babae’ pc</td>
<td>babai</td>
<td>binibini</td>
</tr>
<tr>
<td>you (pronoun)</td>
<td>ikaw pc</td>
<td>ikaw</td>
<td>Ikaw</td>
</tr>
<tr>
<td>TOTAL =30</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Table 3 Similar in First Syllable |
| bad (adjective) | mangih’ npc | marimô/beken y | masama |
| belly (human body) | tiyan’ npc | tian | Tiyan |
| big (adjective) | malag’go npc | mabael | malaki |
| choose (verb) | pinnh’ npc | pilien/pili | Pilin |
| cold (noun) | mahagpay’/ mahag’got npc | maramig/ramig | malamig |
| dirty (adjective) | malum’-mih npc | maburing | marumi |
| dream (verb) | tagainop’ npc | taginep | panaginip |
| earth (noun) | lupa’ npc | lugtâ/kalibotan | Lupa |
| eat (verb) | kaon’ npc | kaen/kain | kumain |
| far (adjective) | malayo’ npc | maraye cao | malayo |
| good (adjective) | marayaw’ npc | mayad n. | mabuti |
| head (human body) | o-u npc | olo/barot | Ulo |
| hot (adjective) | mapaso’ npc | mainit | maanghang |
| house (noun) | bay’ npc | balay | bahay |
| I (proper noun) | ako’ npc | aco | Ako |
| leaf (noun) | dahon’ pc | daon | dahon |
| long (adjectives) | mataas’ npc | malabeg | mahaba |
| near (adjective) | masuok’/daig npc | marapit | malapit |
| neck (human body) | liog’ pc | lieg | Leeg |
| old (adjective) | ma’as’ pc | malam/gurang | matanda |
| one (count term) | isa’ npc | isarâ | Isa |
| plant (plant) | tanom’ pc | tanim | tanim/halam an |
| rat (animal) | ambaw’ pc | ambe | Daga |
| see (verb) | ki’ tanpc | kitó | tingnan |
| sharp (adjective) | maha-it’ npc | materem/tarem | matalas |
| stand (verb) | tindog’ npc | tindeg | Tayo |
| star (heavenly body) | bito-on’ pc | biton | bituin |
| thick (objects) | marakmol’ npc | madamel | makapal |
| thunder (noun) | daog-daog’ npc | daledeg | kulog |
| tooth (human body) | ipun pc’ | ipen | ngipin |
| wide (adjective) | maluag’ pc | mawayang | malapad |
| TOTAL= 31 |

| Table 4 Similar in Second Syllable |
| below (adverb) | hababa’ npc | ibaba | sa ibaba |
| fat (adjective) | tambok’ npc | matambek | mataba |
| squeeze (verb) | piogah’ npc | puguin | Piga |
| that (2p) (pronoun) | iyan’ pc | daya | Iyan |
| TOTAL =4 |

109
### Table 5 Similar in Final Syllable

<table>
<thead>
<tr>
<th>All (adjective)</th>
<th>katan (noun)</th>
<th>tanan</th>
<th>lahat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cook (verb)</td>
<td>lutoon (noun)</td>
<td>loto/paglutô</td>
<td>magluto</td>
</tr>
<tr>
<td>Cry (verb)</td>
<td>tumanis, (noun)</td>
<td>tangis</td>
<td>umiyan</td>
</tr>
<tr>
<td>Fire (noun)</td>
<td>ka’yo (noun)</td>
<td>calayô</td>
<td>apoyp</td>
</tr>
<tr>
<td>Fly (verb)</td>
<td>lulap’ (noun)</td>
<td>lepad</td>
<td>lilipad</td>
</tr>
<tr>
<td>Four (count noun)</td>
<td>upat’ (noun)</td>
<td>apat</td>
<td>apat</td>
</tr>
<tr>
<td>Hand (human body)</td>
<td>lima’ (noun)</td>
<td>alimâ</td>
<td>kamay</td>
</tr>
<tr>
<td>Head/hair (human body)</td>
<td>buhok’ (noun)</td>
<td>bok</td>
<td>buhok</td>
</tr>
<tr>
<td>Heavy (adjective)</td>
<td>bug-at’ (noun)</td>
<td>mabegat</td>
<td>mabigat</td>
</tr>
<tr>
<td>Kill (verb)</td>
<td>piatay’ (noun)</td>
<td>patay</td>
<td>patayin</td>
</tr>
<tr>
<td>Name (noun)</td>
<td>ngan’ (noun)</td>
<td>aran</td>
<td>pangalan</td>
</tr>
<tr>
<td>Rain (noun)</td>
<td>ulan’ (noun)</td>
<td>oran</td>
<td>ulan</td>
</tr>
<tr>
<td>Sleep (verb)</td>
<td>to’-og (noun)</td>
<td>maturog</td>
<td>matulog</td>
</tr>
<tr>
<td>Steal (verb)</td>
<td>tiakaw’ (noun)</td>
<td>panakaw</td>
<td>magnakaw</td>
</tr>
<tr>
<td>That (3rd p) (pronoun)</td>
<td>yadto’ (noun)</td>
<td>dato</td>
<td>iyon</td>
</tr>
<tr>
<td>Walk (verb)</td>
<td>manaw’ (noun)</td>
<td>panaw</td>
<td>lumakad</td>
</tr>
<tr>
<td>Wet (noun)</td>
<td>basah’ (noun)</td>
<td>mabasa adj.</td>
<td>basa</td>
</tr>
<tr>
<td>What (pronoun)</td>
<td>ono’ (noun)</td>
<td>anono</td>
<td>ano</td>
</tr>
<tr>
<td>Wood (noun)</td>
<td>kahoy’ (noun)</td>
<td>caoy</td>
<td>kahoy</td>
</tr>
<tr>
<td>TOTAL = 19</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 6 Non Probable Cognates

<table>
<thead>
<tr>
<th>Above (adverb)</th>
<th>hataas’ (noun)</th>
<th>kabui</th>
<th>saitaas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alive (adjective)</td>
<td>buhi’ (noun)</td>
<td>ubong</td>
<td>buhay</td>
</tr>
<tr>
<td>At (preposition)</td>
<td>in’ (noun)</td>
<td>sa</td>
<td>sa</td>
</tr>
<tr>
<td>Because (conjunction)</td>
<td>amu man yan (noun)</td>
<td>samoro</td>
<td>dahil</td>
</tr>
<tr>
<td>Bird (noun)</td>
<td>manuk-manuk (noun)</td>
<td>lamgam</td>
<td>ibon</td>
</tr>
<tr>
<td>Bite (verb)</td>
<td>kikut’ (noun)</td>
<td>kagat</td>
<td>kagat</td>
</tr>
<tr>
<td>Black (adjective)</td>
<td>hom’ (noun)</td>
<td>mangitit</td>
<td>itim</td>
</tr>
<tr>
<td>Bone (noun)</td>
<td>bukog’ (noun)</td>
<td>tulan</td>
<td>buto</td>
</tr>
<tr>
<td>Breath (breathe)</td>
<td>napas, hagok (noun)</td>
<td>agginawa</td>
<td>huminga</td>
</tr>
<tr>
<td>Buy (buy)</td>
<td>bi’ni (noun)</td>
<td>agbacal/bacal</td>
<td>bumili</td>
</tr>
<tr>
<td>Climb (verb)</td>
<td>dimag’ (noun)</td>
<td>saka</td>
<td>umakyat</td>
</tr>
<tr>
<td>Come (verb)</td>
<td>mari’ (noun)</td>
<td>pakon</td>
<td>halika</td>
</tr>
<tr>
<td>Count (verb)</td>
<td>mag-itong’ (noun)</td>
<td>bilang/numiro</td>
<td>bilang</td>
</tr>
<tr>
<td>Cut/chop (verb)</td>
<td>lyyotod (noun)</td>
<td>girit</td>
<td>putulin</td>
</tr>
<tr>
<td>Dig (verb)</td>
<td>kalot’ (noun)</td>
<td>makali</td>
<td>hukayin</td>
</tr>
<tr>
<td>Dog (animal)</td>
<td>iruh’ (noun)</td>
<td>t’o</td>
<td>aso</td>
</tr>
<tr>
<td>Dry (verb)</td>
<td>tahay’ (noun)</td>
<td>tay</td>
<td>tuyo</td>
</tr>
<tr>
<td>Earthworm (insect)</td>
<td>labod’ (noun)</td>
<td>ulolago</td>
<td>bulati</td>
</tr>
<tr>
<td>Fall (verb)</td>
<td>nahulog’ hug (noun)</td>
<td>lagpak</td>
<td>madapa</td>
</tr>
<tr>
<td>Father (noun)</td>
<td>amah’ (noun)</td>
<td>tatay</td>
<td>ama</td>
</tr>
<tr>
<td>Fear (verb)</td>
<td>mabughah’ (noun)</td>
<td>adlik</td>
<td>takot</td>
</tr>
<tr>
<td>Fish (noun)</td>
<td>isiah’ (noun)</td>
<td>canen-canen</td>
<td>isda</td>
</tr>
<tr>
<td>Flow (verb)</td>
<td>labi’ (noun)</td>
<td>ilig</td>
<td>agos</td>
</tr>
<tr>
<td>Flower (noun)</td>
<td>sumping’ (noun)</td>
<td>borak</td>
<td>bulaklak</td>
</tr>
<tr>
<td>Forest (noun)</td>
<td>kakahuyan’ (noun)</td>
<td>cagueban</td>
<td>gubat</td>
</tr>
<tr>
<td>Fruit (noun)</td>
<td>bunga’ (noun)</td>
<td>protas</td>
<td>prutas</td>
</tr>
<tr>
<td>Grass (noun)</td>
<td>sagbut’ (noun)</td>
<td>elamonon</td>
<td>damo</td>
</tr>
<tr>
<td>Grow (verb)</td>
<td>tubo’ (noun)</td>
<td>kolay</td>
<td>tubo</td>
</tr>
<tr>
<td>He/she (pronoun)</td>
<td>siya’ (noun)</td>
<td>lalaki</td>
<td>lalake/habae</td>
</tr>
<tr>
<td>Hear (verb)</td>
<td>diongo’ (noun)</td>
<td>lamati</td>
<td>dinggin</td>
</tr>
<tr>
<td>Hide (verb)</td>
<td>tiapok’/tiao’ (noun)</td>
<td>palipid</td>
<td>itago</td>
</tr>
<tr>
<td>Hold (verb)</td>
<td>kapotan’/kapot (noun)</td>
<td>awid</td>
<td>hinawakan</td>
</tr>
<tr>
<td>How (adv.)</td>
<td>unohon’ (noun)</td>
<td>marasano</td>
<td>paano</td>
</tr>
<tr>
<td>Hunt (verb)</td>
<td>piangita’ (noun)</td>
<td>pangayam</td>
<td>pangaso</td>
</tr>
<tr>
<td>English Word</td>
<td>Cebuano (noun)</td>
<td>English Meaning</td>
<td>Cebuano Meaning</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------</td>
<td>-----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>husband</td>
<td>bana’ npc</td>
<td>asawang lalaki</td>
<td>Bana</td>
</tr>
<tr>
<td>inside</td>
<td>balaum’/sud npc</td>
<td>sa siled</td>
<td>sa loob</td>
</tr>
<tr>
<td>knowledge</td>
<td>kaintatan’ npe</td>
<td>kinaelaman</td>
<td>karunungan</td>
</tr>
<tr>
<td>lake</td>
<td>lanaw’ pc</td>
<td>leek</td>
<td>Lawa</td>
</tr>
<tr>
<td>laugh</td>
<td>katawa’ pc</td>
<td>magkadlaw</td>
<td>Tawa</td>
</tr>
<tr>
<td>left (adj)</td>
<td>laway’ pc</td>
<td>walà</td>
<td>kaliwa</td>
</tr>
<tr>
<td>louse (insect)</td>
<td>lisa’ npe</td>
<td>kutô/kaymed</td>
<td>Lisa</td>
</tr>
<tr>
<td>man</td>
<td>lilo’ npe</td>
<td>lalaki</td>
<td>Tao</td>
</tr>
<tr>
<td>meat/food</td>
<td>unod’ npc</td>
<td>karmi</td>
<td>Karne</td>
</tr>
<tr>
<td>mosquito</td>
<td>hilam’ npe</td>
<td>namok</td>
<td>lamok</td>
</tr>
<tr>
<td>mother</td>
<td>inah’ pc</td>
<td>nanay</td>
<td>Ina</td>
</tr>
<tr>
<td>mouth</td>
<td>simud’ npe</td>
<td>babà/sungad</td>
<td>Bibig</td>
</tr>
<tr>
<td>needle</td>
<td>jaum’ pc</td>
<td>tingway</td>
<td>karayom</td>
</tr>
<tr>
<td>night</td>
<td>dom’ npe</td>
<td>gabi</td>
<td>Gabi</td>
</tr>
<tr>
<td>nose</td>
<td>ho’ng pc</td>
<td>irong</td>
<td>Ilong</td>
</tr>
<tr>
<td>open (verb)</td>
<td>ho’ng pc</td>
<td>abri</td>
<td>Bukas</td>
</tr>
<tr>
<td>other (adj)</td>
<td>kaibanan’ npe</td>
<td>iba</td>
<td>Iba</td>
</tr>
<tr>
<td>person</td>
<td>ho’ng pc</td>
<td>tao</td>
<td>Tao</td>
</tr>
<tr>
<td>pound/beat</td>
<td>dok-dok’ npe</td>
<td>maháyò</td>
<td>dikdik</td>
</tr>
<tr>
<td>sand</td>
<td>buhangin ’ npe</td>
<td>baras</td>
<td>buhangin</td>
</tr>
<tr>
<td>say (verb)</td>
<td>baytah’/laong pc</td>
<td>makón</td>
<td>Sabi</td>
</tr>
<tr>
<td>scratch (verb)</td>
<td>kiatal’/ki’amas pc</td>
<td>kalot</td>
<td>kamot</td>
</tr>
<tr>
<td>sew (verb)</td>
<td>tali’ npe</td>
<td>bidid</td>
<td>Tahí</td>
</tr>
<tr>
<td>sit (verb)</td>
<td>ingcod’ npe</td>
<td>maguráng</td>
<td>Upo</td>
</tr>
<tr>
<td>skin (human body)</td>
<td>país’ npe</td>
<td>kulit</td>
<td>Balat</td>
</tr>
<tr>
<td>sky (heavenly body)</td>
<td>langit’ pc</td>
<td>maeneyaen</td>
<td>himpapPawid</td>
</tr>
<tr>
<td>small (adj)</td>
<td>anak-anak’ npe</td>
<td>maitê</td>
<td>malit</td>
</tr>
<tr>
<td>sniff (verb)</td>
<td>simhot npc</td>
<td>palangamö</td>
<td>singhot</td>
</tr>
<tr>
<td>spider</td>
<td>laway’ pc</td>
<td>bukaw-bukaw</td>
<td>Gagamba</td>
</tr>
<tr>
<td>spit (verb)</td>
<td>lumurah’ pc</td>
<td>durâ</td>
<td>Luwa</td>
</tr>
<tr>
<td>stab (verb)</td>
<td>diogsoan’/bio’nog pc</td>
<td>bunô</td>
<td>sakask</td>
</tr>
<tr>
<td>suck (verb)</td>
<td>sup-sup’ npe</td>
<td>sipsip</td>
<td>SopsoP</td>
</tr>
<tr>
<td>they (pronoun)</td>
<td>sila’ npe</td>
<td>sanda</td>
<td>Sila</td>
</tr>
<tr>
<td>this (pronoun)</td>
<td>ini’ npe</td>
<td>dia/dague</td>
<td>Ito</td>
</tr>
<tr>
<td>thou (pronoun)</td>
<td>ikaw’ npe</td>
<td>darwa</td>
<td>Ikaw</td>
</tr>
<tr>
<td>three (noun)</td>
<td>to-o pc</td>
<td>tatô</td>
<td>Tatlo</td>
</tr>
<tr>
<td>throw (verb)</td>
<td>ti’luh/liabay pc</td>
<td>pilak</td>
<td>ihagis</td>
</tr>
<tr>
<td>tie (verb)</td>
<td>hukut’ npe</td>
<td>igot</td>
<td>Tali</td>
</tr>
<tr>
<td>true/correct (adj)</td>
<td>amo’ npe</td>
<td>matod</td>
<td>totoo/tumpak</td>
</tr>
<tr>
<td>turn (verb)</td>
<td>lingi’ pc</td>
<td>biring</td>
<td>lumingon</td>
</tr>
<tr>
<td>two (noun)</td>
<td>dua’ pc</td>
<td>darwà/dos</td>
<td>Dalawa</td>
</tr>
<tr>
<td>when (pronoun)</td>
<td>koh’-no pc</td>
<td>inoro</td>
<td>kailan</td>
</tr>
<tr>
<td>where (pronoun)</td>
<td>hanono/hariinn pc</td>
<td>sadin</td>
<td>Saan</td>
</tr>
<tr>
<td>who (pronoun)</td>
<td>hisiyo’ pc</td>
<td>sino</td>
<td>Sino</td>
</tr>
<tr>
<td>wind (noun)</td>
<td>hangin’ npe</td>
<td>angen/eyep-eyep</td>
<td>Hangin</td>
</tr>
<tr>
<td>work (verb)</td>
<td>hinang’ npe</td>
<td>buwát/binuatan</td>
<td>trabaho</td>
</tr>
<tr>
<td>year (noun)</td>
<td>tahon’ npe</td>
<td>dagun</td>
<td>Taon</td>
</tr>
<tr>
<td>yellow (adj)</td>
<td>bianing’ npe</td>
<td>dulaw</td>
<td>Dilaw</td>
</tr>
</tbody>
</table>

TOTAL = 82

<table>
<thead>
<tr>
<th>English Word</th>
<th>Cebuano (noun)</th>
<th>English Meaning</th>
<th>Cebuano Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>and (conjunction)</td>
<td>iban’ npc</td>
<td>At</td>
<td></td>
</tr>
<tr>
<td>breast (human body)</td>
<td>daghal’ npe</td>
<td>dibdib</td>
<td></td>
</tr>
<tr>
<td>chew (verb)</td>
<td>sio’pa/supao’n pc</td>
<td>nguya</td>
<td></td>
</tr>
<tr>
<td>die (verb)</td>
<td>patay’ pc</td>
<td>Patay</td>
<td></td>
</tr>
<tr>
<td>dry up (verb)</td>
<td>piatahay’ npe</td>
<td>patuuyin</td>
<td></td>
</tr>
<tr>
<td>dull (adjective)</td>
<td>pol-pol’ npe</td>
<td>mapurol</td>
<td></td>
</tr>
</tbody>
</table>

Table 7 No Equivalent
It is noted that most of the words that are similar or the same are parts of the body, parts of the house, atmospheric elements, color, basic activities, adjectives and pronouns. Based on inspection method of Gudschinsky (1956), it is possible these words originated from the same roots and can therefore be called probable true cognates.

On the other hand, there are thirty one (31) pairs of words which have similarity only in the first syllable. They are usually adjectives, nouns, verbs, pronouns and objects. Words that have similar sound in the second syllable in Tausug and Cuyunon are four (4) in all. They consist of adverbs, insects and count terms. Lastly, there are nineteen (19) words of Tausug and Cuyunon languages which are only similar in the final syllable. They are found in human body parts, pronouns and verbs.

5. General discussion

Statistical data reveals probable true cognates are attributed heavily on basic words of simple life. It can be better describe as lexicons of the old world before Christianity. In narratology approach of scientific investigation, mention of such words point out to socio-economic condition of life before conquest or even before Chinese came to Jolo and made it the center of commerce. Noun refers to body parts, nature and heavenly bodies are all characteristics of the old world. In fact all the probable true cognates revolve around simple life style with the natural world. Possible link or contact between two cultures occurred can be estimated during the 10th century when the land bridges crossing Palawan to Borneo still existed.

In another investigation, contact between the two tribes could have happened before the spread of Islam or the historical time of Mindanao Islamization following the demise of the Prophet Muhammad - led to the expansion of Islam to Europe, Africa, and Southeast Asia.
6. Conclusion

Socio-economic, political and religious factors that influenced southern Philippine History can be concluded as the primary reasons for the non-intelligibility or big dissimilarity of the Cuyunon language from Tausug language. Contact between tribes could have occurred in the pre-Christianity times when the barter exchange do not use currency but basic necessities of life like food and clothing.

References


For ms of Address in French-Korean and Korean-French Dictionaries from the Perspective of a French Learner of Korean

NAM Kilim¹ JUNG Hae-Yun²
Kyungpook National University, South Korea
¹nki@knu.ac.kr, ²haeyun.jung.22@gmail.com

Abstract
This study examines how forms of address are dealt with in French-Korean and Korean-French dictionaries (respectively, Nouveau dictionnaire coréen-français (say-hanpwul-sacen), Prime dictionnaire français-coréen (pwul-han-sacen)). As bilingual dictionaries are undeniably a second language learning tool, their structure, form and content are altogether and equally important in the transmission of information and the contribution towards shaping the image the user has of the other language, and culture. Focusing on address terminology, I demonstrate that some words are heavily loaded with “shared cultural charge” (Galisson, 1988) just as terms of address are, and this kind of words should be brought into sharp focus when compiling bilingual dictionaries. In order to do so, I first give an overview of lexical entries of selected address terms as they are presented in the two aforementioned dictionaries, and which I have classified into three categories (pronouns, free forms of address and kinship terms). In the following section, I analyse these entries in depth focusing on the shortcomings they present – particularly from the perspective of a French learner of Korean as a foreign language – which are of three types: lexicographic inconsistencies, inaccuracy or imprecision of equivalents and lack of sociocultural information. In a the last section, I reconsider the issue in the light of pragmatics and “lexiculture” (Galisson, 1988) in order to show that forms of address, which are essential components of the cultural identity of the language they belong to, deserve a precise and thorough treatment in bilingual dictionaries on one hand, and on the other, to contribute to the revalorization of the status of bilingual dictionaries.

Keywords: forms of address, bilingual dictionary, lexiculture, French, Korean

1. Introduction

Language is not a mere tool of communication but also the vehicle of one’s cultural identity and social status. Thus, it assumes the role of establishing human relationships, whether horizontal (solidarity, equality) or vertical (power relation, hierarchy) (Brown&Gilman 1960; Kerbrat-Orecchioni 2010). Particularly, the social deictic forms of address epitomize this ambivalent function of language. In Korean, for instance, honorific titles (conching)² that also are a way of addressing people (hoching) constitute a highly complex and hierarchical system while in France titles used as address terms (Monsieur, Madame) have lost most of their honorific status. Besides titles, forms of address also include names, pronouns and kinship terms. Once again, Korean and French have a very distinctive approach, understanding, and usage of address terms. Bilingual dictionaries, which are the companions of language learners, should reflect the cultural gap between the two address systems without, nonetheless, fossilizing the differentiations, because words do not have a predetermined and immutable meaning but instead are unpredictable as being the “one-off realization of a given situation” (Schuwer 1994:1). How to render a correspondence between two different identities is the difficult task of the bilingual lexicographer and falls within the sphere of pragmatics, that is, the theories of the language in contexts that include, inter alia, the deictic forms of address.

Raising the issues of bilingual dictionary (mis)use and cultural gaps, this study aims to assess

² Korean has been transliterated using the Yale Romanization system, except for quotations where the system chosen by the author has been respected
the pragmatic treatment of terms of address in French↔Korean lexicography in order to propose a systematic lexicographical method of presenting pragmatically-cultural information for address terms. The first section surveys a number of forms of address from the point of view of a French learner of Korean and according to their typology as presented in Nouveau dictionnaire coréen-français (sayhanpwulsacen) and Dong-a’s Prime dictionnaire français-coréen (pwulhansacen)³. The main focus of the following section is to analyze the deficiencies and disambiguate the equivalents in French↔Korean lexicography which fails to represent consistently the cultural concepts behind each of these languages. In order to overcome such deficiencies, it will then be necessary to consider the issue, in the last section, within the wider context of cultural pragmatics and bilingual lexicography theories.

2. Typological presentation of forms of address in French↔Korean dictionaries

Forms of address can be classified into three types: personal pronouns, free forms of address (such as titles and names), and kinship terms. This section reviews how the corresponding entries are presented in Nouveau and Prime⁴.

First of all, it is noteworthy to remind that French uses two different pronouns according to context: to address someone: *tu* (you, second person singular subject) to address relatives and friends regardless of their age or social status, and *vous* (you, second person plural subject or polite form of address), the polite way of addressing anyone who is neither a relative nor (yet) a friend. The Korean equivalent of *tu* given by Prime is the personal pronoun *ne* and *vous*⁵ is rendered by *tangsin*. The latter case is a problematic translation which will be discussed in detail in the following section. In Nouveau, the entries for these two equivalents given by Prime present each specific features. First, the headword *ne* is accompanied by no usage information. As for *tangsin*, Nouveau translates it both by *tu* and *vous*.

Another way of addressing somebody is the free form of address, which can either be an honorific title or simply the addressee’s name. Needless to say, the latter can hardly be represented in dictionaries⁶. The French *Monsieur/Madame* used to be honorific titles to designate people of high social status. However, today it is mainly used as to address a stranger or an acquaintance in a polite way. It is worth noting that the treatment of *Monsieur* and *Madame* in Prime is somewhat asymmetrical. *Monsieur* is rendered by more genderless, generic Korean honorific items such as –ssi or –nim (when addressing someone in reference to their occupation), while *Madame* as term of address translates into either *pwuin* (wife) or the archaism *manim*. Only when occupying a high function are they assigned the suffix *-nim*. However, in Nouveau the address forms –ssi, and –nim refer to both males and females and *pwuin* is only referred to as a third person deixis. Interestingly, the headword *–nim* gives as a first example the lemma *sengsayngnim*, which is translated into its literal sense (teacher, professor) and its pragmatic meaning (Sir, Madam, Miss), which is in fact the most common way to address someone in Korean. Despite this, neither *Monsieur* nor *Madame* is rendered by *sengsayngnim*.

Besides honorifics that demonstrate the deference of the speaker toward the addressee,

---

³ Hereafter referred to as Nouveau and Prime respectively.
⁴ Any section in the dictionary entries that does not relate to forms of address or is redundant has been omitted from examples.
⁵ *Vous* as the polite form of address. When *vous* designates the plural form of address, it is rendered by *nehuittul*. This paper only considers the former case.
⁶ In Korean, calling someone by full name is not common. It usually is the case in formal school situation (e.g. when the teacher reads the roll call before the lesson) or when parents scold their children. Among very close friends, Koreans can use the first name to which they add either the suffix –i if it ends with a consonant or –a/ya. Generally speaking, (full) names are used for children or young adults (Lee & Ramsey 2000:233).
Korean also makes use of kinship terms as forms of address for friends or not socially superior adults, where French would simply use first names. In Korean, names are being reserved to address children or close friends of similar age. When addressing real relatives, both languages use the relevant kinship terms, except for relatives of the same generation (siblings and cousins) in French which makes exclusive use of first names. In *Nouveau*, *nwuna* (elder sister for males) put aside, it appears that both real and fictive kinship are represented under all headwords describing siblings, namely *enni* (elder sister for females), *hyeng* (elder brother for males), and *oppa* (elder brother for females). These two usages are dealt with in a particular way. Real kinship terms are rendered by a translation followed (or not) by pragmatic information, whereas fictive kinship terms, having no equivalent in French, are thus rendered by an explanation. Besides sibling terminology, there is in Korean a very elaborate system of denoting uncles and aunts depending on whether the speaker addresses their paternal or maternal, blood-related or in-law, older or younger uncles/aunts. Some of these terms have also given birth to figurative usage, being used as forms of address. Just as for the sibling terms, real kinship meanings are unlikely to be used by non-native Koreans. However, as fictive kinship terms, they are in common usage and this pragmatic implicature is not suitably represented in Korean-French dictionaries. The manner these are being described in *Nouveau* as well as the issues they give rise to are being more amply discussed in the following section.

3. Analysis of the shortcomings present in the entries surveyed

Shortcomings in the lexicographic treatment of forms of address as headwords fall into three categories: (3.1) there is no systematic pattern in describing a headword; (3.2) some significant usages of the headword are either too vague or occulted; and finally, (3.3) the Korean system of address forms being far more complex than the French one, sociocultural information is necessary within the microstructure’s text.

3.1. Lexicographic inconsistencies

In both dictionaries notes on usage, such as contextual information or register labels, is given in double brackets:

(1) **tu**¹ [...] ((chinhan salam-kacok-sonalays salam ttawieykey))
you [...] ((to close friend, relatives, inferiors etc.))

(2) **–ssi**² ((suffixe honorifique))
Mr[Mrs, Miss] ((honorific suffix))

However, other types of information, such as grammatical notes or part-of-speech description, are also presented in double brackets. Moreover, all types of information can be included within the same double brackets without distinction:

(3) **tangsin**¹¹ ((pronom de la 2e personne au singulier → grammatical information, légèrement honorifique → pragmatic information))

Not only does this method lack clarity, but it might also lead the dictionary user to dismiss the mention on usage. By skipping over the pragmatic information légèrement honorifique (semihonorific) the dictionary user may mistakenly use tangsin in situations where deferential forms of address are required.

Another practical issue is the language in which the usage information is given. This type of inconsistency raises the question of whom the dictionary is targeted at. *Prime* provides it in Korean only, clearly indicating that it has been rather designed for Korean natives learning French rather than French learners of Korean. *Nouveau*, on the other hand, provides usage information sometimes both in Korean and French, sometimes either in Korean or French, showing some visible efforts to meet the needs of the majority, that is, both Korean and French users (Heo 2000:84-85). However, the following example reveals that there is no

¹ Examples given in dictionaries have been omitted for more visibility.

116
systematic pattern in presenting pragmatic information in such-and-such a language (simple underlining for Korean pragmatic information and double underlining for French pragmatic information).

(4) enni 1 camay soeur aînée; grande sœur 2 sonwi ((pour une femme, appellation intime d’une femme plus âgée que soi))
nwuna soeur aînée ((pour un homme))
oppa 1 frère aîné (d’une femme) 2 ((appellation intime d’un homme plus âgé pour une femme))
hyeng 1 chincok grand frère, frère aîné 2 namcatul saieyse sonwiuy salamul pwulunun mal (pour les hommes); ancien, aîné

All four headwords of example (4) belong to the category of sibling terms. Three types of inconsistencies can draw our attention here. Firstly, it appears that only the headword nwuna is not given the fictive kinship deictic meaning, failing to indicate to French users that it can also be used metaphorically. Secondly, only enni and hyeng are provided information in both languages. Thirdly, the entry for the headword hyeng plainly specifies what group can make use of this word in French (pour les hommes, for men) but the most important part of the pragmatic information is given in Korean (namcatul saieyse sonwiuy salamul pwulunun mal, term to address someone older between males), making its metaphoric usage rather unclear for French users unlike the headwords enni and oppa. One may argue that French learners may not ever call their own elder brothers hyeng; nevertheless, considering the ongoing globalization phenomenon and the current trends in global education systems, situations where French male learners being led to call hyeng befriended elder Korean males are not to be ruled out. Therefore, the second section of the entry should specify for French users, rather than for Korean users, that it also is a metaphoric form of address and not merely a third person deixis.

3.2. Inaccuracy or incompleteness of equivalents
Besides these practical inconsistencies, equivalents themselves may be vague or not thorough:

(5) tangsin 1 ((pronom de la 2e personne au singulier, légèrement honorifique)) toi, vous
· cennyekun nayka sal theynikka swulun tangsini sa. Je paie le dîner; tu paies ta [une] tournée.
· tangsini mwentey namwy iley kwespanun keyo? Qui es-tu pour te mêler des affaires d’autrui? […] 2 ((pronom de la 2e personne au singulier, utilisé entre le couple)) toi, mon chéri[ma chérie], mon amour […]
you 1 ((2nd person singular pronoun, semi-honoricific)) you, you · Dinner’s on me, so drinks are on you · Who do you think you are? Mind your own business! 2 ((2nd person singular pronoun, used between couple)) you, darling, honey

While this entry from Nouveau includes two different types of pragmatic information of the address term tangsin as an honorific and as a way of addressing one’s partner, it does not represent clearly the different connotations these equivalents carry. First of all, the mention ‘légerement honorifique’ (semi-honorific) has to be clarified. It is commonly acknowledged that tangsin is the more polite counterpart of ne (Hwang 1991:123) as well as the most usual translation to second person pronouns⁸, namely tu (second person singular) and vous (second person polite). But here these equivalents are given without distinction in regard to their usage context. Nonetheless, practical usage of tangsin should be dealt with caution. If, indeed, one considers the examples given by the dictionary (here, in italic), the contrast in tone between the first and the second examples is patently visible. The first example denotes friendliness whereas the second clearly connotes aggressiveness. In fact, tangsin is a social deictic form of address that establishes a descending relation from the speaker to the interlocutor. In other words, this form is being used by a superior to an adult inferior (Sohn 2001:409). Lee & Ramsey (2000:226) have emphasized that the use of tangsin is liable to make the addressee feel

⁸ One explanation can be that tangsin, being the most generic second pronoun, is usually the one used “in advertising and book titles when referring to an unspecified you” (Brown 2011:41)
belittled, *tangsin* being merely a semi-honorific. This explains why, in case of altercation, the speaker will address the interlocutor by the term *tangsin* as a “deliberate signal of disrespect” (Brown 2011:41) since honorifics can be used ironically and metaphorically (Levinson 2011:217). Besides, it is worth noting that there are other second person pronouns which do not feature in *Prime* under the headwords *tu* and *vous*, such as *caney*, *kutay* and *caki*, perhaps because they are more restricted in usage and/or frequency.

Kinship terms, those for siblings as well as those for ‘uncle’ and ‘aunt’, keep assuming new metaphorical meanings and show a strong tendency to be used as forms of address for non-relatives (King 2006:109-110), in particular workers in customer service (Kim-Renaud 2001:36). For example, *enni* and *oppa* are used among shoppers and store clerks, regardless of age ranking9, and either reciprocally or non-reciprocally. In the same vein, the kinship terms *imo* (mother’s sister) and *samchon* (unmarried uncle from father’s side) will be commonly heard in restaurants, bars, and hair salons (Brown 2011:40) when to address the worker. Such phenomena are only a matter of a few decades and even some old school scholars deplore the improper use of *samchon* as a form of address while it originally is a third person deixis (King 2006:109). Whether these aspects justify it, the fact is that the aforementioned pragmatic uses of *enni*, *oppa*, *imo*, and *samchon* are not represented in current French—Korean lexicography.

3.3. Lack of sociocultural information

We have seen in the previous subsection that recently *imo* (mother’s sister) tended to be used to address female workers. Unlike *enni*, *imo* does carry the generational connotation, that is, *imo* is mother’s generation. Beforehand, the address term *acwumma*, or rather its non-contracted form *acwumeni* used to be the term used in these situations. The treatment of *acwumeni* in *Nouveau* is not without raising issues.

(6) *acwumeni* 1 ((chincok)) tante 2 ((ilpan hoching, une femme d’un certain âge dans la rue)) Madame • ajwumeni, annyenghaseyyo? Bonjour, Madame!
aunt 1 ((relative)) aunt 2 ((general address term, of a middle-aged woman in the street))
Madam • Good morning, Madam!

Whereas originally the term designated and addressed one’s aunt, its metaphorical use overrode the literal use, a phenomenon which is not represented here. But more interestingly, *Nouveau* renders in the example *acwumeni* by Madame. Madame in French may have lost its honorific value but it remains the polite way to address female, whether stranger or not close. However, the Korean usage of *acwumeni* is not perceived as polite, as it appeared in an experiment conducted by Yun-Roger (2010:164-165) who investigated how addressees experienced the Korean terms of address. Thus, the equivalent *Madame* without warning about the real sociocultural perception of the term is clearly misleading and can bring the learner of Korean into thorny situations. In *Prime*, the headword *Madame*10 is not given the equivalent *acwumma/acwumeni*. In fact, there is no mention either of other address forms such as the suffix -*ssi*, and the deferential *seasyngsim* (teacher), which all translate as Madame in *Nouveau*. Equivalents and examples painfully lack of sociocultural information as shown below.

(7) -*ssi* ((suffixe honorifique)) monsieur (messieurs); madame (mesdames);
mademoiselle (mesdemoiselles)
Mr[Mrs, Miss] ((honorific suffix)) Mr (Messrs), Mrs, Miss
*sensayngnim* (Monsieur [Madame] le(la)) professeur(e), Monsieur (Madame)
teacher teacher/professor, Sir (Madam)

The lack of cultural information make these two forms of address appear interchangeable where they should be disambiguated. As Brown (2011:40) explains, “although -*ssi* indexes a

9 Nevertheless these forms would more easily apply if the shopper looks young.
10 In the continuity of what is said just above, I will only focus on Madame and dismiss the case of Monsieur.
certain degree of ‘separation’ – and indeed is never used amongst intimates – it can hardly be deemed honorific or deferential” (Brown 2011:40). Originally, ‘[Full Name]+ssi’ was used to address newly hired or junior employees who had no title yet (Yun-Roger 2010:163). A recent phenomenon consisting of using the first name only with the suffix –ssi is broadly observed, especially among women and young people, as a sign of equalizing statuses between speaker and addressee (Kim-Renaud 2001:36). Thus, it is acknowledged that the title –ssi, while remaining polite, contains a rather faint load of deference. As for sensayngnim, the entry is significantly unclear for a French learner. Firstly, the headword should be identified as honorific reference and address term. Secondly, the comma after professeur(e) may lead the dictionary user to think that the following only refers to teachers and is not a separate equivalent. A semicolon would have proven a more judicious choice because in French, while not impossible, students rarely address professors as Professeur11 but instead as Monsieur or Madame. In Korean instead, sensayngnim has become a generic “title of deference for both women and men, married or not”, regardless of their actual profession (Kim-Renaud 2001:30) and it is surprising that this title is absent from both the equivalents of Monsieur (Example 3.a) and Madame (Example 3.b). The extension of sensayngnim’s acceptation to anyone whose situation is not clearly identified is also a recent phenomenon12. As language evolves rapidly alongside society, the decision whether to include these sociocultural changes or not raises the issue of the dictionaries’ limits regarding cultural pragmatic information (Celotti 2002/4:459). The present shortcomings in this matter only reflect the failure to recognize the bilingual dictionary as a vehicle of cultural identities rather than a sole translation tool.

4. Culture, pragmatics and bilingual lexicography

4.1. The bilingual dictionary as epitomizing two views of the world

There is a strong discrepancy in the understanding of propriety between French and Korean. The Korean language is stratified into deferential, polite, blunt, familiar, intimate and plain styles. All these styles represent each a modality of politeness and can be grouped into two generic systems (honorification system and non-honorification system)13. The speech style rigidly relies on the addressee’s status, the speaker’s status and, above all, their relationship. Brown (2011) contends that this schematic system of politeness that governs the Korean language and interpersonal interactions is underpinned by the ideology of the ‘face’ as defined by Brown and Levinson. They define the notion of ‘face’ as the wants either to be recognized or to prevent the potential threat of the others, necessitating thereby ‘politeness strategies’ (Hickey & Vazquez Orta 1994). Thus, by choosing the appropriate style depending on the addressee’s status, the speaker elude the threat of the addressee being neglected in his social status and him/her self being threatened in turn (Brown 2001:76). While Korean politeness is governed by non-egalitarian patterns proper to a highly hierarchical society, French politeness, on the other hand, in the manner of Western ideology is applied “equally regardless of age, rank, sex, etc.” (op.cit. p.80). This deep-rooted divide can be exemplified by the case of the French address title Mademoiselle (Miss). Madame and Mademoiselle only diverge on the marital status criterion, Mademoiselle being an appellation for an unmarried woman (younger or older). In the common language, this term has been extended to the acceptation ‘young woman’. In January 2013 the Council of State announced the suppression of Mademoiselle from administrative documents to equalize the situation between men and woman as Monsieur does not prejudge a man’s marital status14. In fact, the use of Mademoiselle can be

---

11 However, kindergarten and primary school teachers are being addressed as Maître (Master) for male teachers and Maîtresse (Mistress) for female teachers.
12 To the extent that merely 25 years ago a scholar “complained that everyone in Korea now is sônsaengnim” (King 2006:109)
13 Vide Sohn’s (2001:414–415) comparative diagrammes illustrating each system: honorification (diagramme 246) and non-honorification (diagramme 247).
14 Vide the official website of the French public service: <http://www.service-public.fr/actualites/002616.html>. It is further specified that Mademoiselle can still be in use privately.
source of flattery as well as unpleasantness, as *Mademoiselle* connotes rather youth than marital status. That is, calling *Mademoiselle* a woman who is evidently not so young can be, on the one hand, a strategic flattery if she is married, but on the other, it can emphasize that she is outside the standard social model and even hurt her feelings if she is not married. Therefore, the term *Madame* is preferred to address a woman no matter her status or how young she may look\(^{15}\), especially in formal situations or unfamiliar relationship. In sum, this shows the strong tendency in France towards neutrality or equality.

Besides, it is worth noting that such titles usually assume a “phatic function” (Kerbrat-Orecchioni 2010:6), that is to say, they are used to hail or draw the attention of the interlocutor, establishing thereby an “interpersonal relationship” (*ibid.*) or, following Levinson’s words, a “relational social dimension” (2011:206-207) between two participants in a conversation, after which they are dropped in favour of the use of pronouns\(^{16}\). Following Brown & Gilman’s (1960) analysis of the European usage of the pronouns *Tu-Vous*, the French societal model has experienced a major shift. In unequal situations, when the speaker is lower than the addressee (for instance, child and parents, or waiter and customer) the superior receive the pronoun *vous* and the inferior, *tu*. In today’s France children reciprocally address both parents as *tu* and customers reciprocally say *vous* to service providers: the tendency in social relationships shifted from asymmetric power relation to symmetric solidarity. On the other hand, Korean reflects rather plainly the verticality of the relationships between speaker and addressee, especially in parents/children relationships. According to Yun-Roger (2010:162-163) the Korean language of politeness reflects the attitudes proper to the Confucian thought that moulded the Korean society for centuries: the importance of knowing one’s place in the world, the reverential mindset towards education, and the perception of social relations as extensions of familial relationships. Despite Brown and Levinson’s aim to reach the universality of human behaviours with the theory of the ‘face’, one must not neglect indeed the significance of regional identities.

**4.2. Bridging the cultural gaps through ‘lexicultural pragmatics’**

The cultural dichotomy between French society and Korean society renders the task of compiling French↔Korean dictionaries all the more difficult. According to Svensén (2009), there should be ideally four bilingual dictionaries for each of the languages dealt with, serving four distinct purposes (*op.cit.* 14-18):

- Reception (for understanding target language (L2))
- ‘Passive’ translation (for translating L2 to source language (L1))
- Production (for producing L2)
- ‘Active’ translation (for translating L1 to L2)

However, this seems to be rather limited due to practical issues. In today’s society, one crucial factor of making a dictionary is its capacity for being sold; therefore, publishers synthetized all four functions into a hybrid dictionary where L1-to-L2 information and L2-to-L1 information are blended together even under a same headword, resulting in confusing or inaccurate descriptions of the headwords that are hard to translate or have no equivalents in the other language, as it is sometimes the case for terms of address.

Indeed, Galisson (1988:331) explains that, while every word is cultural, some of them are more than the others; that is to say, they are loaded with what he calls ‘shared cultural charge’ (SCC)\(^{18}\). The list of all the words with SCC constitutes the “lexiculture”. As regards terms of address in Korean, we can cite the case of *imo* seen above (3.2) which literally designates a

---

\(^{15}\) Unless it is obvious that she is a minor or still in secondary school.

\(^{16}\) *Vous* always comes after titles of address such as *Monsieur, Madame, Mademoiselle, Maître, Maîtresse* and so on.

\(^{17}\) This concept has been brought to light by Galisson (1988)

\(^{18}\) In French, *Charge Culturelle Partagée* (CCP).
mother’s sister but can also be used as a fictive kinship appellation to address unrelated women who work, for instance, as waitresses or female hairdressers. We have seen that this latter acceptation is not given in *Nouveau*; however, as a feature of the Korean culture and society, its mention would have proven helpful for a French learner not only to understand this aspect of Korean society but also to improve his(her) production of Korean. Thus, lexicultural pragmatics plays a crucial role in bilingual lexicography and can either be presented in a separate note following recent trends (Celotti, 2002/4:458) or constitute a ‘contextualized definition’ (De Surmont 2010:26) in equivalents’ stead as seen with the case of *enni* and *hyeng* in example 4 (respectively, *pour une femme, appellation intime d’une femme plus âgée que soi* (intimate way for a female to address another female who is older than her) and *pour les hommes* (for men)). While French has an equivalent to their literal meaning, it does not have one for its lexicultural acceptation, hence the necessity of a contextualized definition.

There are two treatments of lexicultural information. The first regards the language (that is, which user is it designed for?), the second the typographic form in which it is presented. For *enni*, the contextualized definition is given in French with a heading in Korean. As for *hyeng*, it provides the Korean user with an unnecessary detailed heading *namcatul saiyesey sonwiuy salamul pwulunun mal* (term to address someone older between males) and provides the French user with an insufficient definitory context followed by an attempt of equivalence. However, a Korean user does not need the extended lexicultural information of *hyeng*: the heading could have been as short as it is under the headword *enni*: *sonwi* (someone older). From a French perspective instead, it would have proven much useful. The parsimonious contextualized definition *pour les hommes* (for men), even combined with the translations *ancien, aîné* (senior, elder), does not help the French user to fully grasp the practical use of *hyeng* as an actual term to address unrelated people. Furthermore, it can be said that there is no relevant exemplification showing the word in context for any of these headwords. Indeed, definitory contexts only help the user have a conceptual value of the headword; the realization of it only happens through the act of enunciation (Schwuer 1994:5-6). It would have been useful to include exemplification to the headwords that would put into context the practical usage of the metaphorical meanings (e.g. short dialogues), rather than simple equivalents. From here, we can further go with two observations. Firstly, the equivalents *ancien, aîné* (senior, elder) given under the headword *hyeng* are only ‘partial equivalents’ (Wiegand 2002:245). That is to say, they are semantically equivalent but pragmatically incomplete. *Hyeng* can be translated as *ancien* or rather *aîné* as third person reference but not as term of address. Secondly, the headings *camay/chincok* and *sonwi* should also be provided in French for more clarity for a French user. Schwuer (1994:7) has highlighted the significance of such headings, or as she puts it ‘semantic describers’. They constitute indeed another significant factor in giving a deep insight of a SCC word. Semantic describers, by categorizing the various usages of a lemma, help recreate meaning by creating ‘conceptual domains’ in addition to the ‘semantic basis’ which consists of ‘sets of equivalents’ or contextualized definitions, and notes on usage, which should not be a mere translation of L1 definitions into target language because pragmatic uses of words and cultural divide would inevitably been lost in translation.

5. Concluding remarks: towards a revalorization of the status of bilingual dictionaries

19 Or ‘definitory context’ (ibid.)

20 In fact, *frère* can also be used metaphorically to address a group of compatriots or a close friend; however, the term is usually determined by the possessive *mon, mes* (my) and sometimes by the affectionate attribute *cher* (dear). In addition, this usage of *frère* mainly has a phatic function (*Vide* Section 3.1) and often takes on an ironical connotation. The only other instance when *frère* is used as regular term of address for an unrelated man is to address a member of a religious group, especially a monk. This is also valid for *soeur*. Nevertheless, these usages of *frère* and *soeur* are very different from the lexicultural usage of *hyeng* and *enni*.
Schwuer (1994:5) fairly pointed out that bilingual dictionaries are given a hard time to the extent that L2 teachers often recommend against using them. Despite the fact that lexicographic theories and tools continue being developed, L2 speakers, nonetheless, keep repeating the same translation mistakes. Are bilingual dictionaries inadequate by nature? Are they too complicated to be used by laymen? Should we then invent the *lexicopraxis*, theorizing the methods of using the dictionary? Perhaps we need not go so far. Instead, present shortcomings should be dealt with more precisely. First, the type of information (grammatical, pragmatic etc.) should be clearly defined and its treatment systemized. A dictionary user, whether inexperienced or not, should be able to see the skeleton of the microstructural text before even tackling the meaning of its contents. Secondly and specifically to bilingual dictionaries, they should reflect the cultural realities the two languages bear, whether from a contrastive or a cross-cultural perspective. Nevertheless, a difficulty lies in the restrictions on the size of the dictionary, thereby, on the volume of information. Not everything can feature in the dictionary; therefore, it is incumbent on the lexicographer to choose what lexicultural information should be included or not. The subjectivity of lexicographers also is a delicate question. Through their choices, however based on the most thorough and exhaustive corpus, the language that describes, suggests, or silences the world, is fashioned from their intuition and point of view. As Galisson (1988:333) puts it, “dictionaries are not objectively descriptive, as they pretend it, but subjectively normative, as they do not confess it.”

**References**


Heo, JY. 2000. ‘Un nouveau dictionnaire bilingue coréen-français d’apprentissage.’ 프랑스어교육 제10집, 한국프랑스어문교육학회, 83-101


Dictionaries

*Nouveau dictionnaire coréen-français (say-hanpwul-sacen)*. Société Coréenne de Langue et Littérature française (한국불어불문학회), 한국외국어대학교 출판부 (electronic dictionary)


http://stdweb2.korean.go.kr/main.jsp
The Development of the Dictionary of the Usage Frequency of Basic Korean Vocabulary and Its Application to Korean Lexicology

Seo, Sang-Kyu, Kim, Jinung
ILIIS, Yonsei University, South Korea
inaka@yonsei.ac.kr, borgesk@utexas.edu

Abstract
The purpose of this presentation is to show the methodologies of a quantitative analysis of data applicable to corpus linguistics, lexicography, and Korean language and information by elucidating the process of developing the Dictionary of Usage Frequency of Basic Korean Vocabulary (2014). We carried out the process of extracting information from the corpora and reshaped them into a body of knowledge applicable to linguistic education and research. The number of word forms defined in the Dictionary is 7,203 (including 58 affixes), which covers 14.9% of 48,281 contents words from the corpus of standard Korean words. Interestingly, the number of word tokens of only 7,203 words is 803,820 which cover 76.4% of the corpus of standard Korean words. In other words, 76.4% of the corpus of standard Korean words is annotated with word sense.

Keywords: dictionary, corpus, Korean, vocabulary, frequency

1. Introduction
The purpose of this presentation is to show the methodologies of a quantitative analysis of data applicable to corpus linguistics, lexicography, and Korean language and information by elucidating the process of developing the Dictionary of Usage Frequency of Basic Korean Vocabulary (2014). We carried out the process of extracting information from the corpora and reshaped them into a body of knowledge applicable to linguistic education and research. The number of word forms defined in the Dictionary is 7,203 (including 58 affixes), which covers 14.9% of 48,281 contents words from the corpus of standard Korean words. Interestingly, the number of word tokens of only 7,203 words is 803,820 which cover 76.4% of the corpus of standard Korean words. In other words, 76.4% of the corpus of standard Korean words is annotated with word sense. In particular, this presentation deals with the following topics.

1) A detailed introduction of the background, structure, and characteristics of the Dictionary of Usage Frequency of Basic Korean Vocabulary, which has been compiled from 1998 to 2014
2) Methodologies in selecting basic vocabulary and research on usage frequency
3) Challenges faced while compiling the dictionary and their solutions
4) Questions that arose from the compilation of the Dictionary

2. What is the need for a dictionary of the usage frequency of basic Korean vocabulary?
There are several words which have a singular meaning, but in most cases words are used with multiple meanings. Such polysemic characteristics are more prominent in words which are frequently used in everyday life (or, with very high usage frequencies). To examine the aspects of the homograph entries and senses of words, 300 words with the highest frequencies in the Vocabulary Frequency of Modern Korean Language (1, 2) (Seo, 1998a) as investigated from the Yonsei corpus (approximately 40,000,000 mati) were studied to determine how many homograph entries and senses were defined in the Yonsei Korean Dictionary (1998). The results showed that one word had an average of approximately 1.7 homograph entries of the same part of speech (approximately 3.5 when including different parts of speech), and an average of 7.2 senses. Moreover, when the entries in the Dictionary of Usage Frequency of Basic Korean Vocabulary (2014)
in a dictionary, and through this we can understand the various meanings of a word. However, it is
difficult to know which meaning is more important or more common. In such situations, we can only
rely on experiential judgment or instinct. Generally, when defining a dictionary meaning, the central
(or etymological) meaning is placed first, after which the scope of the meaning is widened to
peripheral (expanded) meanings. Below is an example from the Korean Standard Dictionary.

(1) Sichimi
① quadrilateral horn tied inside the tail feathers of a falcon with written address to declare the owner
of the falcon ≡ danjangpan
② the attitude which pretends not to have done something when they did, or feigning ignorance
(Korean Standard Dictionary, 1999)

In modern Korean language, meaning ② is not only more common, meaning ① is only used in a very
specific context. However, the etymological meaning ① appears first in the dictionary due to general
dictionary recording method. The dictionary description does not match the reality of the language. In
fact, in the ‘corpus of standard Korean words’ (million words), the word sichimi is used 14 times in 11
texts, in each case used with meaning ② i.e., feigning ignorance.

When students are learning the meaning of words during Korean language education, it is very helpful
to decide on basic educational vocabulary. When establishing stages of vocabulary education for the
purpose of language education, the usage frequency of a word is considered as a very important
element. Words included in basic vocabulary are naturally words with very high usage frequencies.
Additionally, words with such high usage frequencies are inevitably polysemic words with a relatively
wide scope of meaning. However, it is questionable whether all meanings of a word in a basic
vocabulary are seen as basic. For example, the noun dui is included in the Korean learner’s vocabulary
list compiled by the National Academy of the Korean Language, but ten different meaning categories
are defined for the word in the Yonsei Korean Dictionary (1998). It is impossible for Korean learners
to understand and memorize the entire list of meanings for the word dui. We need to consider which
meaning is more important, or should be learned first, from the polysemic meanings of words in the
basic educational vocabulary. By investigating frequencies in a large-scale corpus, the Dictionary of
the Usage Frequency of Basic Korean Vocabulary can suggest ‘words which are used at high
frequencies and which appear in diverse language resources’, and present quantitative information on
the senses of the words to indicate the degree of importance objectively within the senses of the words.

3. Methodology used to compile the dictionary of the usage frequency of basic Korean vocabulary

3.1. Procedures in compiling the dictionary of usage frequency

In a dictionary of usage frequencies, the previous research accomplishments of basic Korean
vocabulary studies should be reflected, and entries should be described according to the analysis of
usage frequency of meanings and usages based on corpus. Ample and easy examples should also be
provided. To develop the Dictionary of the Usage Frequency of Basic Korean Vocabulary based on the
results of analyzing the corpus, the following procedures are necessary as a minimum.\(^{22}\)

(2) Elements and procedures needed for developing a dictionary of usage frequencies
   a. Select the basic Korean vocabulary
   b. Establish a balanced corpus
   c. Annotate the morphological information of the corpus
   d. Annotate the semantic information of the corpus
   e. Extract the usage frequencies
   f. Design and apply a model for the dictionary of usage frequencies

Each procedure of (2) is performed in order; (a) seeks to define the range and list of the investigative
subjects, while (b) is done to test the validity of the basic vocabulary list selected in the previous stage

\(^{22}\) This procedure and methodology were established by Seo Sang-Kyu, Nam Yoon-jin, and Jeon Ki-ho
and to compile resources to investigate the usage distribution according to the meanings for each of the words, (c) and (d) provide the shape of the actual investigation process, where the corpus constructed in (b) was reorganized with the added word analyses results of morphological and semantic annotations. Here, it is effective to use an existing dictionary for the criteria for annotating senses; hence, we selected the Yonsei Korean Dictionary (1998), which was the first Korean dictionary to be based on example analyses of actual corpus.

When the procedures of (a) to (d) are complete, this becomes the basis for the quantitative processing of (e). The result is a usage frequency information database for basic vocabulary. Here, (f) signifies the manufacturing of the final results and the determination of the content and scope of the vocabulary information to be extracted from the entire process of the usage frequency dictionary. Chapter 3 examines the procedures of (a) to (d), while (e) and (f) will be examined in Chapter 4, where the structure and characteristics of the final product will be discussed.

3.1.1. Selecting basic Korean vocabulary
For the compilation of this dictionary, a large-scale corpus, various Korean language textbooks, basic vocabulary lists, and dictionaries were analyzed to obtain three types of basic vocabulary lists selected through objective criteria in three aspects. First, these words were elaborately distinguished from common language resources (corpus), Korean language textbooks, basic vocabulary lists from experts, or key vocabulary in learner’s dictionaries. Second, these words were used widely in many language resources from the most elaborately analyzed semantic annotation corpus. Third, these words appeared to be used frequently in new corpuses with completely different compositions, indicating that they are stably used words which are not influenced by the method of collection and/or the structure of the language resource.

The key vocabulary suggested in the Korean Learner’s Dictionary for Foreigners (Seo et al., 2004, 2006) did not distinguish between homograph entries, but this dictionary attained objectivity in vocabulary selection by accurately identifying the usage frequencies of word senses through a more in-depth analysis. Moreover, it resolved the largest obstacle for in the Korean Learner’s Dictionary for Foreigners. Previously it was virtually impossible to distinguish between homograph entries in frequency studies of large-scale corpuses. When at the stage of selecting one from numerous homograph entries to include in a basic vocabulary, the only option was for the researcher to choose instinctively.

3.1.2. Structure of a balanced corpus
The corpus of standard Korean words consists of one million mati23, and the sample size for each text was limited to 5,000 mati in principle. However, in cases where the resource is a transcribed recording of spoken words or a short text, the entire body of the text was included. The smallest sample in this corpus is 506 mati, and the largest sample is 18,437 mati. The finalized version of the corpus of standard Korean words consisted of total 218 texts, with 86.3% of this in the written language and 13.7% in the spoken language form. The actual structure of the corpus of standard Korean words is shown in Table 124.

<table>
<thead>
<tr>
<th>Medium</th>
<th>Genre</th>
<th>Type</th>
<th>Number of text</th>
<th>Number of mati</th>
<th>Proportion</th>
<th>Number of Mati according to genre</th>
</tr>
</thead>
</table>

23 As already identified in Seo (1998b:226), ‘mati’ (number) is a “unit reflecting the amount of corpus material, using [the] word spacing as [a] separator, and [which] is used as a similar concept to the common ‘ejeol’”. However, it is rare to find examples where the regulations of word spacings are thoroughly applied in actual resources, and it is difficult to combine these instances into a consistent principle. Therefore, ‘mati’ in this instance indicates a word unit separated by spacings in actual language resources.”

24 This corpus was designed and structured with one million mati initially, but during the actual annotation process, certain inappropriate material (such as texts with a high percentage of archaic sayings or Chinese poems) were removed or replaced. Ultimately, the final product was confirmed to have approximately 990,000 mati. Refer to Seo Sang-gyu, Choi Ho-cheol, Kang Hyun-hwa (1999), Seo Sang-gyu, Kang Hyun-hwa, Yoo Hyun-gyung (2000), and Seo Sang-gyu (2006).
<table>
<thead>
<tr>
<th>Category</th>
<th>Subject</th>
<th>Count</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Written language</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-academic writing</td>
<td>Humanities</td>
<td>7</td>
<td>3.7%</td>
</tr>
<tr>
<td></td>
<td>social sciences</td>
<td>7</td>
<td>3.7%</td>
</tr>
<tr>
<td></td>
<td>Art</td>
<td>9</td>
<td>4.9%</td>
</tr>
<tr>
<td></td>
<td>Technology</td>
<td>2</td>
<td>1.1%</td>
</tr>
<tr>
<td>Creative writing</td>
<td>Fairy tales</td>
<td>5</td>
<td>2.3%</td>
</tr>
<tr>
<td></td>
<td>Novels</td>
<td>28</td>
<td>14.4%</td>
</tr>
<tr>
<td></td>
<td>Essays</td>
<td>12</td>
<td>6.0%</td>
</tr>
<tr>
<td>Practical writing</td>
<td>current issues</td>
<td>7</td>
<td>3.6%</td>
</tr>
<tr>
<td></td>
<td>information</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private writing</td>
<td>Memoirs/biography</td>
<td>15</td>
<td>7.7%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>218</td>
<td>100%</td>
</tr>
<tr>
<td><strong>Spoken language</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcribed speech</td>
<td>Daily conversation</td>
<td>8</td>
<td>1.2%</td>
</tr>
<tr>
<td></td>
<td>Topic conversation</td>
<td>5</td>
<td>1.9%</td>
</tr>
<tr>
<td></td>
<td>Counseling interviews</td>
<td>1</td>
<td>0.1%</td>
</tr>
<tr>
<td></td>
<td>Lectures/talks</td>
<td>1</td>
<td>0.5%</td>
</tr>
<tr>
<td></td>
<td>Broadcasting news</td>
<td>1</td>
<td>0.4%</td>
</tr>
<tr>
<td></td>
<td>Broadcasting conversation</td>
<td>1</td>
<td>0.5%</td>
</tr>
<tr>
<td></td>
<td>Broadcasting counseling</td>
<td>1</td>
<td>0.4%</td>
</tr>
<tr>
<td></td>
<td>Broadcasting discussion</td>
<td>1</td>
<td>1.2%</td>
</tr>
<tr>
<td>Semi-spoken language</td>
<td>Drama scripts</td>
<td>4</td>
<td>1.9%</td>
</tr>
<tr>
<td></td>
<td>Play scripts</td>
<td>4</td>
<td>2.7%</td>
</tr>
<tr>
<td></td>
<td>Screenplay scripts</td>
<td>3</td>
<td>2.0%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>218</td>
<td>100%</td>
</tr>
</tbody>
</table>

3.1.3. Annotating the morphological information of the corpus
The groundwork for annotating semantic information is achieved by dividing each ‘eojeol’ of the text based on the word subject for semantic annotation and then attaching a basic part-of-
speech label for the divided unit. The process of annotating morphological information is given below.

(3) Process of annotating morphological information
a. Attach a part-of-speech label through an automatic morpheme analyzer program-
   HAM_KTS is used to attach the eojeol analysis (morphological information) automatically to
   all one million mati in the corpus of standard Korean words25
b. Revised part-of-speech label by hand using the semiautomatic annotation program
   ‘Kkoridalgae’

3.1.4. Annotating the semantic information of the corpus
Upon the completion of the morphological information annotation of the corpus in the
previous stage, the annotation of semantic information for the subject words in the corpus is
 carried out. The practical method is to select a foundation dictionary and to distinguish
between different meanings based on how the word is defined in the dictionary. The
dictionary to be used should satisfy a few conditions. First, the categorization of meaning
should be precise and credible. Second, all of the words subjected to annotation should be
included in the dictionary, and all usages which can appear in the corpus should be defined.
Third, it must be a dictionary constructed on the basis of an analysis of actual language
resources. Fourth, it is also very important that there is an electronic version of the dictionary
given that such a dictionary is essential in annotating the actual corpus. Considering the above,
the Yonsei Korean Dictionary (1998) was selected as the foundation for semantic annotation,
and an electronic dictionary was obtained to be used in the program for semantic annotation.
The Yonsei Korean Dictionary (1998) was the only dictionary which satisfied nearly all the
required conditions for a foundation dictionary for semantic annotation; the entire dictionary
was digitalized in September of 2000 and free internet search services were being provided at
the time of annotation (http://kordic.britannica.co.kr). At the same time, for an efficient and
accurate semantic annotation process, we developed and used a semi-automatic program
which searches the dictionary and attaches tags.

The subjects for semantic annotation in principle were actual vocabulary items centering on
basic vocabulary, with some suffixes included. In total, 7,203 words (including 58 affixes)
were published in the Dictionary of the Usage Frequency of Basic Korean Vocabulary, and
although this is only 14.9% of all words in the corpus of standard Korean words, when the
usage examples are added, it becomes 803,820, which is as much as 76.4% of the entire
corpus. In other words, 76.4% of the entire corpus was separated by meaning and was
annotated.

3.2. Problems in the process of compiling the usage frequency dictionary and solutions
3.2.1. Determining the credibility of the basic vocabulary
In section 3.1.1, the importance of obtaining a basic vocabulary list for the compilation of this
dictionary was pointed out. Regarding the selection of basic vocabulary for Korean language
education, there should be rational answers to questions as to which resources and
information are used for the selection, what the basis is for determining the importance of
certain words above other words, and whether a list of 1,000 or 5,000 words can be compiled
systematically with only the knowledge in our heads.

To establish a ‘basic Korean vocabulary’ for the Dictionary of the Usage Frequency of Basic
Korean Vocabulary, we initially obtained three sub-lists with independent characteristics.
Corpus material (A), a Korean language textbook (B), and a basic vocabulary list of experts
or key vocabulary in learner’s dictionaries (C) have different characteristics because they use
different objective selection methodologies. The basic vocabulary list from the three sub-lists

25 ‘KTS’ was developed by the Natural Language Processing team at the Korean Advanced Institute for
Science and Technology (KAIST), and ‘HAM’ was developed by Professor Kang Seung-sik at
Kookmin University.
(approximately 2,000 words) is combined (in other words, \( A \cup B \cup C \)) to formulate the ‘total basic vocabulary’. On the other hand, the words which are included on all three sub-lists (in other words, \( A \cap B \cap C \)) are referred to as ‘critically important vocabulary’ in this paper. This list is based on the three sub-lists but is further distinguished by introducing quantitative information and utilization concepts related to each item.

Discordance between the analysis of the dictionary and the corpus analysis
The Dictionary of the Usage Frequency of Basic Korean Vocabulary selected the Yonsei Korean Dictionary (1998) as the foundation dictionary for semantic annotation. Through the semi-automatic semantic annotation program ‘Chamteut’, the semantic annotator performed the process of attaching semantic annotations to the corpus, which had been morphologically annotated. During this process, there were numerous examples where despite confirmation in the corpus, the meaning of an item was not defined in the dictionary. In this case, the dictionary database was revised and supplemented. In Figure 1, the dictionary information of the verb geurida is shown, and when a word sense is clicked, a small window opens at the top to show the usage examples recorded in the dictionary. The word sense, indicated as ‘x’ on the bottom, shows that it was added during the semantic annotation stage. Therefore, it did not stop after extracting the usage frequency of the relevant word through semantic annotation but revised and supplemented the dictionary content by applying the results of the actual usage analysis.

Figure 1 Dictionary screen of the semantic annotation program ‘Chamteut 2.1’ (geurida)

4. Structure and characteristics of the dictionary of the usage frequency of basic Korean vocabulary

There are certain conditions necessary for the usage frequency dictionary to be realistically helpful in the teaching and learning of Korean, or to assist learners in their efforts to understand the characteristic meanings of Korean vocabulary items. First, it is desirable for the actual usage of the basic vocabulary to be expressed through a statistical analysis of all of the examples in a large-scale standard corpus. Second, the basic vocabulary item and its importance along with its weight considering various homonyms should be easily identified. Third, the quantitative information of the meaning senses of each word should be shown so as to be utilized in Korean language education. To do this, it is desirable to heighten the level of user convenience by considering the dictionary structure, structural elements, and method of expressing the frequency. The following paragraphs will examine how the above conditions
are reflected in the structure of the Dictionary of the Usage Frequency of Basic Korean Vocabulary (2014).

4.1. Entries
From the corpus of standard Korean words, the words which appear in ten or more texts and with a frequency of 15 times or more (5,162 words) were entered into this dictionary along with their homographic entries, reaching a total of 7,203 words.

<table>
<thead>
<tr>
<th>Entry</th>
<th>Text Frequency</th>
<th>Freq1</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>“gamda1”</td>
<td>129</td>
<td>81</td>
<td>62.8%</td>
</tr>
<tr>
<td>“gamda2”</td>
<td>27</td>
<td>16</td>
<td>59.3%</td>
</tr>
</tbody>
</table>

Figure 2 Part of the Dictionary of the Usage Frequency of Basic Korean Vocabulary

From these, entries with three black or white stars next to the part-of-speech label are ‘basic Korean vocabulary’. These consist of only actual vocabulary (including some affixes) items, numbering 2,962. ‘Text’ indicates how many texts in which the relevant entry appears (used) out of all 218 text samples comprising the subject corpus, and when appearing in all texts, the maximum number is 218. ‘Freq1’ is the appearance (use) frequency (in other words, the absolute frequency) of the relevant entry shown in the subject corpus. The percentage inside the parentheses shows the ratio of the frequency from the total sum of frequency of the entry (in other words, the relative frequency). Therefore, in Figure 2, “Freq1=81(62.8%)” for ‘gamda1’ means that this entry was used 81 times in the entire corpus, covering 62.8% of the total use of the entry ‘gamda’. ‘Freq2’ is the frequency of the relevant meaning (word sense), and the percentage inside the parentheses shows the ratio of the relevant entry for the total frequency (Freq1). This ratio shows which meaning is used more often from the multiple meanings of a single entry.

4.2. Usage frequency and importance
The basic Korean vocabulary indicated by a combination of black stars (★) and white stars (☆) is based on the ‘key vocabulary list of the Korean learner’s dictionary’ published by Seo (2003a). This vocabulary list was used in the Korean Learner’s Dictionary for Foreigners
(2004, 2006) as 2,975 key vocabulary words (basic Korean vocabulary). By applying a semantic annotation analysis of the corpus with the meaning and usage frequency research results carried out while compiling the usage frequency dictionary, the final 2,962 words were selected as basic Korean vocabulary items. An indication of the importance of the basic vocabulary (three black stars) and the number of words can be seen in the following table.

### Table 2 Importance and Distribution of Basic Korean Vocabulary

<table>
<thead>
<tr>
<th>Degree of importance</th>
<th>Number of words</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>★★★ (3 lists)</td>
<td>1,128</td>
<td>38.0%</td>
</tr>
<tr>
<td>★★☆ (2 lists)</td>
<td>377</td>
<td>12.7%</td>
</tr>
<tr>
<td>★☆★ (2 lists)</td>
<td>242</td>
<td>8.2%</td>
</tr>
<tr>
<td>☆★★ (2 lists)</td>
<td>184</td>
<td>6.2%</td>
</tr>
<tr>
<td>★☆☆ (1 list)</td>
<td>287</td>
<td>9.7%</td>
</tr>
<tr>
<td>☆★☆ (1 list)</td>
<td>312</td>
<td>10.5%</td>
</tr>
<tr>
<td>☆☆★ (1 list)</td>
<td>436</td>
<td>14.7%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2,966</td>
<td>100%</td>
</tr>
</tbody>
</table>

Words with a black star in the first place 26 commonly appear at a high frequency in various large-scale corpora and are words which are included on the high-frequency key vocabulary list (around 2,000 words). In others words, these words are widely used in modern Korean language. Words with a black star in the second place 27 commonly appear as key vocabulary items in approximately 30 types of Korean language textbooks, including textbooks from major Korean language education institutes, showing that they are included in the Korean language education key vocabulary (around 2,000 words). Words with a black star in the third place 28 show that these words appeared in 14 different basic vocabulary lists published by Korean language education experts and in major learners’ dictionaries (around 2,000 words).

The purpose of this list is to find words that are commonly recognized as important based on various lists compiled through the ‘subjective decisions’ of experts or dictionary publishers. Therefore, the 1,128 words with black stars in all three places in Table 2 are used widely at a high frequency in everyday life, are dealt with importantly in Korean language textbooks, and are recommended by experts and learners’ dictionary publishers.

### 5. Conclusion

We examined the process of developing the Dictionary of the Usage Frequency of Basic Korean Vocabulary, which had been implemented since 1998, and the characteristics of the dictionary structure. We also examined the necessary elements to be considered in the annotation and manufacturing of corporuses, which will become more active in the future. This dictionary does not stop at simply suggesting a single detailed method to utilize corporuses but also shows that we should deeply understand the characteristics of actual language use and distributional information which had relatively ignored thus far, while also taking interest in

---

26 Words of importance labelled with ①②③⑤ refer to this type.
27 Words of importance labelled with ①②④⑥ refer to this type.
28 Words of importance labelled with ①③④⑦ refer to this type.
quantitative and qualitative analyses of this information. Presently, based on the results of analyzing the reality of the meanings and usages of basic vocabulary items, it is possible to seek research methodologies which can deepen vocabulary and grammar research. Finally, we suggest several new questions which arose while developing this dictionary.

1) For words which are used as two or more parts-of-speech, are the actual usages in meaning always the same, or does the central use differ according to the part-of-speech? 2) For synonym or antonym pairs, do the frequencies of use mechanically correspond to each other in actual usage frequency? 3) For abbreviations and original words, can the usage frequency be mechanically predicted? 4) For words in a derivative relationship, is the aspect of usage frequency equally maintained, or can it differ?

References

Problems and Countermeasures in Chinese Lexicography Modernization

KANG Shiyong¹, DUAN Haifeng²
Ludong University, China
¹kangsy64@163.com, ²haifeng-duan@126.com

Abstract
Chinese lexicographical modernization has the following problems. 1) Lacking lexicography modernization theory. 2) It is difficult to form scale and achieve sharing because corpus development sacks a unified standard. 3) Low level of corpus processing and limited tagged language information cannot meet the demands of lexicography automation. 4) Few kinds of electronic dictionaries, small amount of information, lack of innovation and competitiveness. Accordingly, these dictionaries have no enough competitiveness. The causes for these problems might contain: 1) A lack of foresight consciousness, 2) inadequate project approval of scientific research, 3) lack of reason on scientific research direction, 4) imbalance on the main part of scientific research, and 5) low-level repetitive construction. Several promising strategies are as follows: 1) establishing a high level platform for national lexicography modernization with the improvement of top-class design and long-term investment; 2) establishing original theories suitable for Chinese lexicography by learning from advanced experience of foreign countries, 3) establishing effective mode of dictionary industry and creating boutique products in this era by adapting to the market demands and strengthening cooperation between universities and enterprises; 4) building environment for cultivating lexicography talents by changing the assessment mechanism; 5) holding frontier conference to highlight the issue of lexicography; 6) establishing lexicography major to offer related courses in the field of lexicography and cultivate advanced talents; 7) establishing national professional society and promoting the development of lexicography industry pragmatically and effectively.

Keywords: lexicography modernization, theories on lexicography, corpus, computer-aided lexicographical system, electronic dictionary

Lexicography modernization includes the modernization of lexicography theory, means (corpus, databases, computer-aided compiling system) and carrier (electronic dictionary). From the current status, with the beneficial exploration of Lexicography modernization, we have made some theoretical and practical achievements, including putting forward the concepts of "the Dream of Dictionary Power" and industrialization of dictionaries (Liu Hairun, Kang Shiyong, 2013), developing a batch of corpus (for example, the State Language Commission Corpus) and computer-aided compiling systems (Kang Shiyong et al., 2012), as well as some electronic dictionaries (for example, "Modern Chinese Grammar information Dictionary" of Peking University). But as a whole, compared with western developed countries, there is a big gap and some certain problems that are needed to be carefully solved.

1. Existing Problems

1.1. Lagging lexicography modernization theory
We are short of original and advanced lexicography theories. Taking Computational Lexicography Theory as an example, it is basically in the stage that only introduces lexicography modernization theory of modern western developed countries, which is not well digested, absorbed and used by ourselves, and even no innovation in the Theory. To introduce and create something new, western countries constantly absorb new theories from linguistic theory, language learning theory, cognitive science and other aspects, forming original lexicography modernization theory, which is used to guide the lexicography practice. For example, American WordNet, which has broad effects in the word, is set to use synset to represent lexical concept and describe lexical matrix, that is to say, establish a mapping
relation between the form and the meaning of a word. This design idea comes from "human vocabulary memory" theory of psychological linguistics, which attempts to abstract language regularity from all kinds of natural discourse to imitate the human brain to build the method of vocabulary memory (Zhang Yihua, 2004). However, FramNet is built on the basis of Case Grammar and Frame Semantics. Frame Semantics is a way of describing lexical meaning and grammatical structure. This theory is derived from such a hypothesis: in order to understand the meaning of a certain word in any language, we must understand its conceptual structure and semantic frame first, because these two provide the context that lexical meaning in the language and the methods used in discourse. Frame Semantics assumes that linguistic unit can select and highlight certain aspects or instances of semantic frame in a certain way through its locating language structure, therefore, lexical meanings and functions can be explained by describing the characteristics of lexical underlying semantic frame and these approaches; however, the frame is an intuitive idea that can formalize the relationship between semantic features and syntactic structure in the results of lexical analysis. Semantic frame expresses the context including a variety of predicates, supporting words and other conceptual roles in the form of schema. Each of the items above is frame elements and the semantic items of predicates correspond to the associated elements of the word (Zhang Yihua, 2004).

Corpus Pattern Analysis is a way that projects lexical semantic to a specific lexical text, aiming at building the pattern data of English verb for the research of computational linguistics, language teaching and language cognition. Its theoretical basis is "Theory of Norms Exploitations". This theory has absorbed the Generative Lexicon Theory and Generative Semantics Theory and other theories. Theory of Norms Exploitations says that the meaning of an isolated word is very vague, or it has no "meaning" and just "meaning potential". The information stored in human brain is not isolated words, but lexical pattern or phraseological pattern and characteristics of different categories that are closely related to the pattern. In order to understand the meaning of a word in verbal communication, the key is to know its pattern and distinguish the norms pattern and the exploitation pattern. To know how to get the meaning of a word, it is necessary to analyze the location distribution and the context of the words: valence and match (Zhang Yihua, 2012). We lack such original, forward-looking theory, and some good ideas can not receive any response for a long time. For example, inspired by WordNet, the article "Research on New Century Oriented Online Dictionary Compilation" published by Mr. Xu Shiyi in 2001 thinks that we should take advantages of network technology to develop and compile a network-like dictionary on the basis of human cognitive characteristics of vocabulary, weaving the concepts from human brains into a network, with each word on a node of the network to achieve connection between each other through a variety of relationships. Even now, we still do not have such a kind of language engineering conducted practically based on Chinese, but the Institute of Computational Linguistics from Peking University finished the Chinese version of WordNet, and formed the Chinese Concept Dictionary (CCD).

1.2. Difficulties to form scale and achieve sharing because corpus development lacks a unified standard

A series of dictionaries like international sample dictionaries of Oxford, Longman, Cambridge, and Collins were compiled on the basis of large scale and unified tagged sharable corpus, in which the representatives are the Brown Corpus established by Brown University in 1960s, LOB Corpus that was established by Lancaster University in Britain in 1970 and transferred to Norway in 1976, Birmingham Corpus (COBUILD) co-established by University of Birmingham in UK and Collins Publishers Ltd., British National Corpus (BNC) of Oxford University Press and so on. Under the guidance of a certain theory, the relatively influential corpus in western countries are tagged and developed according to the same criteria and corpus developed by different agencies in different periods can be compatible, coming together to form a large-scale tagged corpus, and achieve sharing. However, corpus developed by China seems a lot, in fact, each one is isolated. For so-called "innovation", each corpus has a set of development and tagging standard, which is not compatible, not sharable and results in the failure of forming a large-scale tagged corpus and low level repeating.

1.3. Low level of corpus processing and limited tagged language information
With the continuously improvement of western corpus’s processing level and increasingly plentiful linguistic information from part-of-speech tagging, collocation information to syntactic constituent, semantic role, semantic pattern and pragmatic environment, from soup to nuts, corpus datamation can be achieved finally, which has provided a basic guarantee for computer lexicography automation and made it possible for automatic dictionary generation. However, our corpus tagging stays in shallow level, and the most are raw corpus with no or little linguistic tagging information. Only participle and part-of-speech tagging are tagged in slightly large scale, syntactic constituent tagging, semantic role tagging, semantic pattern tagging, sense tagging and pragmatic information tagging are tagged a little. Since, different from English, Chinese is a language lack of morphologic change with no information tagging, corpus datamation cannot be achieved; thus Chinese lexicography automation will become an empty word.

1.4. Few kinds of electronic dictionaries, small amount of information, lack of innovation and competitiveness

There are many kinds of electronic dictionaries in western countries with rich information, some of which are almost synchronous with the paper dictionary and some have enlarged amount of information by the use of modern information technology, which has become an irreplaceable tool. It is more valuable that online lexical resources in true sense have emerged, such as WordNet, FramNet, Word Sketch, DANTE and Corpus Pattern Analysis, which are widely used all over the world. “Maybe WordNet is the most important and most widely used lexical resource for natural language processing system so far” (Zhang Yihua, 2004). But some of our limited electronic dictionaries are basically the simple version of related paper dictionaries on the network, some adding a little new information at most, with no innovation. So far, relatively influential electronic dictionaries in China are Modern Chinese Grammar information Dictionary and Chinese Concept Dictionary (CCD) developed by Institute of Computational Linguistics in Peking University and HOWNET, developed by Dong Zhendong and Dong Qiang. Although Web Dictionary and Baidupedia are influential, we can see that they are open systems where each person in different qualities can edit, and contents in fact are good and bad, thus the quality cannot be guaranteed. The whole domestic electronic dictionaries are lack of competitiveness.

2. Reasons of Existing Problems

2.1. A lack of foresight consciousness

On one hand, lack certain foresight consciousness and ideas, and ignore the tremendous effect caused by the combination of computer and lexicography in highly developed information age in the future, and lack recognition of the applicability and importance of the computational lexicography or computer language information processing technology in future lexicographical studies and lexicography compilation; on the other hand, government departments lack adequate attention and emphasis to the development of dictionaries, especially bilingual dictionaries. Therefore, we should actively call on the relevant government departments to pay proper attention to the research on lexicography modernization theory, and the research and development of dictionary corpus and databases, striving to make our career of lexicography modernization and dictionaries lead the world like Britain and the United States, and making the dictionary industry become the important backbone of national culture industry to obtain social benefits and economic benefits.

2.2. Inadequate project approval of scientific research

It can be reflected from the capacity of project approval of scientific research to some extent that whether a discipline and a professional field being taken seriously. Scientific research projects of lexicography and lexicography modernization have no place in National Natural Science Foundation of China (NSFC) and Social Science Planning Projects, as well as the "project guidance" of scientific research projects in provincial level, lack of independence because of walking in a fine line of other disciplines only. Therefore, we should call on national functional departments to add project approval for the study of computational lexicography moderately to Natural Science Funds Projects in national and provincial level and Social Science Planning Projects.
2.3. Lack of reason on scientific research direction

Nowadays, all kinds of dictionaries are not regarded as scientific achievements or considered in low scientific research coefficient during the assessment, promotion and other aspects by most of the universities and research departments, which directly results in insufficient enthusiasm and low motivation for the academic research of lexicography of a number of academic scholars, who are unwilling to participate in the work and research task in lexicography or lexicography field; it is for the love of the work that someone still makes an effort in this area reluctantly. A large number of researchers with considerable academic ability and quality are needed to participate in the research, because dictionaries should not only carry out objective description for natural language, but also make specification for the semantics, grammar and pragmatics of the language. In China, lexicographers were considered to be "saint", they lived a good life with high position and high income and were respected by people. It is still like this in western countries, but not like this in China. Someone even jokingly says that "let the condemned prisoner to compile the dictionary". Thus, domestic universities and research departments should change their research orientations, take the research achievements of different types of dictionaries and lexicography as an important chip and indicator to measure the level and ability of the researchers, to arouse their enthusiasm and give full play to the potentials in lexicography study, lexicography and other aspects.

2.4. Imbalance on the main part of scientific research

Britain firstly puts funds into all areas of lexicography modernization, especially into the construction practice of dictionary corpus, and then uses the corpus for machine translation research. Because machine translation is research work with pure exploratory nature, which is very challengeable and difficult to break in a short time, with large investment but maybe no economic benefits for a long time. However, under the circumstance of shortage of research funds, we put most of the limited funds into machine translation over a long period of time, instead of the construction, research and development of dictionary corpus and dictionary databases, which need less money with relatively quick effect and good economic benefits. In future, our functional departments and academic circle should focus on changing the imbalance, appropriately increasing funds support for the construction of lexicography research and dictionary corpus.

2.5. Low-level repetitive construction

In Britain, development projects of large commercial dictionary corpus are results by the cooperation between universities, scientific research institutions and presses. With collaborative division of labor, clear goals and enough funds, high quality development of corpus can be conducted continuously and effectively for they give full play to the initiative to produce great economic and social benefits. Currently, we are repeating low-level construction to a certain extent. In addition, lack of guidance and coordination by functional departments, communication and cooperation between departments and with so much re-handling of small workshop style, a lot of manpower and material resources are wasted and high quality research achievements and dictionaries are hard to be available. Therefore, large national language research institutions and excellent presses should make a medium-term and long-term plan for the development of lexicography modernization, and construct and develop some forward-looking and efficient projects like dictionary corpus, dictionary databases, computational lexicon, computer-aided lexicography system and dictionary automatic generation system by uniting with domestic dominant academic strength, With the formation of several influential corpus or dictionaries in signature style or fist style internationally, other researches can improve the efficiency by avoiding repetition of low level construction on the basis of them.

3. Improvement strategies

3.1. Establishing a high level platform for national lexicography modernization with the improvement of top-class design and long-term investment
Dictionary industry in Britain, United States, Japan, France and other countries has become an important part of their national GDP. Everybody knows that how much money the British, American and Japanese have made in the world, especially in China by English dictionary and electronic dictionary. With the rapid development of China, a new round of "Chinese fever" has been formed in the international community. However, "Chinese lexicography fever" has not developed with the "Chinese fever" springing up. We did not make money in the world by Chinese dictionaries like the British and American, on the contrary, many foreigners learn Chinese by the use of Chinese dictionaries compiled by their own country instead of ours. Why? The dictionaries compiled by us are unavailable for them to use. One of the important reasons is that we lack a platform for lexicography modernization and Chinese dictionaries that are suitable for learners from different countries on country-specific corpus. Therefore, we should recognize the importance of the establishment of national platform for lexicography modernization from the height of national cultural industry development, do top-class designs well, including corpus design and management, corpus tagging system, development of the computer-aided compiling system, development and publishing of electronic dictionary, funds investment and coordination of all departments. In order to establish cultural resources with an international influence, cooperation between universities, research institutions and publishers should be unified and coordinated by a competent department, and consider it as a normal work continuously and constantly.

3.2. Establishing original theories suitable for Chinese lexicography by learning from advanced experience of foreign countries
We all know that European and American dictionaries have accounted for the most market share of English dictionaries and related bilingual dictionaries worldwide. Accumulating nearly thirty years' successful experience on the development and utilization of dictionary corpus, dictionary databases, electronic dictionary, Web Dictionary, electronic lexicon, computer-aided lexicography system, they have formed a more stable development pattern with more advanced development ideas. However, the development and utilization of our academic research on computational lexicography, corpus, databases construction and practical electronic dictionary is still in exploratory stage, and lexicography method also stays in relatively traditional operational level. Therefore, it is a shortcut to actively learn certain theory research and practical experience from foreign countries. We should organize the professors of Chinese study, lexicography, language teaching, cognitive psychology, computer application and other fields to cooperate whole heartedly to establish original theories suitable for Chinese lexicography, guiding the practice of Chinese lexicography modernization on the basis of absorbing and digesting foreign advanced experience and theories.

3.3. Establishing effective mode of dictionary industry and creating boutique products by adapting to the market demands and strengthening cooperation between universities and enterprises
In Britain, with active cooperation, Collins Publishers Ltd. and University of Birmingham organized a research and development team with nearly a hundred people who mainly come from the University of Birmingham to build the famous COBUILD corpus. In addition, the famous Canadian “Bilingual Dictionary Project” also resulted from the cooperation of many universities and relative departments, which took more than a decade. Thus it can be seen that in modern lexicography circle, it is difficult to make influential modern achievements without seeking cooperation actively. At the moment, interdisciplinary research is the development direction of modern linguistics and lexicography, because many research tasks are difficult to complete only by lexicography or computer technology. Since universities, research institutions and publishing houses have their own advantages; complementary advantages should be formed by them. Therefore, to create boutique dictionaries in this era, we should understand the market demand, adapt to market needs, integrate multi-power, seek cooperation with enterprises, and establish an effective mode of dictionaries industries.

3.4. Building environment for cultivating lexicography talents by changing the assessment mechanism
Currently, a variety of assessment mechanisms have restricted the development of lexicography. On one hand, the assessment mechanisms of all levels of management departments should produce achievements as soon as possible, which determine all benefits, making most scholars seek quick success with low level repetition, instead of doing research quietly for original achievements. Thus, the research and development cycle of projects related to lexicography should be longer, which is helpful to achieve comprehensive results. On the other hand, cultivate talents with ability to do great business. This kind of talents should have noble professional and dedication spirits, and can endure the loneliness, because lexicography work needs to be done slowly and slow work yields fine products; they should have comprehensive knowledge and update knowledge constantly, which also needs time. This requires changing the assessment mechanism, not only counting the dictionaries as research achievements to address the worries of scholars, but also creating environment for scholarship and great career. Besides, periodical grade standard used for assessment and promotion has many disadvantages that everything is evaluated by SCI, SSCI, EI and CSSCI. But in fact, research articles in lexicography field are impossible to be found in SCI, SSCI and EI. Representing the highest level of domestic lexicography field, the journal *Lexicographical Studies* of Chinese Lexicography Society is excluded from CSSCI, because shortage of researchers in this field, the quoting rate of the articles in this journal is far behind that of journals of other disciplines, which has greatly defeated the enthusiasm of researchers. In response to this phenomenon, the national authorities should introduce appropriate incentives to mobilize the enthusiasm of researchers in lexicographical research field.

3.5. *Establishing lexicography major to offer related courses in the field of lexicography and cultivating advanced talents*

Lexicography modernization requires interdisciplinary talents, who should possess knowledge and skills in linguistics, lexicography, philosophy, history, science and technology, cognitive psychology, computer and information processing. It is so difficult for existing majors to cultivate such interdisciplinary talents. Therefore, lexicography major should be set to offer courses including the knowledge of theories and skills in related field. Even if lexicography major cannot be set, orientation under the related majors should be established to cultivate senior specialized personnel. Also relative training courses can be offered for personnel in this field to improve their professional knowledge.

3.6. *Holding frontier conference to highlight the issue of lexicography*

The basic task of academic conferences or forums is to concentrate on issues of particular fields. In the future, lexicography frontier conferences or forums should be held regularly by national lexicography society and computational lexicography circle, putting forward a frontier issue or the issue that is urgently in need of solution for discussing every conference to solve some practical problems. At the same time, each of the participants can walk in the forefront of international academic development by understanding the knowledge in computational lexicography field or domestic and international academic trends.

3.7. *Establishing national professional society and promoting the development of dictionary industry pragmatically and effectively*

Professional academic institution is one of the foundations for discipline development. It is gratifying that the leaders of Chinese Dictionary Society is keenly aware of the urgency of the situation, active promoting the establishment of Technology Professional Committee for Lexicography Modernization, advocating lexicographers in young and middle age to master computer skills and applied to the research, compilation and publishing of dictionaries. But so far as the status, the function and effectiveness of the professional committee is not great enough, national professional organization of "Lexicography Modernization Society" should be set up to give the personnel and funds support, coordinating the research, development and promotion of lexicography modernization field in an united and coordinating way to effectively promote the development of dictionary industry in China.

Contents above are our views on the problems, reasons and countermeasures of lexicography modernization in China. We welcome comments on these issues.
References
Hennie van der Vliet: Department of Language and Communication, Vrije Universiteit, Amsterdam (hd.vander.vliet@let.vu.nl) THE REFERENTIEBESTAND NEDERLANDS AS A MULTI-PURPOSE LEXICAL DATABASE
Optimizing Procedures for THE Making of Bilingual Dictionaries And The Concept Of Linking Contrastive Lexical Databases
白崇远.微软光盘版ENCARTA电子全科全书的特点及其利用.辞书研究,2000（04）
常爱东.中美电子百科全书对比分析.辞书研究,2001（06）
常爱东.电子辞书散论.辞书研究,2001（04）
徐时仪.面向新世纪的网络词典编纂刍探.《中国辞书论集2001》，陕西人民出版社，2002
赵铁军,李生.知识处理与电子词典的设计.辞书研究,1995（01）
张国强.关于工具书数字化发展趋势的几点思考,辞书研究,2001（03）
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）
章宜华,电子词典编纂的一般原则,辞书研究,2000（04）
章宜华,计算词典学与新型词典.上海辞书出版社,2004
章宜华,黄建华.电子词典的现状与发展趋势.辞书研究,1996（06）
章宜华.国际辞书现代化技术的新理念：辞书语料数据化.辞书研究,2012（02）
张绍麒.辞书与数字化研究.上海辞书出版社,2005
中国辞书学会学术委员会.中国辞书论集2001，陕西人民出版社，2002
中国辞书学会学术委员会.中国辞书论集2002，四川辞书出版社，2002
中国辞书学会学术委员会.中国辞书论集（第7辑），外语教学与研究出版社，2002
乐嘉民，亢世勇.辞书编纂现代化研究.上海辞书出版社，2009
亢世勇、王兴隆、谢晓燕.我国计算机辅助辞书编纂系统初步调查研究，辞书研究，2012（3）
刘海润、亢世勇.我国辞书编纂现代化理论研究调查分析，长江学术，2013（2）
Where Do We Go from Here?

Ari Kernerman
Kernerman Publishing, Tel Aviv, Israel
kernerman@zahav.net.il

Abstract
Fifty years of trying to prove to teachers why monolingual dictionaries are better language-learning tools for their students than any other kind of dictionary, and thirty years of trying to convince lexicographers that bilingualised dictionaries are lexicographically and pedagogically superior to any other type of language-learning dictionary, did not achieve their goals. During all this time, simple bilingual dictionaries went their merry way, overwhelmingly dominating both print and digital dictionaries, in sales and in use, without the need for any particular promotion.

There is still no general agreement on what information language-learning dictionaries should contain, or what should be the make-up of an entry. While monolingual English-learners' dictionaries were enjoying their heydays, basking in bliss, and semi-bilingual dictionaries were striving for recognition, simple bilingual dictionaries were (and still are) the language learner's choice. It seems that six different advanced learners' dictionaries, all written by the cream of the professional crop, incorporating every gimmick imaginable, such as corpus lexicography, were unable to win over the EFL market. Nor did semi-bilingual dictionaries succeed in doing so. What is the reason why language learners prefer a simple bilingual dictionary to a sophisticated monolingual or semi-bilingual learner's dictionary?

I will discuss these and other problems pertaining to the conflict between the goals of lexicographers and actual dictionary use.

Keywords: monolingual; bilingual; semi-bilingual; foreign-language-learning dictionaries; foreign-language-using dictionaries

We have here before us two questions. The first question is: "Where did we go wrong?" And the second question is: "Where do we go from here?"

I will try to answer the first question myself in the course of this presentation. I will solicit suggestions from the audience regarding the answer to the second question. But first we should define the word "here". Some people might not be sure what "here" means here.

"Here" means here a particular situation in which a small number of academics are still trying to do something that most dictionary users are not at all interested in. Here, is a situation in which thousands of dollars, maybe millions of dollars, are being spent over the years by many universities, in as many countries, on lexicographic R&D (Research and Development.) The particular R&D that our lexicographers have been involved in has been leading them in the wrong direction for fifty years, ever since lexicography became an academic subject.

Excuse me for throwing a bombshell, but let's start from the beginning. For hundreds of years there was no problem. People studied foreign languages, and even had word lists with translations, or lexicons, to help them understand the meanings of the foreign words they were trying to learn. Then, about 80 years ago, a handful of Englishmen went to Japan to teach English. They were innovative individuals – so innovative that because of them, many of you now have jobs.

These men said that teaching a language entirely by means of translation is bad. In that, they were right. They said that foreign languages should be taught entirely in the foreign language. In that, they were only partly right.
Teaching a language using only the language being taught they called "The Direct Method", and teaching a language as had been done until that time, by translating everything into the learner's mother tongue, they called "The Indirect Method".

Of course, if you're going to explain the meaning of a word in the language being taught, called "L2", the mother tongue being called "L1", then you have to explain that meaning using very simple language - if possible using only words and expressions that are already familiar to the user. Not only should the words be familiar, but the syntax and the grammar used should also be familiar, such that new words should be explained within their own linguistic environment.

So according to the Direct Method, a simple list of words with their translations was not adequate for language learners. Students needed a "proper" dictionary in which all the words in the L2 that they were studying were explained in L2, using only easy explanations, or even synonyms if possible. How to use these words was exemplified by means of example sentences or phrases, written in easily comprehensible language, and without employing difficult or unfamiliar grammar. Thus was born the learner's dictionary.

All this started with a handful of English teachers, who subsequently became dictionary compilers, or what we call today, "lexicographers".

Longman published The New Method English Dictionary, by Michael West and James Gareth Endicott, way back in 1935 – exactly 80 years ago. The explanations were written in a defining vocabulary, but there were not yet examples of usage. E.C. Parnell wrote English-Reader's Dictionary. And in 1948 Oxford University Press (OUP) published the first version of The Advanced Learner's Dictionary of Current English, written by A.S. Hornby, E.V. Gatenby and H. Wakefield. Other dictionaries that were published around that time were Idiomatic and Syntactic English Dictionary, The Progressive English Dictionary, and An English Reader's Dictionary, all written by pioneers in the field such as Palmer.

Until now all that I have been saying is common knowledge. But knowing about something is one thing; realizing its full significance is another.

At that time, air travel and the rapid development of communications were making the world smaller. There was need for a global language to connect people throughout the whole world. A war was raging – not between countries but between languages - a war between English and French. Which one would become the dominant language? English won out, as you all know. It is not important now why and how, but the English language won the battle, and became the universal lingua franca. Since then, people all over the world have been clamoring to learn English. English Language Teaching & Learning became a big business. It became one of the UK's biggest exports – worth billions of pounds. And just as it is today, it was money that made the world go round. Some publishers saw a golden opportunity to make a lot of money, so they jumped on the bandwagon to publish English learners' dictionaries. ELT dictionary publishing became a big business, with publishers competing frantically among themselves. All of the dictionaries had the same target audience, and more or less the same headwords, so the publishers competed on such features as the general appearance of the dictionary, the make-up of the entries, and whatever gimmicks they could devise.

The pioneers were Longman and OUP. OUP's flagship dictionary, now called Advanced Learner's Dictionary, was very accurate, and had the advantage of emanating from a prestigious university. The dictionary was well promoted, so it enjoyed great success.
Longman's flagship was Longman Dictionary of Contemporary English. Longman were expert at marketing, and were able to gain the approval of the ministries of education in many countries, so that their dictionaries became mandatory for students to purchase. So the Longman name too became well known.

HarperCollins Publishers came out in 1987 with a novel idea for their English learners' dictionary. They innovated the concept of basing their dictionary entirely on a word corpus, called Bank of English, with Collins COBUILD Dictionary (now called Collins COBUILD Advanced Learner's English Dictionary). It was a hard fight, competing against such prestigious publishers as Oxford and Longman. But HarperCollins succeeded in instilling into the minds of lexicographers the importance of using word corpora in selecting example sentences and phrases, so Oxford and Longman had no alternative but to eventually base their English learners' dictionaries also on corpora.

English teachers and entire education systems were brainwashed into believing that corpora-based dictionaries were superior as language-teaching tools to non-corpora-based dictionaries, so the concept of word banks was able to make considerable headway all over the world, achieving a very respectable status, not only in the area of lexicography but in linguistics in general.

It was difficult now for other publishers to enter the ELT dictionary market. Cambridge University Press (CUP), who were prestigious monolingual dictionary publishers, joined the show in 1995 with Cambridge International Dictionary of English, now called Cambridge Advanced Learner's Dictionary. It is also a fine dictionary, but CUP found it extremely hard to compete.

Seven years later, in 2002, another prominent and highly respected publisher, Macmillan, joined in with their version of virtually the same product, Macmillan English Dictionary for Advanced Students. They were late in joining the club, and, I believe, had an even harder time than CUP.

Finally, an American publisher did their thing, with Merriam-Webster's Advanced Learner's English Dictionary, in 2008, sadly late.

At the same time, a very small, insignificant publisher, in another part of the world, started to publish a different sort of learner's dictionary – the semi-bilingual dictionary. The first ones appeared in 1986 – almost 30 years ago – for speakers of Hebrew and for speakers of Arabic, and within a decade local language editions appeared in 34 countries. This novel type of learners' dictionary consisted of entries that had the same basic elements as in monolingual learners' dictionaries, viz., headword, part-of-speech, definition in simple language, examples of usage, and, if deemed useful, additional information such as spelling tips, or warnings about misuse.

The semi-bilingual method realized that the Direct Method in foreign language teaching had both advantages and disadvantages. In general, the semi-bilingual approach combined the positive features of both the Direct Method and the Indirect Method, while avoiding the drawbacks of both.

But it is difficult during the course of a lesson to teach only in L2, to explain, to penetrate the minds of students, to know whether or not they really understood what was said, or maybe even misunderstood it, and to verify comprehension, during the course of a lesson.

Besides, the truth of the matter is that learners want to know their own L1 equivalent, in order to be certain that they have correctly understood the meaning of that word, and this is what semi-bilingual dictionaries provided.

I will not spend more time singing the praises of semi-bilingual dictionaries. You all know what they are, and all of you have your own opinions about them.
So you had available on the market numerous excellent monolingual learners' dictionaries, as well as, in many countries, semi-bilingual learners' dictionaries.
But the language learners wanted neither. All they wanted was simple dictionaries, even if they were just pocket dictionaries, so long as they gave them the translation. They were not interested in getting a wealth of information. True, many teachers recommended monolinguals or semi-bilinguals. But when it was left to them, students demonstrated their preference very explicitly when purchasing their dictionary. Their preference has not changed during all this time. Students are in a hurry, and there is no faster route to finding the meaning of a word than that of using a bilingual dictionary – whether printed or digital.
Yet, we continue to spend all our time and money trying to improve dictionaries that the masses do not want! Vast sums of money, as well as whole lifetimes of lexicographers, have been invested in monolingual and semi-bilingual learners' dictionaries, both of which are rejected by most language learners. Why don't we spend more time trying to improve bilingual dictionaries? Why is this ASIALEX conference concerned more with satisfying lexicographers than it is with satisfying dictionary users?
Besides, who says that all dictionaries that are in a foreign language are used by people who are learning that language? What about those dictionary users who are not learning the language but are simply using it? All they want is the very same thing as what the language learners want – THE SHORTEST ROUTE TO UNDERSTANDING THE MEANING.
Ladies and gentlemen, I began by posing two questions: Where did we go wrong? and, Where do we go from here?
I tried to answer the first question. I don't know the answer to the second question. So I turn to my audience. Ladies and gentlemen, where do we go from here?
Lexicography as a Teaching Tool: A Hong Kong case study

James Lambert
National Institute of Education (NIE), Singapore
james_lambo@hotmail.com

Abstract
Dictionaries have been put to a number of pedagogical uses, including lessons on how to use dictionaries (e.g. Chi 1998), and lessons utilizing pre-existing dictionaries for a range of educational goals (e.g. vocabulary building). However, less common is the use of the practice of lexicography as a teaching tool. This paper details an example of using lexicography in the teaching of World Englishes. In March of 2014, three classes of students studying World Englishes at City University of Hong Kong were required to research and draft two dictionary entries on items of Hong Kong English. This lexicographical assignment had a number of educational benefits. Specifically, the students gained hands-on experience of dictionary writing, which provided insight into the difficulties faced by lexicographers, and hence clearer understanding of the reliability of dictionaries. The project also helped students hone their sense of the overlap between Hong Kong English and Cantonese. Finally, the entries written by the students document over 60 items of Hong Kong English that have never been treated in other lexicographical texts on the variety (such as Cummings & Wolf 2011), highlighting the need for more work in this area. Similar projects could be utilized for practically any variety of English as part of World Englishes studies.

Keywords: lexicography; pedagogy; World Englishes; Hong Kong English

Introduction
In the field of metalexicography there seems to be very little literature on the teaching of lexicography. Pedagogical interest in lexicography seems to be largely concerned with how students use dictionaries (e.g. Al-Salami & Hattab 2008), the teaching of dictionary skills in the classroom (e.g. Chi 1998), or the use of existing dictionaries to improve vocabulary or achieve other educational goals (e.g. Gonzales 1999). Lexicography itself – that is, the art and craft of dictionary writing – is a subject that has been taught at tertiary level for some decades in various universities worldwide. Examples include lexicographer David Blair who taught courses on lexicography at Macquarie University, Sydney from 1987 until 2001, and Professor Julie Coleman, at the University of Leicester, who regularly has a module in which students create their own subject-restricted lexicons which are then published online: samples can be viewed via the university website (Coleman). What seems to be scarce, however, are scholarly articles on how such courses are run and the theoretical issues involved. One of the few articles to discuss educational benefits of teaching students to be lexicographers is Battenburg and Lant (2003). They note that such projects required students to apply knowledge from a range of linguistics subject areas (14), and also argue that lexicographical projects are a type of ‘situated learning’ in which abstractions are ‘grounded’ (8) since students have to face the same questions that lexicographers have to face (13). In response to the dearth of material on using lexicography as a teaching tool, this paper provides details of a lexicographical assignment conducted at City University Hong Kong in 2014.

Assignment design
As part of a course on World Englishes run in 2014 by the Department of English at City University of Hong Kong, students were set a lexicography assignment by the course coordinator, Associate Professor Chris Jenks. There were three tutorial groups totaling 60 students in all. Students were English majors in their third year, mostly but not exclusively of Cantonese background. The assignment description was brief: ‘Students are responsible for two dictionary entries of Hong Kong English.’ The primary goal of this assignment was not to teach lexicography per se, but rather to engage the students in a practical research project that would link to the course aims and objectives, namely: to raise the students’ consciousness of
the socio-political, cultural, economic, and ethical aspects of the global spread of English; to equip students with critical concepts needed to examine the status, functions, and features of world Englishes; and to allow the students to discover some aspect of the role of English in Hong Kong, with particular reference to the description of Hong Kong English (adapted from Jenks 2014).

As co-teacher, I devised a range of teaching materials to supplement the assignment description. A detailed example of the type of dictionary entry required was supplied as a handout (Appendix 1), along with a description of the required lexicographical fields (Appendix 2), and a table of etymological types (Appendix 3). The sample entry (Appendix 1) was based on the *Oxford English Dictionary* model, that is, the dictionary on historical principles (Murray et al. 1888-1928). The sample entry consisted of a headword, pronunciation in the International Phonetic Alphabet, part-of-speech label, definition, several citations, an etymology, and usage notes. Student assignments had to follow this basic entry structure, with a minimum of two citations (from different sources, and preferably some years apart). Students were also encouraged to find the earliest citation possible. If the etymology included Cantonese etymons, the Chinese characters had to be given followed by a transliteration in parentheses. Cantonese has no official Romanization system, so for consistency sake, the Yale Romanization system was stipulated for all student dictionary entries, and copies of Chik and Ng (2000) were made available to students in tutorials. For etymons in Mandarin, pinyin was stipulated. If required, transliterations for other non-English scripts could be negotiated individually with the teacher.

**Teaching lexicography**

Students were essentially given a ‘crash course’ in how to be a lexicographer over a number of tutorial classes, beginning with lexicographical data collection. As a starting point, students were prompted to discuss how lexicographers go about creating dictionaries. Considering that no lexicographer can conceivably know all the words of a language, students were asked to ponder what data lexicographers used to inform them about words, leading to discussion of various aspects of typical lexicographical research such as reading programs, citation collection, and corpus linguistics. To illustrate how word meanings could be inferred from context, students were given citations of an unfamiliar word from which they had to construct the meaning. The word chosen was the Australian slang term *chook* which means ‘chicken.’ They were presented with the following citation first.

[chook n.

‘If you don’t want ’em, you can sling ’em to the chooks.’

1932 ‘Miles Franklin’ *Old Blastus of Bandicoot* v. 59]

From this, students could infer that the word *chook* was a noun and countable, though little else as could be reasonably inferred as the text does not say what is being slung to the chooks. From the next citation, the students were able to guess the meaning ‘chicken,’ as the chooks are in a *hen-house*.

[chook n.

‘Good garden, Archie, and a big hen-house full of chooks.’

1966 Bruce Beaver *You Can’t Come Back* 111]

More citations of the target word were supplied, from which the meaning became abundantly clear. This exercise gave students an insight into the way in which the meaning of words is linked to context, how lexicographers work with unfamiliar lexis, and the informational sources that lexicographers use to construct dictionaries. Also, it gave the students insight into the type of citational material they were required to collect for their assignment. Sample entries from the *Oxford English Dictionary* (Murray et al. 1888-1928) were discussed in class, focusing on the functions of displaying the citational evidence in a dictionary entry.
In terms of etymologies, apart from the use of transliteration, students were expected to categorize terms according to the type of word formation (e.g. agglutination, compounding, clipping) or word transmission (e.g. borrowing, loan translation/calque). The various processes of lexical generation were discussed in class, with group-work activities in which students had to generate their own examples. To further assist students when doing their assignments, a handout of various word formation categories was provided (Appendix 3). Finally, a certain amount of class time was devoted to assisting students with the selection of Hong Kong English lexical items for their assignments. Each student had to select two lexical items and register them with the teacher on a first-come-first-served basis. Students were free to choose whatever lexical items they liked so long as they satisfied two stipulations: (a) the lexical items had to be Hong Kong English, so far as the student and the teacher were aware, and (b) each lexical item had to be unique from those of the other students. It was anticipated that this would generate 120 separate dictionary entries (2 for each of the 60 students). In order to assist students in the selection process there was a discussion of the various semantic fields and subject areas that are likely to contain localized lexis (e.g. local types of food, local culture, flora and fauna, transport infrastructure, educational and political terminology, slang, etc.). While many students furnished their own lexical items, I also provided a list of localized lexis, based on my own Hong Kong English citation database, from which they could select items if they were having trouble finding one themselves.

Results of the assignment
A total of 113 separate items of Hong Kong English were lexicographically described among the dictionary entries drafted by the students (see Appendix 4). This is slightly lower than the projected 120 possible entries (2 for each of 60 students). The shortfall was partially due to overlap, despite stipulation (b) mentioned above, and partially a result of failure to submit the assignment. These were compared against the Hong Kong English lexis in three pre-existing lexicographical treatments of Hong Kong English: the Grolier International Dictionary (2000), Bolton (2003), and Cummings and Wolf (2011). A total of 92 lexical items were not found in the Grolier, 92 were not in Bolton, and 73 were not in Cummings and Wolf. An impressive total of 67 terms were recorded afresh by the students, having appeared in none of the previous lexicographical treatments of Hong Kong English, revealing that there is much work yet to be done on the lexis of Hong Kong English.

In terms of etymologies, by far the largest source language was, understandably, Cantonese. Students detailed 37 direct borrowings, 37 loan translations/calques, and 6 partial loan translations. The next most frequent source language was English, with neologistic compounds accounting for 15 etymologies, with 9 other English-based terms from a variety of word formation processes such as initialism and semantic extension. Mandarin was not a prominent source of etymons with only 2 borrowings and 2 loan translations. Other terms were of mixed origin, some with complicated or uncertain etymologies. The term mahjong is neither Cantonese nor Mandarin, and has not been sufficiently accounted for yet in other etymological sources. The term missy appears to be a Chinese Pidgin English term from English miss + the Cantonese phonotactic suffix -eel/-y, though perhaps owes something to the pre-existing British English diminutive of miss.

The student contributions were of variable quality, though generally very well done. Some in-class tutorial time was set aside for workshopping the entries. The poorer assignments were turned in by students (known locally as ‘deadline fighters’) who decided to forgo this opportunity and draft their final dictionary entries the night before the due date. With the students who worked on drafts in class, a number of aspects of lexicography proved challenging and counterintuitive. With regard to definition-writing students, as a result of having spent years padding out written assignments in order to meet word-length requirements, were somewhat bemused to be told that fewer words were better than a long, wordy definition. Occasional errors occurred with grammatical category labelling, categorizing etymons, and IPA transcription. Some etymons were not transliterated with the Yale system, and some entries overly relied on information from Wikipedia. Nevertheless, the high quality of the resulting entries meant that students were able to compare their own work
favorably with that of Cummings and Wolf (2011), a published academic text from the most prestigious university in Hong Kong, which contains less detailed entries and lacks transliterations of Chinese etymons.

Problems encountered
Some of the Chinese students had difficulty comprehending the need to Romanize the Chinese etymons. Living in a Chinese script environment they naturally could read the pronunciations of Chinese characters. And although the Hong Kong linguistic landscape has Romanized Chinese, for example in business names, bilingual Chinese students had expressed to me on many occasions that they generally ignore English text, whether transliteration or translation. To put it another way, they are largely blinkered to the English texts in their own linguistic environment. In order to dramatize the need for Romanization for readers with no knowledge of Chinese, students were provided with the following etymology:

From Hindi बीस्ती ‘water carrier,’ from Persian بهشت ‘heavenly.’

Without transliterations, the Hindi and Persian scripts had no meaning to the students. When transliterations were subsequently supplied (Hindi bhishē / Persian bihīshē), the students felt that the etymologies were more readable, and were able to identify better with someone (i.e. a potential reader of their own dictionary entries) who could not read Chinese.

By far the most challenging task for students was how to determine if a word was Cantonese or English. As is typical in language contact situations, although a great deal of lexis is clearly the preserve of one or other of the languages involved, there is a certain amount of lexis that exists in a gray area between the two. Take for instance the terms *bon voyage* and *gesundheit*, which although used in English for many years still retain French and German spellings and pronunciations. Are they truly English words? In the Hong Kong context there are a great many such terms. First, there are borrowings. These include English words borrowed into Cantonese, such as English *taxi* > Cantonese 的士 (dī-sī) (see Chan and Kwok 1982), and Cantonese words borrowed into English, such as Cantonese 點心 (dīm sām) > English *dim sum*. There are also loan translations, such as English *add oil*, an encouragement, from Cantonese 加油 (gà yàuh), from 加 (gà) *add* and 油 (yàuh) *oil*, literally meaning ‘put fuel into your vehicle,’ but figuratively ‘to persevere and work hard.’ There are also cases involving bilingual puns, for example, the expression *delay no more*, which is used punningly to express the Cantonese phrase 屁你老母 (diū nēi hōi mǒuh), literally, ‘fuck your old mother.’ In Cantonese, initial *n* and *l* are frequently interchanged (Campbell and Wong 2015: 36), so that this expression is also pronounced diū léi hōi mǒuh. Importantly, the expression *delay no more*, written in the Roman alphabet, is commonly used in both Cantonese-language contexts and English-language contexts.

Another difficulty faced by the students was in determining if a certain lexical item was part of wider English or restricted to Hong Kong. Student opinions on this matter were by and large correct, however, with a number of expressions their initial impressions were incorrect. For instance *flea market* was thought to be a local Hong Kong term on the basis that the Cantonese term for *flea market* is 購物場 (tiu jōu sīh chëuhng), literally 購物 (tiu jōu) *flea* + 市場 (sīh chëuhng) *market*. However, the Cantonese is a calque on the English term, which dates back to the 1890s (*Warwick Examiner* 1891: 2), not the other way around. Meanwhile, *pawn shop* dates back to 1735 (Pope: 128), before the English language had come to Hong Kong. A similar misapprehension was made with *egg tart*, which was believed to have originated in Hong Kong and Macau. However, the term *egg tart* dates back only to 1975 in Hong Kong (Sung: 18), and occurs much earlier in Western sources (e.g. *Sunday Times* 1914: 31).

Within this melee of intersections between English and Cantonese, the students, being themselves bilingually fluent, were able to navigate with perfect ease in communicative contexts where the provenance of a certain term or expression matters little. Yet, they encountered difficulty when it came to taking a linguistic stance and deciding if a term was (Hong Kong) English or not – as the assignment required them to do. In order to provide the
students with a rule of thumb they could use to decide if a term was English, irrespective of its origins, at least two citations in which the lexical item was used in a wholly English context needed to be found. For instance, in the text ‘Add oil, the road ahead is long’ (Dotland.net blog), the expression add oil is used in an English-only context, and thus was acceptable for the assignment. Whereas the following example of the term, ‘Add Oil 加勵發放正能量’ (Oriental Daily), is in a wholly Cantonese context and thus represents a borrowing of English into Cantonese. As work on the assignment progressed, students found that many of the terms they first selected proved unlocatable in English-language contexts, and were only found in Cantonese contexts, which was somewhat of a revelation to them. Such terms had to be discarded and other terms selected.

Finally, students were generally poor at finding early citations, with 70 of the 113 terms (61.9%) being antedated by my own records, usually by many decades (see columns Date 1 and Date 2 in Appendix 4), in large part due to the lack of easily accessible resources for such research.

Discussion
The primary goal of the lexicographical assignment described was to enhance the students’ understanding of various concepts in the World Englishes course they were taking. Obvious educational benefits were that students were given a clearer sense, through practical experience, of the intersection between Hong Kong English and Cantonese. Additionally, the hands-on experience of dictionary writing provided insight into the difficulties faced by lexicographers and hence a strong position from which a more critical view of dictionaries could be taken. The student entries also documented over 60 items of Hong Kong English that have never been treated in other lexicographical works on the variety, highlighting the need for more work in this area.

As this assignment was only one of three in the course, it was necessarily brief. One obvious way in which the assignment could have been enhanced would have been to publish the finalized dictionary entries, either online or as a booklet (the word count for total entries was approximately 20,000 words). Since dictionaries are so widely regarded as having the power to standardize and legitimize lexis, lexicographical publication of the student entries has the potential to bring into focus questions about the validity of the so-called ‘new varieties of English’ and thus feed into the wider discussion that is of core concern to World English studies. Although this assignment focused on Hong Kong English, similar lexicographical projects could be utilized for any variety of English as part of World Englishes studies. In addition, the use of lexicography as a teaching tool offers abundant possibilities for a vast range of linguistic and other language-related studies.

References
Chan, Mimi and Helen Kwok. 1982. A Study of Lexical Borrowing from English in Hong Kong Chinese. Hong Kong: Hong Kong University Press.

29 ‘Add Oil to encourage the payment of positive energy.’


Jenks, Chris. 2014. *EN2323 Syllabus*. Hong Kong: City University of Hong Kong.


The Sunday Times. 1914. 10 May. Perth, Australia.


The Warwick Examiner and Times. 1891. 12 Sept. Queensland, Australia.

### Appendix 1 Sample dictionary entry handout

**ASSIGNMENT 1 – SAMPLE DICTIONARY ENTRY**

fortune stick / ˈfɔrtʃuːn ˈstk/ noun (count) one of a set of numbered bamboo slips which corresponds to a message concerning the future, drawn by a supplicant in Taoist divination


1904 Jessie Juliet Knox *Little Almond Blossoms: A Book of Chinese Stories for Children* 142. Ho chan thought that while they were there, they had better try throwing the fortune sticks, as he knew it would please them, so he picked up a tall round box, full of bamboo sticks, and explained it to them.

1938 Thomas Handforth *Mei Li* [unpaginated]. Mei Li ran happily down the hill. Surely, if the fortune sticks said she was going to rule a kingdom she would.

1944 *Contemporary Chinese Stories* 134. Finally mother returned, and the medicine prescribed by the fortune sticks was soon brewing over the stove.

1977 Paul Bowles *Things Gone and Things Still Here* 37. After the painful heat outside, everything in the temple suddenly seemed a symbol of the concept of coolness — the stone floor under my bare feet, the breeze that moved through the shadowy interior, the bamboo fortune sticks being rattled in their long box by those praying at the altar[.]

1992 *Islands: An International Magazine* Mar/Apr 120. Adding to the babble at Wong Tai Sin was the incessant rattling of fortune sticks, one of the most popular methods of divination.

2000 Mimi Chan *All the King’s Women: The Story of a Hong Kong Family* 32. Ah Ching went to the temple on her behalf on every major festival to kou chim, that is to ask one’s fortune through the use of fortune sticks.
Origin: Compound noun formed with English elements. Compare Cantonese 求 (kàuh) divination using fortune sticks, from 求 (kàuh) pray for + 签 (chim) slip of bamboo.

Usage: Not commonly used in spoken English, though not uncommon in newspapers and travel books.

Appendix 2  Format of dictionary entries handout

ASSIGNMENT 1 – FORMAT OF DICTIONARY ENTRIES

headword – in bold
pronunciation – in IPA (International Phonetic Alphabet)
part of speech – traditional grammar
definition – normally one sentence
citations (you need only 2; from different sources, some years apart)
etymology (see below)
usage (see below)

citation
  • year – in bold
  • author’s name (if present)
  • book/magazine/newspaper/website title
  • date (if relevant)
  • page number (if present)
etymology
  • type of word formation (see below)
  • language of origin
  • if from Chinese:
    o Chinese characters
    o Yale Romanisation in parentheses, in italic
    o meaning, and literal meaning
usage
  • make some notes on
  o who uses the word
  o what types of sources it is found in
  • what frequency it has (see sample entry)

Appendix 3 Word formation handout

ASSIGNMENT 1 – TYPES OF WORD FORMATION

1. loanword: a word borrowed from another language (e.g. cliché, from French)
2. agglutination: adding affixes (e.g. unhelpfulness, from un + help + ful + ness)
3. compound: adding to words together (e.g. shoelace, takeaway)
4. hybridised compound: compounding different languages (e.g. iceberg)
5. calque: a literal translation of word or phrase from another language (e.g. Milky Way, from Latin via lactea)
6. semantic extension: adding a new meaning to a word (e.g. gay ‘homosexual’)
7. grammatical extension: changing the part of speech to give new meaning (e.g. to film)
8. eponymy: using a proper noun: (e.g. xerox, Australian, Confucianism)
9. blending: mixing two words together (e.g. smog, from smoke + fog)
10. initialism: from initial letters of a phrase (e.g. USA from United States of America)
11. acronym: an initialism that makes a pronounceable word (e.g. pin from personal identification number)
12. **clipping:** cutting off part of a word (e.g. *ad* from *advertisement*)
13. **onomatopoea:** copying a real world sound (e.g. *boom, pop*)
14. **back-formation:** removing ‘seeming’ affixes from existing words (e.g. *edit* from *editor*)

### Appendix 4  Hong Kong English lexis dictionary entries

The following table lists the headwords of the dictionary entries submitted by students. **Date 1** is the earliest citation provided by students, while **Date 2** is an antedating based on the earliest citational evidence in the author’s own database (based on citations in Hong Kong English sources or in a Hong Kong English context). Column **G** refers to the *Grovier International Dictionary* (2000), column **B** refers to Bolton (2003), and column **C&W** refers to Cummings and Wolf (2011): an asterisk (*) indicates that the term is present in that source (totals at column bottom). Unless otherwise stated, calques are English loan translations from Cantonese.

<table>
<thead>
<tr>
<th>Term</th>
<th>Date 1</th>
<th>Date 2</th>
<th>Etymology</th>
<th>G</th>
<th>B</th>
<th>C&amp;W</th>
</tr>
</thead>
<tbody>
<tr>
<td>add oil</td>
<td>2011</td>
<td>2007</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>add value machine</td>
<td>2012</td>
<td>2000</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ah cha</td>
<td>2002</td>
<td>1958</td>
<td>Cantonese + Hindi</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albert Yip</td>
<td>2010</td>
<td>-</td>
<td>mistranslation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>amah</td>
<td>1997</td>
<td>1903</td>
<td>English &lt; Portuguese</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Amah Rock</td>
<td>1974</td>
<td>1928</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bamboo theatre</td>
<td>2004</td>
<td>1909</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bird nest soup</td>
<td>2008</td>
<td>1743</td>
<td>English</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>black face</td>
<td>2008</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>blow water</td>
<td>2009</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>body check</td>
<td>2001</td>
<td>1979</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>body itchy</td>
<td>2007</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>brainwash</td>
<td>1958</td>
<td>-</td>
<td>calque (Mandarin)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bubble tea</td>
<td>2010</td>
<td>1993</td>
<td>calque (Taiwanese)</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bun mui</td>
<td>2001</td>
<td>1990</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bun tower</td>
<td>2005</td>
<td>1962</td>
<td>part calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bus uncle</td>
<td>2006</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cage home</td>
<td>2008</td>
<td>1995</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Canto-pop</td>
<td>2005</td>
<td>1989</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cha chaan teng</td>
<td>2011</td>
<td>2002</td>
<td>Cantonese</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>char sui bao</td>
<td>2010</td>
<td>1968</td>
<td>Cantonese</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>chiu</td>
<td>2011</td>
<td>-</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chur</td>
<td>2013</td>
<td>-</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coffin home</td>
<td>2011</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>collect skin</td>
<td>2011</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>colour wolf</td>
<td>2007</td>
<td>1990</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>columbarium</td>
<td>2010</td>
<td>1958</td>
<td>English &lt; Latin</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dai pai dong</td>
<td>1971</td>
<td>1967</td>
<td>Cantonese</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>delay no more</td>
<td>2006</td>
<td>1976</td>
<td>bilingual pun</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dim sum</td>
<td>2001</td>
<td>1952</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ding ding</td>
<td>2002</td>
<td>1992</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>doubly non-permanent resident child</td>
<td>2012</td>
<td>-</td>
<td>part calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dragon boat</td>
<td>2002</td>
<td>1904</td>
<td>calque</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>dragon’s beard candy</td>
<td>1991</td>
<td>1975</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPS</td>
<td>2006</td>
<td>1985</td>
<td>English, initialism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>face green green</td>
<td>2010</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>face trade</td>
<td>2009</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fat choi ho si</td>
<td>1998</td>
<td>-</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>flea market</td>
<td>2012</td>
<td>1974</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>food paradise</td>
<td>2003</td>
<td>1974</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>free size</td>
<td>2013</td>
<td>1973</td>
<td>Japanese &lt; English</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fruit money</td>
<td>2012</td>
<td>1992</td>
<td>calque</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Term</td>
<td>Year</td>
<td>Date</td>
<td>Language</td>
<td>Type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>------</td>
<td>-------</td>
<td>--------------</td>
<td>---------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fung shui</td>
<td>2003</td>
<td>1797</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gai daan jai/tsai</td>
<td>2002</td>
<td>-</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gei wai</td>
<td>1997</td>
<td>1988</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>goddess</td>
<td>2011</td>
<td>-</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>green hat</td>
<td>2005</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gawai mui</td>
<td>2006</td>
<td>-</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gweilo</td>
<td>1998</td>
<td>1972</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hak Se Wai</td>
<td>2005</td>
<td>-</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hand letter</td>
<td>2008</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Holland bank cheque</td>
<td>2004</td>
<td>-</td>
<td>bilingual pun</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hong Kong foot</td>
<td>1994</td>
<td>1926</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hong Kong kid</td>
<td>2010</td>
<td>-</td>
<td>part calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hot dog</td>
<td>2005</td>
<td>-</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jetso</td>
<td>2006</td>
<td>-</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jockey challenge</td>
<td>2008</td>
<td>1979</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kai lan</td>
<td>1990</td>
<td>1939</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kowtow</td>
<td>1899</td>
<td>1804</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lang mo</td>
<td>2009</td>
<td>-</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>leftover woman</td>
<td>2009</td>
<td>-</td>
<td>calque (Mandarin)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>locust</td>
<td>2012</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lose face</td>
<td>2000</td>
<td>1834</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mafoo</td>
<td>1977</td>
<td>1863</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mahjong</td>
<td>1987</td>
<td>1921</td>
<td>Chinese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>maid room</td>
<td>2004</td>
<td>1965</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mainlander</td>
<td>1998</td>
<td>1906</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>milk tea</td>
<td>2011</td>
<td>1966</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>missy</td>
<td>2003</td>
<td>1848</td>
<td>English/Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MK</td>
<td>2007</td>
<td>2006</td>
<td>English, initialism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mo lei tau</td>
<td>2004</td>
<td>1990</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monkey King</td>
<td>1996</td>
<td>1951</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>monster parent</td>
<td>2010</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>moon cake</td>
<td>2001</td>
<td>1905</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no money no talk</td>
<td>2000</td>
<td>1981</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Octopus</td>
<td>1997</td>
<td>-</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>one-woman brothel</td>
<td>2009</td>
<td>1980</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>parallel trader</td>
<td>2009</td>
<td>1977</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>people mountain people</td>
<td>1998</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>poon choi</td>
<td>1997</td>
<td>1994</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pork chop</td>
<td>2005</td>
<td>2003</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese egg tart</td>
<td>2006</td>
<td>1994</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>princess syndrome</td>
<td>2009</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pseudo-model</td>
<td>2009</td>
<td>-</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>public light bus</td>
<td>2000</td>
<td>1969</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>San He Hui</td>
<td>1976</td>
<td>-</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sandwich class</td>
<td>1997</td>
<td>1993</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shahe-fen</td>
<td>1997</td>
<td>-</td>
<td>Mandarin</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shark fin soup</td>
<td>1989</td>
<td>1924</td>
<td>part calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sifu</td>
<td>2000</td>
<td>1972</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>siu mai</td>
<td>2011</td>
<td>1957</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>siu yeh</td>
<td>2006</td>
<td>1981</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>small house</td>
<td>1999</td>
<td>1972</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>snakehead</td>
<td>2007</td>
<td>1963</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>soaked in salt water</td>
<td>2012</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>street sleeper</td>
<td>1998</td>
<td>1919</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>subdivided flat</td>
<td>2011</td>
<td>1990</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>taipan</td>
<td>1956</td>
<td>1834</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taobao</td>
<td>2006</td>
<td>-</td>
<td>Mandarin</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tong lau</td>
<td>2010</td>
<td>2003</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tong sui</td>
<td>2009</td>
<td>-</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tongzhi</td>
<td>2000</td>
<td>-</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>topping off</td>
<td>2011</td>
<td>1962</td>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>villain hitting</td>
<td>2013</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wet market</td>
<td>2011</td>
<td>1983</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wide gate</td>
<td>1994</td>
<td>-</td>
<td>part calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wife cake</td>
<td>2005</td>
<td>1984</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wing chun</td>
<td>1972</td>
<td>1969</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Year 1</td>
<td>Year 2</td>
<td>Language</td>
<td>Note</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>--------</td>
<td>--------</td>
<td>----------</td>
<td>------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wonton</td>
<td>1981</td>
<td>1969</td>
<td>Cantonese</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XO sauce</td>
<td>1991</td>
<td>-</td>
<td>English</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yellow joke</td>
<td>2006</td>
<td>-</td>
<td>calque</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yuan yang</td>
<td>2003</td>
<td>-</td>
<td>Cantonese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yum cha</td>
<td>1988</td>
<td>1916</td>
<td>Cantonese</td>
<td>*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Total** 113

<table>
<thead>
<tr>
<th></th>
<th>20</th>
<th>21</th>
<th>40</th>
</tr>
</thead>
</table>

153
Abstract

Advanced and intermediate EFL learners often use the so-called “Big Five” English monolingual learner’s dictionaries (Cambridge Advanced Learner’s Dictionary, “Collins COBUILD” series, Longman Dictionary of Contemporary English, Macmillan English Dictionary for Advanced Learners, and Oxford Advanced Learner’s Dictionary of Current English). Learners rely heavily on these dictionaries’ examples as well as their definitions to learn the proper use of the words they look up, but Japanese EFL learners often find the examples too difficult and become reluctant to use monolingual dictionaries even if they are encouraged to do so. Typical factors that make the examples difficult for them include high levels of vocabulary, sentence length, and grammar, the last of which we focus our attention on.

We extracted our database of examples from the electronic versions of the Big Five dictionaries which allowed us to see what grammatical items are typically used in the dictionaries. The sizes of the examples are about 600,000 to 800,000 words in each dictionary. The grammatical items analyzed in our research are based on the list developed in a survey of the textbooks and reference books widely used in Japanese junior and senior high schools, which means it covers the grammatical items most often studied by Japanese students in class. This paper will reveal which grammatical items are frequently and infrequently used in dictionary examples and discuss how we can prepare them for the use of monolingual EFL dictionaries.

Keywords: learners’ dictionaries, database of examples, grammatical items, Japanese EFL learners

1. Grammatical items in EFL dictionaries

We expect EFL learners’ dictionaries to offer examples to users that are not too difficult and readable enough. Ishii (2011a) studied their examples in terms of vocabulary levels and confirmed that this is generally the case except for some examples of difficult headwords and the examples in COBUILD6. However, we also need to examine whether they are appropriate for learners in terms of grammar. This leads to the following two research questions: (1) Are there any characteristic uses of grammatical items in EFL dictionaries? (2) How can Japanese learners of English be better prepared to use monolingual EFL dictionaries?

In order to answer these questions, we have selected those grammatical items that are important for Japanese learners of English (Section 2) and have analyzed the full-text data of the examples (Section 3). We have also analyzed EFL textbooks and have compared the frequencies of the grammatical items (Section 4). Through these analyses we will identify and discuss the characteristics of examples in EFL dictionaries in terms of grammar (Section 5).

2. Selection of grammatical items

The previous literature on grammatical items in terms of CEFR (Common European Framework of Reference for Languages: Learning, Teaching, Assessment) levels include A Core Inventory for General English (North et al. 2010) and English Profile: Introducing the CEFR for English, Version 1.1 (2011). The former gives items such as the following for the
A1 level: “have got” (British), imperatives (affirmative/negative), past simple, “going to”, “I’d like”, adjectives (common and demonstrative), comparatives and superlatives, adverbs of frequency, and very basic intensifiers (“really” and “very”). The latter offers the following items for the A2 level, for example: sentences with clauses joined by “that”, descriptive phrases introduced by a past participle, simple sentences using infinitives, other infinitives, some modals, and so on. Looking at these items, we soon notice that some of them are difficult to clearly define to check their frequency in corpora because some of them are vocabulary items rather than grammatical ones and classifications are sometimes not detailed enough. In addition, although one of our aims is to focus on Japanese learners of English, the granularity of the classifications sometimes make them unsuitable for application to the Japanese EFL environment.

In order to solve these problems, we decided to select grammatical items widely recognized and accepted in Japan. We first consulted a list of grammatical items developed by Professor Hiroshi Sano at Tokyo University of Foreign Studies and his colleagues (Sano Lab 2005). They conducted a survey of the Japanese textbook market and studied 31 best-selling textbooks (six junior high school texts, eight high school “English I” texts, eight high school “English II” texts, and nine writing texts) and four widely used reference books (Minn et al. 2005). The criteria for inclusion into the list were as follows: (1) items that commonly occurred in all of the main textbooks, and (2) items that may not necessarily have occurred in all of the textbooks, but were common in several of the texts or determined to be highly useful for language learning. From this study, 144 grammatical items were chosen. The 144 items included grammatical items from the simple “I am + noun” to the more difficult S + V + C, emphasis syntax, or subjunctive past perfect. Basically, the 144 items were in the affirmative form, but for each grammatical item, 13 other subordinate items such as negative forms or interrogative sentences were made. Therefore, multiplying 144 items by 14 sentence types resulted in 2,016 items. Of those, however, 696 patterns did not actually exist as proper English sentences, so after deleting them, the total number was 1,320 sentence patterns. These patterns were defined following the CQL (Corpus Query Language) format using word forms, lemmas and POS tags.

In this current study, this list was adapted for our use. Some of the items were either too detailed or not detailed enough. So we divided “present/past passives” into two different items while we integrated items “comparative adjectives ending in ‘-er’”, “comparative adjectives with ‘more’”, “comparative adverbs ending in ‘-er” and “comparative adverbs with ‘more’” into one item, COMPARATIVES, and added lexical items such as “better” here, for instance. Also, some items were lacking (e.g., REFLEXIVE PRONOUNS and “used to”), so they were added in our new list. In addition, some items were difficult to define in terms of word forms, lemmas and POS tags, so their CQL expressions’ accuracy was not high enough; they were modified in order to achieve the highest precision that was possible. As for sentence types, our new list mainly focuses on affirmative patterns alone, but some questions and negatives that seem basic and important, especially for beginners to intermediate learners, were included (e.g., AFFIRMATIVE TAG QUESTIONS following negative sentences and NEGATIVE IMPERATIVES). Furthermore, the new list makes use of versatile regular expressions as the CQL which Sano Lab’s list used was available only on a particular in-house system and not available elsewhere.

The next step was to consult the Core Inventory and to add some items not covered in Sano Lab’s list. For example, we added such items as SIMPLE PAST, AUX + PERFECT + PASSIVE (e.g., ‘could have been done’), “used to” and INDIRECT SPEECH. The final number of our items is 157, part of which is shown in Table 1.
3. Use of full-text data of examples

In this section, we will describe how we can obtain and analyze the whole text data of dictionary examples. The following five dictionaries are examined in this study.

- COBUILD6: Collins COBUILD Advanced Dictionary of English. (2008) (The 7th (2012) and the 8th (2014) editions do not offer a CD/DVD-ROM which makes it impossible to use the data in the way described below.)
- OALD8: Oxford Advanced Learner’s Dictionary of Current English, 8th ed. (2010) (The 9th edition (2015) was not available when we carried out this study. It is not clear whether its accompanying DVD-ROM can be used in the same manner as we adopted in this study.)

We first need to extract XML-formatted data from the CD/DVD-ROMs accompanying the printed editions. The details of how to extract the XML data are explained in Ishii (2011b), but to put it briefly, we can just install the program and obtain the XML data for COBUILD6, and for the other four dictionaries the necessary XML data can be obtained by decompressing the files that are compressed, but not encrypted, in the same format. We then created a script which extracts examples from the XML files and removes unnecessary markup tags. The analysis conducted in this paper is based on the data on the CD/DVD-ROMs, which are different from the printed editions of each dictionary in some cases. Some dictionaries have additional entries on the electronic versions, but they are not used in the following analysis.

---

A few examples of the websites that provide the programs to retrieve the XML data are given below (as of May 20, 2015):
- Ruby: https://gist.github.com/tmaeda/948740/
- Python: https://github.com/superfan89/IDMSKconv/

The script used for this purpose is available on http://www.ishi-i.net/p/dictionarycorpus.html. The version of the script we used is 20150226.
The size of the data thus obtained is about 3,655,000 words in total. Table 2 lists the details.

Table 2 Approximate numbers of tokens in each example data

<table>
<thead>
<tr>
<th></th>
<th>CALD4</th>
<th>COBUILD6</th>
<th>LDOCE5</th>
<th>MED2</th>
<th>OALD8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>802,00</td>
<td>698,000</td>
<td>733,000</td>
<td>615,000</td>
<td>807,000</td>
</tr>
</tbody>
</table>

Finally, we POS-tagged and lemmatized the data with CLAWS, and converted each file into an XML format, which is an extended format from the one used in the BNC (Figure 1). The resulting data looks like the sample given in Figure 2.

Figure 1 Sample of the BNC (line breaks omitted)

```
<s n="2"> ... <w c5="PNP" hw="he" pos="PRON">he</w> <w c5="VBD" hw="be" pos="VERB">was</w> <w c5="VVG" hw="drive" pos="VERB">driving</w> <w c5="PRP" hw="through" pos="PREP">through</w> <w c5="NP0" hw="lydsett" pos="SUBST">Lydsett</w> <w c5="NN1" hw="village" pos="SUBST">village</w> <c c5="PUN">.</c></s>
```

Figure 2 Sample of our final data

The regular expressions for searching each grammatical item correspond to this format: ‘c7="PPGE"’ for POSSESSIVE PRONOUNS (15) and ‘c5="(VBB/VBZ)[^<>]+[^<>]+</w> ?’ for PRESENT PROGRESSIVE sentences (22), for example.

4. Comparison with EFL textbooks

In order to answer our research questions, we need to look at the use of grammatical items in EFL textbooks. We digitized “English Communication I” textbooks for first-year (tenth-grade) Japanese high school students (“JHS1” in Table 3) and CEFR-based EFL textbooks targeted at learners of B1/B2/C1 levels published by foreign publishers. We scanned the textbooks and processed them with some OCR software to get the texts, which were then POS-tagged and lemmatized and finally reformatted into the same XML format as we used for our dictionary examples. The details of the digitized textbooks are summarized in Table 3.

Table 3 Data size of our CEFR-based textbook corpus

<table>
<thead>
<tr>
<th></th>
<th>Number of textbooks</th>
<th>Number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>JHS1</td>
<td>24</td>
<td>185,000</td>
</tr>
<tr>
<td>EFL B1</td>
<td>27</td>
<td>441,000</td>
</tr>
<tr>
<td>EFL B2</td>
<td>24</td>
<td>517,000</td>
</tr>
<tr>
<td>EFL C1</td>
<td>9</td>
<td>253,000</td>
</tr>
</tbody>
</table>

32 The figures are based on the number of <w> elements in the reformatted data (see Fig. 2). Please note that each multi-word unit such as “in accordance with” is counted as one token, not three tokens, in this study.

33 The numbers of words are based on the number of <w> elements in the final XML data. Those words given in vocabulary lists are excluded from these numbers.
5. Data and discussion

5.1. Frequency of each grammatical item
We counted the frequency of each grammatical item in the dictionary examples and textbooks. Part of the results is given in Table 4.

Table 4 Sample of the frequency table (per million words)

<table>
<thead>
<tr>
<th>ID</th>
<th>Grammatical item</th>
<th>JHS 1</th>
<th>EFL B1</th>
<th>EFL B2</th>
<th>EFL C1</th>
<th>CALD D4</th>
<th>COBUG D6</th>
<th>LDOC E5</th>
<th>MED 2</th>
<th>OAL D8</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>FREQUENCY ADVERBS</td>
<td>2,21</td>
<td>3,55</td>
<td>3,23</td>
<td>2,34</td>
<td>1,997</td>
<td>1,442</td>
<td>1,629</td>
<td>1,655</td>
<td>1,393</td>
</tr>
<tr>
<td>27</td>
<td>PAST PERFECT</td>
<td>1,25</td>
<td>638</td>
<td>1,46</td>
<td>1,76</td>
<td>1,746</td>
<td>2,952</td>
<td>2,120</td>
<td>2,345</td>
<td>1,950</td>
</tr>
<tr>
<td>49</td>
<td>COMPARATIVES</td>
<td>4,09</td>
<td>4,35</td>
<td>4,94</td>
<td>4,87</td>
<td>2,809</td>
<td>2,982</td>
<td>2,955</td>
<td>2,883</td>
<td>2,629</td>
</tr>
<tr>
<td>59</td>
<td>might</td>
<td>280</td>
<td>549</td>
<td>1,00</td>
<td>754</td>
<td>359</td>
<td>387</td>
<td>372</td>
<td>289</td>
<td>178</td>
</tr>
<tr>
<td>13</td>
<td>What ...?</td>
<td>885</td>
<td>2,01</td>
<td>1,50</td>
<td>2,19</td>
<td>442</td>
<td>205</td>
<td>563</td>
<td>491</td>
<td>543</td>
</tr>
</tbody>
</table>

It will be of interest to see the difference among the five dictionaries, but in this paper we will focus on the difference between dictionary examples and textbooks, adhering to our research questions.

5.2. Items used less frequently in dictionary examples
In order to answer our first research question (Are there any characteristic uses of grammatical items in monolingual EFL dictionaries?), we looked through the results and found a few items that are less or more frequently used in EFL dictionary examples than the textbooks for intermediate to advanced learners. Those items that occur less frequently in dictionary examples include FREQUENCY ADVERBS (Figure. 3), COMPARATIVES (Figure. 4), some MODALS (“can”, “could”, “may”, “might” and “would”), TAG QUESTIONS, SUBORDINATE CLAUSES and INDIRECT SPEECH constructions. The purpose of the examples in dictionaries is to illustrate how the headwords are actually used, but the lexicographers need to be constantly conscious of space restrictions. They are required to strike a balance between natural English and short examples. It is often the case that items such as FREQUENCY ADVERBS, COMPARATIVES, MODALS and TAG QUESTIONS can be omitted without causing unnaturalness. SUBORDINATE CLAUSES often require almost double the length of simple clauses, and the same holds true about RELATIVE PRONOUNS (NONRESTRICTIVE), RELATIVE ADVERBS, PARTICIPIAL CONSTRUCTIONS and SUBJUNCTIVES that are all used less frequently in the dictionary examples. INDIRECT SPEECH constructions can be avoided simply by giving the utterances or questions. Also, MODALS are often associated with subjective understandings of situations, which may need more contextual information, i.e., more space.
In addition to the space limitations, text types may be a factor for the difference; AFFIRMATIVE IMPERATIVES and some types of QUESTIONS (“Do you ...?”, “Why ...?”, “What ...?”, “Which ...?” and “How ...?”) occur less frequently in dictionary examples, perhaps because these items are more favored in spoken language. These may be some of the plausible reasons why the dictionary examples have less of the above items than textbooks which have less space limitations and which include various text types.

5.3. Items used more frequently in dictionary examples

In order to answer our second research question (How can Japanese learners of English be better prepared to use monolingual EFL dictionaries?), we should look at those items that are more frequently used in dictionary examples. The most noticeable ones include complex structures such as PRESENT PERFECT PASSIVE and PAST PERFECT PASSIVE (Fig. 5) and EXISTENTIAL “there” + PERFECT (Fig. 6). Although we cannot say for sure what the reasons for these high frequencies are, it is a fact that some grammatical items, especially complex structures, are typically used in dictionary examples. To make sure that the users of the dictionaries can fully understand their examples and benefit from the information therein, teachers should know which items are typically used in monolingual EFL dictionaries and teach their students accordingly. Another possibility is that dictionaries themselves offer information on the grammatical items they often employ in their examples (and in their

---

34 Negative imperatives are, on the contrary, used more frequently in dictionary examples, the reasons for which are of interest but beyond the scope of this paper.
It should not be a wild idea because such a list of grammatical items is just like a list of a defining vocabulary offered in most monolingual EFL dictionaries.

6. **Summary and future studies**

We have seen some characteristic features of the examples of the “Big Five” EFL monolingual dictionaries. Although they all function as a model for learners, they tend to use certain grammatical items more frequently or less frequently than EFL textbooks that learners use to study English for their classes. Getting more aware of these facts can lead to learners’ better use of these dictionaries.

What distinguishes this research is the fact that it is based on the complete data of all examples, not on just fragmentary samples. It is true that the data presented in this paper is not completely accurate; processing the data was carried out automatically and some manual data editing and elaboration of the patterns for the grammatical items would be necessary to achieve more accuracy. There is also skewness in the data resulting from different policies taken in each dictionary or textbook; for example, the COBUILD dictionaries, especially, differ greatly from others in grammar as well as the length of examples and vocabulary used.
in them. Considering these aspects would enable us to make a fair and constructive comparison of the examples in EFL dictionaries, as well as reveal more features characteristic of these dictionaries.

Acknowledgments

This work was partially supported by Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (A) #24242017 and JSPS Grant-in-Aid for Young Scientists (B) #26770201.

References


35 For a discussion on the vocabulary levels of the examples of the five dictionaries (some of which are previous editions), see Ishii (2011b).
Where is the place for \textit{slang}? An analysis of labeling of restrictive word usage in monolingual English learners’ dictionaries

Susanna Bae$^1$, Young-kuk Jeong$^2$
International Graduate School of English, South Korea
$^1$susannabae@gmail.com, $^2$young@igse.ac.kr

Abstract
Among a range of labels in English monolingual learners’ dictionaries, restrictive usage labels such as \textit{slang}, \textit{very informal}, \textit{impolite} or \textit{offensive} are very important in that they alert learners to serious potential risks resulting from the inappropriate use of words with these indication markers. The question, however, is whether they are easily and clearly understood by target dictionary users. The present study focuses on the use of \textit{slang} in the major monolingual English dictionaries for advanced learners and critically evaluates their functions against those of other closely related labels. Dictionary analysis indicated that there are considerable inconsistencies in the treatment of these usage labels. A survey questionnaire also showed that some of the labels are likely to create confusion among the target users. Suggestions are made as to ways of increasing the consistency and comprehensibility of restrictive usage information in dictionaries.

Keywords: dictionary criticism, restrictive usage labels, monolingual learners’ dictionaries of English, slang, impolite taboo

1. Introduction

Slang is an elusive term. The most recent edition of \textit{Oxford Advanced Learner’s Dictionary} (9$^{th}$, 2015) provides the following definition for the word: “Very informal words and expressions that are more common in spoken language, especially used by a particular group of people, for example, children, criminals, soldiers, etc.” From this simple definition alone, the term’s multifaceted characteristics quickly emerge—a high degree of informality, spoken mode, user group affiliation, and usage sometimes restricted to members of particular social strata.

Other major learners’ dictionaries offer similar definitions, among which are the definition from \textit{Merriam-Webster Learner’s Dictionary}, which points out another distinctive feature of slang: “Words that are not considered part of the standard vocabulary of a language […]”, and that of \textit{WordNet 3.1}, which goes further to include yet another characteristic: “Informal language […] often vituperative or vulgar.” The definition excerpts from these two dictionaries add to slang the attributes of being non-standard, vulgar and/or offensive. However, mere vulgarism is not slang. As Coleman (2012: 21) points out, swearing is not necessarily slang although slang words are often associated with swearing words.

Moreover, slang is not only informal, but it also has various pragmatic effects and socio-linguistic functions. According to who uses it to whom in what occasion, slang makes speech more affectionate, humorous, disrespectful, playful, or pithy, as in the definition from \textit{American Heritage Dictionary}: “A kind of language occurring chiefly in casual and playful speech, made up typically of coinage and figures of speech that are deliberately used in place of standard terms for added raciness, humor, irreverence, or other effect.” In addition to the ‘playfulness’ characteristic, the AHD’s definition points to another very significant characteristic of slang expressions—that they usually involve the creation of new words or new meanings for existing words. Like newly coined words in general, some slang words that once were obscure to most people may evolve to become an established part of the language, whereas others may have currency only for a short time. As Hornby (1963: xi) admitted,
“Nothing more quickly dates than slang”. This constant change in the linguistic status of slang vocabulary means to the lexicographer the on-going revision of entries to which the label slang is attached.

2. Strengths and weaknesses of slang as a dictionary label

From the definitions discussed in the previous section, it can be summarized that the following characteristics constitute slang vocabulary:

- spoken mode, colloquial—when people speaking together;
- very informal;
- non-standard, or deliberate substitution for a standard equivalent;
- used by particular professional or social groups;
- vulgar, inelegant;
- impolite, vituperative, swearing;
- coinage, neologism, creation of new meanings to current words;
- often with intended effects of playfulness, humor, irreverence, ellipsis, vividness, etc.

Slanginess may consist of the combination of all or some of those characteristics, but it still distinguishes itself from any one of them. Thus if we call a word slang, we may have to decide why it is slang rather than a very informal, new, or impolite word. If slang vocabulary is placed in the overlapping space between humorous and impolite, for instance, this may pose a problem to an English learner. Without knowing exactly why and how a slang word may sound, he or she would feel unsure as to whether it has more of an informal and relaxed connotation, a disrespectful connotation, or both in equal measure. Another problem is that these characteristics are not inherent in a word, but only identifiable in light of the context in which it is used. Moreover, some of the characteristics are within the realm of objective criteria (such as register, mode, regions, nonconformity to standard, and group membership), whereas it is very difficult to objectively judge how much a word is vulgar, humorous, or playful. Coupled to these is the task of currency check. As mentioned earlier, each of slang words follows its own path into informal, standard, dated, or obsolete. For instance, log on was a slang jargon among computer enthusiasts during 1960s (Coleman 2012).

Technological advances in communication media and the spread of social network services make outdated the idea of slang as chiefly spoken language. For one thing, online communication is a hybrid between spoken and written language. From this perspective we may need to modify the view about slang as being both off and online colloquialism. Secondly, nearly instantaneous connection between people through online media speeds up the dissemination of slang vocabulary into wider population, making it very difficult to identify a particular social group to which a slang word belongs. If we cannot identify its main users, the word would better be categorized as (very) informal rather than slang (Coleman 2012).

The peculiarity of slang is also apparent in existing classifications of dictionary labeling system. Hausmann (1989, cited in Svensén 2009: 316) regards slang as a socio-cultural status marker along with such labels as popular and vulgar, and differentiates it from the other register and normativity categories. Similarly, according to Miller’s (2011) summary of literature on classification of dictionary labels, slang is grouped under the same status heading as the labels dialect, jargon, and taboo, amongst others. On the other hand, Svensén (2009: 327-328) cautiously proposes a continuum of ‘style’ level on which slang is placed somewhere between colloquial and vulgar. Landau (2001) and Atkins and Rundell (2008), in contrast, isolate the slang label from all the other labeling categories, presumably in recognition of its nonconformity to any one specific criterion. In short, depending on how much importance we attach to certain aspects of slang, it can either be put under the category
Dictionary labeling of slang is notoriously inconsistent […] It is easier to spend one’s time constructing theories to explain why there is no such thing as slang than to identify and harness its peculiar energy and spice and stick a label on it. […] Slang is a useful concept. Since there are no external criteria for identifying slang, we must support efforts to establish them. […] In day-to-day decisions, words are labeled slang by lexicographers or their advisers because the words are deemed extremely informal. This is unsatisfactory; slang is not simply very informal usage. But until we have agreed criteria by which to judge them, slang and informal words will appear in more or less free variation in dictionaries.

To paraphrase what Landau argues, if it is less than satisfactory to attach the slang label to words just because they are very informal, the opposite practice may also be unsatisfactory, that is to label as very informal words that would more aptly be regarded as slang. The labeling system of a dictionary is centrally an editorial decision; it depends on the dictionary’s purpose and intended users. We may, to some extent, compromise accuracy for simplicity when it comes to dictionaries for EAL learners; but before doing so, we should first ask whether slang is still a useful concept that helps English learners improve their knowledge and control of English vocabulary. The issue of establishing objective criteria then follows. If the slang label is considered to have its own functions distinguished from other similar labels’, and if this discrimination is considered to be relevant to target dictionary users, then it should be consistently and judiciously applied so that its intended message can be correctly communicated to the users.

The present study is motivated by the need for a critical evaluation of slang and related labels from an EAL learner’s perspective. For this purpose, it first investigates the use of the slang label in the OALD eighth and ninth editions and compares it to other closely related labels—informal, very informal, impolite, offensive, taboo, etc.—commonly included in monolingual English dictionaries for learners. The dictionary analysis mainly involves the collection of entries labeled as slang from Section B of OALD8, as well as the comparison of these to the same entries in different learners’ dictionaries. Also, in an attempt to assess the efficiency of using the slang and other related labels, a questionnaire survey examining the understanding
of slang and other related labels was sent out to a group of adult Korean advanced learners of English. The following research questions guided the study:

1) Are the characteristics of slang other than the degree of formality present in the slang entries in OALDs?
2) How do Korean advanced learners of English understand the slang and related labels occurring in the major English monolingual learners’ dictionaries?

In the following sections the main data analysis procedures are presented along with their results. In light of what has been found from this study, the final section suggests ways to improve the process of identifying and describing slang/very informal entries in dictionaries for language learners.

3. Slang entries from the major monolingual learners’ dictionaries

The entry comparison was carried out mainly in order to find out what slang means in different dictionaries and how it is related to other labels. One-to-one comparisons between dictionaries were not possible because all five dictionaries adopted different labeling policies. CALD and OALD maintain the slang label, but the former also has very informal along with informal, while the latter, while not having very informal, specifies that slang refers to ‘very informal language’ in its user guide. This may suggest that CALD assigns a particular function to slang which is differentiated from that of very informal, whereas OALD may use slang more loosely as a substitute for very informal. By contrast, MEDAL substitutes very informal for slang, which its user guide states is ‘used in very informal situations and mainly among people who know each other well.’ It adds that ‘Some dictionaries use the label slang’. LDOCE and COBUILD, on the other hand, have only the informal label, without further specifying the degree of informality. Therefore, it was decided to begin with slang entries from OALD, against which the equivalent entries in the other learners’ dictionaries are compared.

Slang entries were collected within the section B of OALD8 by means of the right-core field search function available on its CD-ROM version. The headword search by register labels yielded 82 slang entries in total. The count of an entry unit was based on each individual sub-sense of a lemma. For example, the number 4 sense of baby, which is labeled as slang, is counted as one entry whereas the lemma bitch has three sub-senses, among which only the second and third are marked as slang, and which must thus have two separate entries. If slang is given in the left-core field (as in the noun lemma bugger), each of its sub-senses was counted as one individual entry. The entries collected from OALD8 were compared to those of OALD9 to see if any revisions occurred: the labels for the entry Black Maria were modified from slang; old-fashioned to informal; old-fashioned, and thus it was excluded from the comparison. The remaining entries were then compared to corresponding entries from the other four dictionaries—LDOCE5, COBUILD6, on-line CALD and MEDAL. Seven out of the OALD’s slang entries were not found in any of the other dictionaries, thus they were also excluded, which resulted in 74 entries in total to be analyzed.

The analysis of the distribution of labels for these entries showed the following patterns: 1) the slang words (of OALD) were neither considered slang nor even as very informal in the other dictionaries; 2) Nearly half of the OALD’s slang entries were accompanied with the label taboo. We will look at each of these findings below.

The first comparison was made between OALD and CALD, and it showed that there were only eleven corresponding slang entries found in CALD. As for the very informal label, CALD had even fewer—only four entries were labeled as very informal. This pattern was not very different from the comparison to MEDAL. Only five entries were labeled as very informal in MEDAL. All the rest were either labeled as simply informal/spoken, or marked
with other derogatory labels (i.e. offensive, not polite, impolite, rude, taboo) throughout the four dictionaries. Only the following OALD slang entries agreed with the two other dictionaries. Table 1 shows the entries in which the three dictionaries are marked as either slang or very informal:

<table>
<thead>
<tr>
<th>Slang entries from OALD</th>
<th>Accompanying labels (OALD)</th>
<th>CALD</th>
<th>MEDAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>beak(^f)</td>
<td>old-fashioned</td>
<td>old-fashioned slang</td>
<td>old-fashioned informal</td>
</tr>
<tr>
<td>beat it</td>
<td>slang</td>
<td></td>
<td>spoken</td>
</tr>
<tr>
<td>berk</td>
<td>old-fashioned</td>
<td>slang</td>
<td>old-fashioned informal</td>
</tr>
<tr>
<td>bint</td>
<td>slang</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>bird(^d)</td>
<td>offensive (8) (\rightarrow) old-fashioned (9)</td>
<td>slang</td>
<td>informal</td>
</tr>
<tr>
<td>a bit of all right</td>
<td>slang</td>
<td></td>
<td>very informal</td>
</tr>
<tr>
<td>bitch(^3[difficulties])</td>
<td>very informal</td>
<td>very informal</td>
<td></td>
</tr>
<tr>
<td>bleeding</td>
<td>taboo</td>
<td>very informal</td>
<td>impolite</td>
</tr>
<tr>
<td>bloody</td>
<td>taboo</td>
<td>offensive</td>
<td>very informal</td>
</tr>
<tr>
<td>blow job</td>
<td>taboo</td>
<td>- [no entry]</td>
<td>very informal</td>
</tr>
<tr>
<td>blow(^15) n.</td>
<td>slang</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bog off</td>
<td>taboo</td>
<td>slang</td>
<td>spoken impolite</td>
</tr>
<tr>
<td>boob(^1)</td>
<td>very informal</td>
<td></td>
<td>informal</td>
</tr>
<tr>
<td>boss adj.</td>
<td>slang</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>brass monkeys</td>
<td>informal</td>
<td></td>
<td>very informal</td>
</tr>
<tr>
<td>BS(^3)</td>
<td>taboo</td>
<td>very informal</td>
<td>impolite</td>
</tr>
<tr>
<td>have/take a butcher’s</td>
<td>old-fashioned</td>
<td>slang</td>
<td>informal</td>
</tr>
</tbody>
</table>

Granted, we cannot draw conclusions about which dictionary is better in terms of the labeling of very informal/slang words based upon this simple distribution table alone; indeed it is quite apparent, even within this small collection of entries, which were noted to fall more solidly under the very informal or slang labels, that characteristics of slang beyond very informal are difficult to find. In the case of OALD, it seems that the slang label is used very arbitrarily from simply very informal words to barely informal ones. It is very hard for an English learner who looks up boob, for example, in OALD9 to understand why the word is slang when it means ‘a woman’s breast’ whereas it is just North American English when it means ‘a stupid person’. Similarly, it is very difficult to find a convincing reason why bloody remains a slang word when ‘everyone knows it, and most people use it, particularly when they are trying not to be offensive’ (Coleman 2012: 22). The exceptions are few: the entry for berk (a stupid person) gives enough information to a dictionary user as to why it is more of a slang word, telling us that it originates from the rhyming slang, Berkeley/Berkshire Hunt, which referred to a cunt (a stupid person) during the 1930s. To have a butcher’s is another example of rhyming slang, and beak (an insulting word for judge) came from 18th century criminal slang. Among the words analyzed, only these last words possess the low frequency and very limited age-group usage which qualifies them as slang rather than simply (very) informal expressions. As for bird (referring to a girl), which was previously labeled as ‘slang; sometimes offensive’, it has been modified to slang; old-fashioned in OALD9. But if bird is a
slang word for a girl, a learner may wonder, for what reason is chick labeled only as old-fashioned, without any other restrictive labels.

Another very prominent pattern was the label’s co-occurrence with the taboo label. In OALD, nearly half of the slang entries (36 among 74, or 41 among the original 82 entries) were accompanied by the taboo label. For taboo or offensive entries, LDOCE has a scale of offensiveness ranging from ‘not polite’ to ‘taboo’, and this corresponds with MEDAL’s continuum of ‘impolite—offensive’ scale. As to CALD, it abandoned the taboo label and instead adopted the single label ‘offensive’. On the contrary, OALD and COBUILD use the ‘offensive’ and ‘taboo’ labels to describe different types of language, rather than to describe the degree of offense of words of the same type. COBUILD minutely distinguishes the speaker’s intended purpose (i.e. insult) from the hearer’s potential reaction (i.e. shock), and assigns the ‘offensive/very offensive’ labels to the former, and the ‘rude/very rude’ to the latter. According to COBUILD, words which describe sexual organs or excretion can thus be marked as (very) rude, while insulting words referring to people’s characteristics, race, religion, etc. are more likely to be marked as (very) offensive. Likewise, the OALD’s paired labels ‘offensive—taboo’ have the same function, with the former describing words which are ‘used by some people to address or refer to people in a way that is very insulting, especially in connection with their race, religion, sex or disabilities’, and the latter describing words which are ‘likely to be thought by many people to be obscene or shocking’. It is interesting to see that taboo has become a less frequently used labeling term, just like slang. Three of the dictionaries do not use taboo at all, and LDOCE, which is one of the two that still include taboo, uses it very rarely compared to OALD.

Table 2 summarizes the list of OALD’s ‘taboo+slang’ entries and the distribution of offensive/impolite labels provided in the other dictionaries:

<table>
<thead>
<tr>
<th>Taboo+slang entries from OALD</th>
<th>LDOCE</th>
<th>COBUILD</th>
<th>CALD</th>
<th>MEDAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ball^ v.</td>
<td>informal, not polite</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>balls^ [nonsense]</td>
<td>informal, very rude</td>
<td>offensive</td>
<td>impolite</td>
<td></td>
</tr>
<tr>
<td>balls^ [courage]</td>
<td>not polite</td>
<td>informal, rude</td>
<td>offensive</td>
<td>impolite</td>
</tr>
<tr>
<td>Balls^ [exclamation]</td>
<td>not polite</td>
<td>-</td>
<td>offensive</td>
<td>impolite</td>
</tr>
<tr>
<td>balls^ [testicles]</td>
<td>not polite</td>
<td>-</td>
<td>offensive</td>
<td>impolite</td>
</tr>
<tr>
<td>balls-up</td>
<td>informal</td>
<td>informal, rude</td>
<td>offensive</td>
<td>impolite</td>
</tr>
<tr>
<td>balls sth up</td>
<td>informal</td>
<td>not polite</td>
<td>informal, rude</td>
<td>offensive</td>
</tr>
<tr>
<td>bang^ v.</td>
<td>not polite</td>
<td>-</td>
<td>offensive</td>
<td>-</td>
</tr>
<tr>
<td>bastard^ [unpleasant]</td>
<td>taboo</td>
<td>informal, very rude</td>
<td>offensive</td>
<td>offensive</td>
</tr>
<tr>
<td>beat off</td>
<td>informal, not polite</td>
<td>-</td>
<td>offensive</td>
<td>-</td>
</tr>
<tr>
<td>bloody</td>
<td>-</td>
<td>rude</td>
<td>very informal</td>
<td>impolite</td>
</tr>
<tr>
<td>blow job</td>
<td>informal</td>
<td>-</td>
<td>offensive</td>
<td>very informal</td>
</tr>
<tr>
<td>bog off</td>
<td>informal</td>
<td>-</td>
<td>slang</td>
<td>impolite</td>
</tr>
<tr>
<td>bollocks^ [nonsense]</td>
<td>informal</td>
<td>informal, rude</td>
<td>offensive</td>
<td>impolite</td>
</tr>
<tr>
<td>bollocks^ [testicles]</td>
<td>informal</td>
<td>informal, rude</td>
<td>offensive</td>
<td>impolite</td>
</tr>
<tr>
<td>Bollocks^ [exclamation]</td>
<td>informal</td>
<td>informal, rude</td>
<td>-</td>
<td>impolite</td>
</tr>
<tr>
<td>BS^</td>
<td>not polite</td>
<td>-</td>
<td>very informal</td>
<td>impolite</td>
</tr>
<tr>
<td>bugger^ [insulting a man] n.</td>
<td>not polite</td>
<td>informal, rude</td>
<td>offensive</td>
<td>impolite</td>
</tr>
<tr>
<td>bugger^ [sympathetic] n.</td>
<td>not polite</td>
<td>-</td>
<td>informal</td>
<td>impolite</td>
</tr>
<tr>
<td>bugger^ [difficulties] n.</td>
<td>not polite</td>
<td>informal, rude</td>
<td>offensive</td>
<td>-</td>
</tr>
<tr>
<td>bugger^ [don’t care] v.</td>
<td>not polite</td>
<td>informal, rude</td>
<td>offensive</td>
<td>impolite</td>
</tr>
</tbody>
</table>
As can be seen in Table 2, OALD’s taboo+slang words roughly coincide with COBUILD’s ‘rude’, LDOCE’s ‘not polite’, CALD’s ‘offensive’, and MEDAL’s ‘impolite’ labels. However, there are two points that merit special attention here. First, the degree of restriction suggested by the combination of slang and taboo to a word may be much greater than the word really deserves. As taboo may mean either an extremely offensive word or a very rude word, this semantic overload may further confuse a learner, unless he or she carefully examines the dictionary’s user guide to see what this label means. Secondly, in the other dictionaries, there is a tendency for the impoliteness markers to have priority over the register markers. The only exception to this is COBUILD, which provides the register (i.e. informal) and status (i.e. rude) labels independently. In such a case, it would be more justifiable to give the most important information the greatest prominence; if a word is sure to sound impolite, then comments about its informality may be redundant. It was found that OALD has only a few ‘offensive’ entries compared to ‘taboo’ entries—the search within section B yielded only six entries marked as ‘offensive’: baldy, bird, blonde, brunette, bulldyke, and banana republic—and all but one (bird) had no accompanying ‘informal’ or ‘slang’ markers. There may be a lot of taboo entries, too, which do not need the additional slang marker.

### 4. Understanding of slang and very informal labels in Korean learners of English

A questionnaire survey was sent out to a group of Korean learners of English and received 75 responses. Some of the participants were postgraduate students in English Language Teaching and related fields, and all the rest had master’s degrees in ELT and applied linguistics. 84% of the respondents answered that they are familiar with English monolingual dictionaries, and 83% were frequent users of MLDs. In order to find out how they understand the slang and very informal labels, the questionnaire included thirteen statements about each of the ‘very informal’ and ‘slang’ labels, and asked the respondents to tick whichever option on the scale best represented their opinion. The numerical score 0 was given to ‘strongly disagree’, 1 to ‘disagree’, 2 to ‘agree’ and 3 to ‘strongly agree’. The questionnaire contained the following statements for each question, which were originally written in Korean:

**Question A:** When a certain word in your dictionary has the usage label ‘very informal’, what characteristics do you assume the word has? Click on the radio button that best applies to you:

<table>
<thead>
<tr>
<th>Word</th>
<th>Usage Label</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>bugger</td>
<td>informal, rude</td>
<td>offensive, impolite</td>
</tr>
<tr>
<td>buggered</td>
<td>informal, rude</td>
<td>offensive, impolite</td>
</tr>
<tr>
<td>bugger all</td>
<td>not polite</td>
<td>offensive, impolite</td>
</tr>
<tr>
<td>bugger me</td>
<td>not polite</td>
<td>offensive, impolite</td>
</tr>
<tr>
<td>I'll be buggered</td>
<td>not polite</td>
<td>offensive, impolite</td>
</tr>
<tr>
<td>I'll be buggered if...</td>
<td>not polite</td>
<td>offensive, impolite</td>
</tr>
<tr>
<td>buggery</td>
<td>law</td>
<td>specialized, legal</td>
</tr>
<tr>
<td>bullshit</td>
<td>informal, rude</td>
<td>offensive, impolite</td>
</tr>
<tr>
<td>bullshit v.</td>
<td>informal, rude</td>
<td>offensive, impolite</td>
</tr>
</tbody>
</table>
Question B: When a certain word in your dictionary has the usage label ‘slang’, what characteristics do you assume the word has? Click on the radio button that best applies to you:

Statements:
1) It is a spoken word.
2) It is an informal word.
3) It is a non-standard word.
4) It is a word that a certain group of people mainly uses.
5) It has a lot of newly coined words.
6) It is a word that people in close relationships use.
7) It is a word that people in the same generation use.
8) It is a vulgar word.
9) People who use this word seem trendy and up-to-date.
10) It gives people discomfort.
11) It is rude.
12) It cannot be used in a formal situation.
13) It must not be used under any circumstances.

Table 3 summarizes the participants’ responses to each of the statements made about the labels ‘very informal’ and ‘slang’. In order to find out where the difference occurs in terms of the respondents’ perception, the Mann-Whitney U test was used:

<table>
<thead>
<tr>
<th>Statement</th>
<th>Question A: very informal</th>
<th>Question B: slang</th>
<th>Asymp. Sig. (2-tailed)**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean*</td>
<td>s.d.</td>
<td>mean*</td>
</tr>
<tr>
<td>Statement 1</td>
<td>2.62</td>
<td>.613</td>
<td>2.65</td>
</tr>
<tr>
<td>Statement 2</td>
<td>2.78</td>
<td>.53</td>
<td>2.75</td>
</tr>
<tr>
<td>Statement 3</td>
<td>1.56</td>
<td>.952</td>
<td>2.60</td>
</tr>
<tr>
<td>Statement 4</td>
<td>1.79</td>
<td>.838</td>
<td>2.60</td>
</tr>
<tr>
<td>Statement 5</td>
<td>1.63</td>
<td>.82</td>
<td>2.08</td>
</tr>
<tr>
<td>Statement 6</td>
<td>2.37</td>
<td>.697</td>
<td>2.36</td>
</tr>
<tr>
<td>Statement 7</td>
<td>2.07</td>
<td>.73</td>
<td>2.20</td>
</tr>
<tr>
<td>Statement 8</td>
<td>1.52</td>
<td>.753</td>
<td>2.17</td>
</tr>
<tr>
<td>Statement 9</td>
<td>1.39</td>
<td>.779</td>
<td>1.58</td>
</tr>
<tr>
<td>Statement 10</td>
<td>1.75</td>
<td>.823</td>
<td>2.08</td>
</tr>
<tr>
<td>Statement 11</td>
<td>1.29</td>
<td>.684</td>
<td>1.84</td>
</tr>
<tr>
<td>Statement 12</td>
<td>2.31</td>
<td>.81</td>
<td>2.43</td>
</tr>
<tr>
<td>Statement 13</td>
<td>.51</td>
<td>.583</td>
<td>1.17</td>
</tr>
</tbody>
</table>

*0-3 point scale; **grouping variable: label

The results showed that the high degree of informality and colloquialism entailed with the slang/very informal labels is clearly understood by the respondents, as seen in the results of their response to statements 1, 2, 6, and 12. However, the respondents understood the slang label to be quite distinct from the very informal label in many respects. In particular, the options including nonconformity to the standard language (statement 3), strict social restriction of its usage (statements 4, 10, and 10), vulgarism (statement 8), swearing or rudeness (statement 11) were significantly and clearly perceived by respondents thanks to the slang label. On the other hand, while the respondents generally agreed that slang and very informal words usually involve neologism, the difference was not significant. Raciness or
playfulness, which are claimed to be typical characteristics of slang words, were not effectively perceived either, even by proficient users of English.

At this point, we may discuss the function of ‘very informal’ compared to the ‘informal’ label. While the slang label is semantically too overloaded, the very informal label is semantically nearly empty. In spite of its apparent clarity, the very informal label seemed to fail to convey to most of the respondents what is special in being ‘very’ informal. For learners, it may be very difficult to appreciate the quality of being ‘very informal’, and differentiate it with that of being ‘informal’. We may need to establish objective external criteria for identifying not only slang, but also very informal.

5. Suggestions and conclusion

Simply attaching the slang label to an entry may not do justice to the ‘spice and peculiar energy’ (Landau 2001: 240) the word sends to the hearer. A satisfactory description of the meaning of a slang word requires clearer criteria for identification as well as more diverse defining strategies on the part of the lexicographer. It was mentioned earlier in this paper that some of the certain characteristics may serve as a more objective guideline for identifying slang. These include register, mode, regions, deviation from the standard, user group, and currency. These characteristics may be supported rather objectively. In deciding whether a word can be regarded as slang, we may ask the following questions:

- Can we identify the social group it belongs to?
- Is there evidence from off- and online sources relating its register and mode?
- Is there a distinctive tendency in its use in terms of its national or regional boundaries?
- Is there a characteristic of unconventional meaning, grammar, or typography?
- How frequently it is used?
- Is the word used widely enough so as to become a part of the unmarked informal?
- Was it used during a particular period of time in the past?

For other more subtle characteristics which are also highly context dependent, the glossing policies proposed by Dalzell, Victor and Williams (2002) would be a good guide. In their report on the revision of Eric Partridge’s Dictionary of Slang and Unconventional English, the three slang lexicographers advised including the following information in a slang entry (adapted from Dalzell, Victor and Williams 2002: 333):

- The main users of the term, on a cultural, demographic, and/or geographical basis;
- Etymology;
- Illustrative examples of usage;
- Information concerning the tone of voice or gestures used when uttering the word;
- Degree of stigmatization or taboo;
- Synonyms, antonyms, and collocates;
- Identification of the entry by figure of speech (such as metaphor, euphemism, rhyme, etc.)

Some of these elements would be very relevant to English learners’ dictionaries. For English learners who do not have a firsthand feeling for the nuance of a given slang word, illustrative examples showing various contexts of its use as well as its associated paralinguistic behavior would be very invaluable. Since a word’s status as slang is sometimes contingent upon by and to whom it is spoken, the information about the usual speaker and its typical pragmatic force may at times also be required. In this respect, the definition of baby from OALD is informative: ‘A word used to address sb, especially your wife, husband or lover, in a way that expresses affection but that can be offensive if used by a man to a woman he does not know.’
Slang is regarded as the deliberate substitution of the standard terms for some added effect. Then it would be pedagogically justifiable to show, where possible, its more standard or less offensive equivalents in a slang entry, as in the entry for bugger something up: ‘a more polite, informal way of saying this is foul something up, mess something up or bungle something’ (OALD9). Currently this practice is applied more often to offensive entries, rather than slang ones, but slang entries would also certainly benefit from such kind of treatment.

This study was carried out to examine whether slang is a useful term as a label for English learners’ dictionaries. It is still felt that more questions remain than answers given. Labeling is a petty, ‘unsexy’ labor for which rewards are incredibly poor: no dictionary sells for its elaborated labeling system. There is no shortage of arguments, too, against preserving labels carrying semantic information (Meer 2008). The slang label may be ranked highest in terms of semantic information; but it is still questionable whether ‘very informal’ is a satisfactory alternative. Thus, we were only left with the options of tentatively proposing a more objective set of criteria for identifying slang words and more systematic and comprehensive defining principles for these words. During the course of the study, we have come across another emerging issue—the use and effectiveness of ‘impolite’ ‘offensive’, and ‘taboo’ labels, which we leave for future research and debate.

References

[Dictionaries]
AHD. American heritage dictionary. Available online at https://www.ahdictionary.com/
Merriam-Webster’s learner’s dictionary. Available online at http://www.learnersdictionary.com/

[Other publications]


The Research on the Name Information in the Commercial Press Learner’s Dictionary of Contemporary Chinese

He Shan
Beijing Normal University, China
heshan2014@mail.bnu.edu.cn

Abstract
Cultural information is an indispensable part of Chinese learner’s dictionary; however, the definition of cultural information is still not clear, which involves difficulties for the process of dictionary compilation. Existing researches regard cultural information as an independent item which is dealt with separately in the dictionary but the treatment cannot solve all the problems, for example, should a real person’s name be included in the examples? Which surname-characters are suitable to be headwords? It is proposed in this paper that cultural information should be an indispensable element of learner’s dictionary. Name information is taken as an example in the discussion of how to integrate cultural information into different sections of CFL learner’s dictionary. By analyzing name information in the Commercial Press Learner’s Dictionary of Contemporary Chinese, the effects of the absence or improper use of name information are identified as: (1) examples cannot help dictionary-users obtain the meaning of headwords; (2) that examples are not natural which prevents users from obtaining the usage of headwords; (3) that examples are tedious which therefore hinders stimulating user’s interests.

Keywords: name information, cultural information, learner’s dictionary

1. The context of this study
1.1. Culture and language
Everybody has the sense of culture, but there is no easy way to answer the question “what is culture”. We may take culture as the way of life of a specific community which involves how they think, how they speak, how they act, etc. Instead of offering a precise definition of culture, some scholars simply describe culture as a broad concept that embraces all aspects of the life of man. (Seelye 1984:26)

Although the concept of culture is vague, “who are they” is defined by the culture of their community. There are a great number of scholars who have tried to describe the concept more specifically by indicating a few broad categories. For example, a threefold classification was suggested by Hammerly (1982) which distinguishes factual culture, behavior culture and achievement culture. In China, there is a fourfold classification which contains material culture, behavior culture, system culture and spirit culture. (Zhang Dainian et al. 2004) However, those classifications cannot solve the problem of the range of cultural items.

There are also many researchers have tried to reduce the nature of the concept of culture to manageable proportions. Zhang et al. (2004) introduced Chinese culture systematically in 18 categories:

1) The historical and geographic environment of Chinese culture
2) The economic base of Chinese culture
3) The social and political stricture of Chinese culture
4) The development of traditional Chinese culture
5) The multi-cultural integration in China
6) Chinese language
7) Chinese science and technology
8) Chinese education
9) Chinese literature
10) Chinese art
11) Chinese historiography
12) Traditional Chinese ethics and morals
13) Religions
14) Philosophy
15) Characteristics of Chinese culture
16) Spirit of Chinese culture
17) Traditional Chinese value system
18) Modernization of traditional Chinese culture

From this catalog we can see almost everything in our life is about culture. Among all these aspects, language is particularly important because it reflects culture systematically. It is not only an important feature of human culture, but also the carrier of culture and an indispensable component of it. The relationship between culture and language is a key of our discussion.

Wilhelm von Humboldt (1838:46) said, “The mental individuality of a people and the shape of its language are so intimately fused with one another……Language is, as it were, the outer appearance of the spirit of a people; the language is their spirit and the spirit their language; we can never think of them sufficiently as identical.” Language is not an isolated system but a path, through which we can get to know a people and their life. The cross-culture communication will not be completed only with the help of grammar and vocabulary, because communicating is all about the sharing of information, emotion and thoughts which are expressed in a specific cultural way, only by knowing the cultural background, both sides of communication can fully understand each other without any misunderstandings.

1.2. Cultural contents in the foreign language teaching
Culture determines the way we understand the nature, society and ourselves, so people from different cultures have different standpoints, if one side or both sides of cross-culture communication cannot empathize, there will be misunderstandings even conflicts. Therefore, great importance has been attached to the culture in foreign language teaching or culture teaching. Humboldt (1838:60) said: “To learn a foreign language should therefore be to acquire a new standpoint in the world-view hitherto possessed, and in fact to a certain extent is so, since every language contains the whole conceptual fabric and mode of presentation of a portion of mankind.” Watson-Gegeo (1988) also held that “When we learn a second language, we are learning more than a structure of communication; we are also learning social and cultural norms, procedures for socialization…”

The goals and contents of culture teaching have been discussed in many articles and researches, such as “culture with a small c” which means way-of–life rather than “Culture with a capital C” which means formal culture should be involved in foreign language teaching. Stern (1992:218) described the goals of culture teaching as proficiency, cognitive, and affective objects. In his cultural syllabus, 6 aspects of culture teaching are as following: 1) Places, 2) Individual persons and way of life, 3) People and society in general, 4) History, 5) Institutions, 6) Art, music, literature, and other major achievements. Zhang Zhan-yi et al. (1994:113-123) summarized the goals of culture teaching as cultural understanding and cross-culture communication. They also listed some points of cross-culture-communication failure, including temporal words, color words, addressing, taboos, and euphemism… Wei Chunmu et al. (1992:54-60) listed 114 cultural items which should be taught, but the term of cultural syllabus was not mentioned.

(1) Physiological needs: food and drink, greeting, shopping…
(2) Interpersonal Relationship: address, dating, invitation, helping…
(3) Emotion and attitude: modesty, compromise, apology…
(4) Views and opinions: comment, suggestion, discuss…
(5) Privacy: age, income, personal issues…

…..
After more than 20 years, there is still no official cultural syllabus of Chinese teaching which causes some difficulties in TCFL. Zhang Ying (2009:93-100) pointed out that it is very urgent to formulate a syllabus of Chinese culture. The general principle—teaching cultural content should help to overcome prejudice and stereotypes which can prevent learners from coming to terms with the reality of the target culture (Stern 1992:215) also applies to the syllabus compiling.

1.3. Culture information in the learner’s dictionary

There are “CULTURE” column in the Oxford Advanced Learner’s English-Chinese Dictionary. In the CULTURE column, British culture and American culture are introduced in detail. The appendix of The Commercial Press Learner’s Dictionary of Contemporary Chinese (CPLDCC) contains “grammar key points of contemporary Chinese”, “the most common 200 surnames in China” and other information about history, geography of China. It is believed that all those information will help Chinese learners to know more about China and do better in their daily communication with Chinese people.

1.3.1. The necessity of including cultural information in the learner’s dictionary

As a useful tool for FL learners, learner’s dictionary provides information of the sense of words and grammar in detail. The characteristic that helps users with decoding and encoding makes learner’s dictionary an important supplement of classroom teaching. What needs to be paid attention to is that inputting is the base of both decoding and encoding, so the quality of inputting which depends on the contents of the learner’s dictionary is the key. In fact, cultural information is determinative of the quality of a dictionary, but that it is everywhere makes it an element which is easy to be neglected.

When talking about contents of a dictionary, entries may be the first one that you come up with, because they are the basic information units in the dictionary which can be separated into formal comment and semantic comment. The former contains lemma, morphology, pronunciation; the latter contains pragmatic-semantic information and example. (Hartmann 2001:60) In the researches of Chinese dictionary, semantic comment has been attracted much attention, the accuracy and comprehensiveness is what matters the most. However, more and more compilers and scholars stress that cultural content should be an indispensable part of learner’s dictionary. (Hu Mingyang et al.1982:146-147; Li Yimin 1989; Li Ergang 2002:171-175; Zhang Hong 2009)

1.3.2. The presenting of cultural information in the learner’s dictionary

He Jianing (2001) suggested that English-Chinese dictionary should include cultural information in macrostructure and microstructure. In the level of macrostructure, cultural information can be presented though entry, inset and appendix; in the level of microstructure, it can be presented in definition, example, gloss, figures, column and cross-reference. Other researches were also conducted within a similar framework, such as Wei Xiangqing (2005), Wu Ping (2006), Yan Lizhi (2006), Miao Lanbin (2013) and Wu Jiangshuai (2014).

The basic framework is valuable, but cultural information in those studies are regarded as the main subject of an entry, a column or some other units which means it is dealt with independently which goes against what we have discussed about the nature of culture—it is about everything. The awareness of culture is inside every native speaker, so every speech and behavior contains cultural information, which is essential for cross-culture understanding and communication. We believe it is better to take cultural information as an element that involved in every aspects of learner’s dictionary. In the past, entries are considered as the flesh; similarly, cultural information can be compared to the blood which relates to the well-functioning of each part of a dictionary.

2. Research object
There is no widely accepted cultural syllabus which makes dictionary compilation difficult, but it is not easy to research on all the cultural information in the learner’s dictionary, because it is a large systematic work, so we take name information in the Commercial Press Learner’s Dictionary of Contemporary Chinese as the research object. The concept of name information refers to names of Chinese people (full names, surnames and forenames) and appellations which are related to names. Before discussing about the research object, the importance of name information has to be indicated by dealing with the following issues: (1) the cultural value of name information; (2) the necessity of including name information.

2.1. Name information as a cultural item
Name information is regarded as both linguistic and cultural symbol which has distinct features that reflects Han people’s value, morality, philosophy, and the relationships in Chinese family and society. The distinct features and cultural value of name information have been revealed in many researches: Zhang Hui (2005) concluded the following characteristics: (1) the great importance Han people attach to names, (2) the aesthetic taste showed in names, (3) Confucianism and Daoism thoughts in names, (4) the rhetorical means used in names, and (5) taboos, which can be illustrated with examples: when a new-born baby was given a name, he/she would not be named after their parents or grandparents to show their respect to the seniority in the family, babies of the same generation might share a same character in their given names to show the status of their generation in the family and so on. Zhang Dianen (2007) and Wu Biyu (2008) illustrated the cultural meaning of Chinese names by analyzing the origin of surnames and the ways of name-giving. For example, most parents would choose those characters that have optimistic meaning, beautiful shape, pleasant pronunciation to be the name of their child; some parents would decide the name according to the date of birth and the eight characters of a horoscope of the baby. It is not hard to see that name means a lot to Chinese people, though the names you can know much about their perspectives of clan ethics, their life value and their understanding of nature and human society. Huang Pengfei et al. (2012) indicated that the order of Chinese name is a symbol of “We Culture”, because Chinese surname is in front of the forename, whereas the English name is in the opposite way which makes it a symbol of “I Culture”. As an important cultural item, name information are suggested to be involved in CFL teaching, and be used to help with CFL learning.

2.2. Names information and CFL teaching and learning
Huang et al. (2012) pointed out that name information is one kind of valuable cultural resources which provide them with an easier way to get to know the nature of Chinese character. Awareness and knowledge of it will help CFL learners to understand Chinese culture further and better. Bai Zhaxia (2006) took giving CFL learners appropriate Chinese names as an effective way to create a good atmosphere for immersing them into Chinese language and culture.

The reasons why CFL learners should familiarize themselves with Chinese name information are as following: (1) being familiar with Chinese names is time-consuming, because the number of Chinese family names is numerous, the characters that can be used as given names are countless, (2) recognizing Chinese names information is difficult, because CSL learners cannot separate Chinese names in the text by Capital letters, spaces or different font as in other languages, (3) name information is a cultural item that CFL learners are interested in which was showed in survey and interview conducted by Wang Ruihua (2011).

3. Name information in the CPLDCC
We examined name information in the CPLDC to discuss the methods of integrating cultural information into CFL learner’s dictionary. The main deficiencies are as following: (1) name information is incomplete; (2) name information is insufficiently used; (3) the quality of name information is limited.

3.1. name information is incomplete

3.1.1. Special reference sections
The term “special reference sections” refers to insets, notes, columns, appendixes and other reference sections which are designed for special items.

There is no inset, index or column which introduces name information, but some notes and synonym columns illustrating the usage of kinship terms can be found in the CPLDCC which is irrelevant with name information.

3.1.2. Headwords
65.5% of characters in the Appendix 8 are included in the dictionary, but there is no sense of “family name” in those entries; the rest 26.5% of characters which are usually used as family names cannot be found in the dictionary. (He Shan 2014) The criteria of selecting surname-characters and arranging senses are not clear—whether a surname-character being included is not decided by their rankings or the number of their senses.

3.1.3. Senses
The sense “can be used as a surname” can be found in the entries of some surname-characters, but the criteria of arranging senses are not clear—neither frequency, polyphony, nor the number of senses is considered.

3.1.4. Examples
There are only 12 surnames (王, 张, 李, 马, 刘, 司马, 诸葛, 陈, 赵, 徐林, 西门) in the examples; over 93% of them are “张”, “王” and “李”, among which the frequency of “王” is 273 times, consisting 66.7%, which is far more than that of others. Family names are not only few in numbers, but also few in forms. For example, “王” appeared in the form of “小王” for 169 times, which is over 40% of totality; the frequency of “小王”, “老王” and “王老师” is more than half of the sum. (He, 2014)

Full names of famous person include “李白”, “杜甫”, “屈原”, “鲁迅”, “巴金”, “张大千”, and there are 10 Chinese names in the text of the CPLDCC. Other full names and forename-related addresses are “安娜”, “小红”, “王华”, “小明”, “小东”, “小文”, “张明”, “张文”, “张英”, “张全英”, “李光”, “小华”, “王老师”, “张老师”, most of them are two-character names, and there is only one three-character name. The proportion does not correspond to reality.

3.2. Name information is insufficiently used
Lots of cultural information is presented in the appendixes, such as the table of Chinese previous dynasties era, China’s administrative division table, the table of major countries and regions in the world, the table of “Tiangan” and “Dizhi”, the 24 solar terms table, the table of China’s nationalities, kinship terms table, table of common measure words and punctuation table. The Appendix 8 is a table of the most common 200 surnames, but there is no index or cross-reference that leads to the appendix which means there is little chance that it is referred to, so a great deal of name information cannot be used effectively by dictionary-users. Therefore, it is necessary to build links between them to make the contents easy to find, such as “see also” marks.

3.3. The quality of name information is limited
Szende (1999:217-219) thought examples might not only reflect the linguistic competence of a community, but also its culture and way of life. However, existing examples are focusing on illustrating headword’s lexical and grammatical meaning. (He 2001) Examples should
concern more about cultural depth, vividness and interestingness which corresponds to our study that the quality of name information is directly related to examples’ validity. Improper name information will affect examples in two ways: (1) cannot help dictionary-users obtain the meaning of headwords; (2) the lack of natural style will prevent users from obtaining the information of context; (3) tediousness examples are not helpful in stimulating user’s interests.

3.3.1. Cannot help dictionary-users obtain the meaning of headwords

(1) 【开会】老李开了三个小时的会。P393
(2) 【揩】小王把桌子揩干净了。P394
(3) 【筐】老李正在编竹筐。P415
(4) 【辣椒】四川人特别喜欢吃辣椒。P421

The positions or occupations of “老李(Lao Li)” and “小王(Xiao Wang)” are unknown, so the relationship between the subject and the action is not clear. Addresses which imply the positions or occupations of subjects rather than useless name information will make the meaning of headwords will be more understandable. In example (4), “四川人” is better than a specific name or address. So name information should be used prudently in examples.

(5) 【齐名】他是与巴金齐名的大作家，是著名画家，跟张大千齐名。P549
   →老舍/茅盾是与巴金齐名的大作家，齐白石是著名画家，跟张大千齐名。
(6) 【文学家】他是文学家，也是教育家。P734
   →韩愈/叶圣陶/苏格拉底(Socrates)是文学家，也是教育家。

The reference of subject “他” is ambiguous which cannot provide detailed information for dictionary user to understand the lexical meaning of the headwords. Different from a personal pronoun, real names will make【齐名】and【文学家】better understood.

(7) 【不朽】大诗人李白写下了大批不朽的诗篇。P57
(8) 【端午节】传说端午节是为了纪念中国古代的爱国诗人屈原（Qu Yuan）。P178
(9) 【根据】电影《祝福》是根据鲁迅的小说改编的。P247
(10) 【文学家】鲁迅是中国著名的文学家。P734
(11) 【祖宗】孔子是儒家学说的老祖宗。P894

Real names in these examples provide complete context which enriches the meaning of sentences and makes their understanding of headwords much more precise.

3.3.2. That examples are not natural prevents users from obtaining the usage of headwords

(12) 【搭配】小王和小李搭配参加羽毛球比赛。P123

In this example【搭配】means “小王” goes together with “小李” in the badminton match, but from name information in the example the point that makes them a team cannot be inferred, so the usage of the headword is not properly showed.

(13) 【之类】现在像小王、小张之类的成功人士并不太多。P925

When taking somebody or something as an example, the headword【之类】is used. What should be paid attention to is that somebody or something is usually very typical which cannot be showed by name information in this sentence. It is better to use well-known names.

(14) 【叫】他叫张三︱我们这儿没有叫李四的。P357

It is a common sense that the term “张三李四” is used to refer to an unidentified person. The definition of “张三李四”in The New Chinese-English Dictionary (Wu Guanghua, 2003) is: “this one and that one; a Mr. X or Y; any Tom, Dick or Harry; any man in the street; anybody; Brown, Jones and Robinson; either one or the other; Mr. so-and-so; somebody...” Therefore, the lack of authenticity will not help users obtain the right usage of the headword.

(15) 【尊称】王先生被大家尊称为“王老”。

P896
The second sentence explains that “先生” is a respectful form of address which is used in the first sentence to explain another respectful form of address—“王老”. Does it mean “王老” is a more respectful form? The example is improper which may confuse dictionary users.

Name information is not just about somebody who does not matter being the subject of a sentence. The background knowledge and context attaching to the name or address is helpful for proving more complete reference which is for the benefit of dictionary users.

3.3.3. That examples are tediousness hinders stimulating user’s interests

Every experienced CFL teacher will confirm that interesting and vivid examples will motive students in the classroom, so it can be inferred that the interestingness and vividness of examples are more significant for paper dictionary.

<table>
<thead>
<tr>
<th>Example</th>
<th>Chinese</th>
<th>Pinyin</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>(16)</td>
<td>【潦草】(liao2 cao3)小王写字很潦草，谁都看不清楚。</td>
<td>(zhao3 cao3)</td>
<td>Xiao Wang's handwriting is very untidy, nobody can read it.</td>
</tr>
<tr>
<td>(17)</td>
<td>【了不得】(liao3 bu4 de0)可了不得，小王的孩子丢了。</td>
<td>(liao3 bu4 de0)</td>
<td>He is really great, Xiao Wang's child is lost.</td>
</tr>
<tr>
<td>(18)</td>
<td>【了解】(liao2 jie3)领导让我来这儿了解一下儿小王的情况。</td>
<td>(liao2 jie3)</td>
<td>The leader asks me to come here to find out about Xiao Wang.</td>
</tr>
<tr>
<td>(19)</td>
<td>【咧】(lie3)小王今天很高兴，一直咧嘴笑。</td>
<td>(lie3)</td>
<td>Xiao Wang is very happy today, he can't stop smiling.</td>
</tr>
<tr>
<td>(20)</td>
<td>【列】(lie4)小王被列为参加会议的代表。</td>
<td>(lie4)</td>
<td>Xiao Wang is listed as a representative to attend the meeting.</td>
</tr>
<tr>
<td>(21)</td>
<td>【再说】(zai4 shuo1)小王很想参加这次英语竞赛, 再说她的英语也很好, 就让她去吧。</td>
<td>(zai4 shuo1)</td>
<td>Xiao Wang wants to participate in this English competition, again, her English is also very good, let her go.</td>
</tr>
<tr>
<td>(22)</td>
<td>【在，】(zai4)小王在吗？——他不在，上课去了。</td>
<td>(zai4)</td>
<td>Is Xiao Wang here? —— He is not here, he went to class.</td>
</tr>
<tr>
<td>(23)</td>
<td>【在场】(zai4 chang3)事故发生时小王正好在场，看得很清楚。</td>
<td>(zai4 chang3)</td>
<td>When the accident happened, Xiao Wang was just present, he saw it clearly.</td>
</tr>
<tr>
<td>(24)</td>
<td>【在于】(zai4 yu2)小王的成功在于坚持。</td>
<td>(zai4 yu2)</td>
<td>Xiao Wang's success lies in perseverance.</td>
</tr>
<tr>
<td>(25)</td>
<td>【在，】(zai4)小王出生在1985年。</td>
<td>(zai4)</td>
<td>Xiao Wang was born in 1985.</td>
</tr>
<tr>
<td>(26)</td>
<td>【在，】(zai4)小王在音乐上很难有才能。</td>
<td>(zai4)</td>
<td>Xiao Wang has little talent in music.</td>
</tr>
<tr>
<td>(27)</td>
<td>【攒】(zan3)小王虽然工作了好几年, 但是没攒下多少钱。</td>
<td>(zan3)</td>
<td>Xiao Wang worked for many years, but he didn't save much money.</td>
</tr>
</tbody>
</table>

Example (16)-(20) are found in Page 446-447, and example (21)-(27) are form the Page 874 to 875; all of them contain the same name information “小王” which appears so frequently that could make users feel bored. If real names or made-up full names are regarded as obstacles of reading, at least some other surnames besides “王” should be used.

It is not only a technical problem of writing or editing name information in the examples, but also a problem of perspective: firstly, it cannot be denied that “小/老+ surname” are a commonly used way of address, but there is no illustration of their pragmatic information in the dictionary, because users are subjectively assumed that they fully mastered the usage of them. Secondly, the chance them are used in everyday life is not that big, especially for school students who constitute a great part of CFL learners. Because of the repetition of name information, the style of examples is not vivid, less life-oriented and practical. This will not contribute stimulating learner’s interests.

4. Conclusion

Name information and other cultural information are helpful for Chinese learners to learn the Chinese language and culture better; and appropriate name information will make examples more real, natural and vivid. Although the problem of including cultural information in the learner’s dictionary has been paid much attention to, most of existing researches took it as the main subject of entries, insets, notes and columns, but name information as an indispensable element in the whole dictionary are ignored. The deficiencies of including and making use of name information were discussed in this paper, and the following suggestions are put forward:

(1) The compiler of CSL learner’s dictionary should concern more about Chinese name information and other cultural information such as places, festivals, food and so on.
(2) It is important to set up a series of criteria of selecting and editing material of a dictionary which is clear and scientific.

(3) The methods of adding name information in the examples should be more scientific and discreet to make the lexical and grammar meaning better presented, and take the advantage of it to make the examples more natural and interesting.

Reference:
Bai Zhaoxia. 2006. Name Culture and TCFL. Journal of Yunnan Normal University. 4(4):61-64
He Shan. 2014. A Research on the Name Information in the Commercial Press Learner’s Dictionary of Contemporary Chinese. The 7th Seminar of Graduate Students in the Beijing Area of Teaching Chinese as a Foreign Language Processing. 44–50. Peking University, Beijing, China.
Wu Ping. 2006. On the Survey Concerning the Necessity to Integrate Cultural Information into English Learner’s Dictionaries——An Empirical Study. Journal of University of Shanghai for Science and Technonology. 28(3):61-66


Zhang Zhanyi, Bi Jiwan. 1991. How to Understand and Reveal Cultural Elements in CFL Teaching. language teaching and linguistic studies. 4:113-123
Two Innovations in Pedagogical Dictionaries

Geoff Toister,
GT-Lex: www.gt-lex.weebly.com, Israel
gat178@012.net.il

1. Introduction

Before I begin to discuss the subject of this paper I should like to outline the trajectory of my professional career. I am a retired teacher of English as a foreign language, with 26 years of classroom experience. The major part of those years was spent in teaching intermediate and upper-intermediate students preparing for the Israeli matriculation exam, known as the Bagrut. Some years after retiring from teaching, I became a lexicographer. Since then I have compiled entries for a prominent dictionary publisher, disambiguated entries for another and published a hybrid dictionary and vocabulary development workbook entitled Bagrut Lexicon. I will have more to say about this later.

2. Rationale

My motivation for writing Bagrut Lexicon was that I always felt that I was giving my students less than what they needed to develop their vocabulary. My teaching career spanned the latter years of the audio-lingual method in ELT and the early years of the communicative method, but when it came to the teaching of vocabulary, I felt that the materials writers and coursebook publishers on whom I depended were always in two minds. Their products always seemed to alternate between favouring incidental acquisition of words embedded in reading texts and explicit instruction in discrete lexical items for use in a specific situational context. These days, learning a foreign language is still popularly conceived as a matter of learning how to manipulate its grammar. Some smartphone apps recently released by reputable ELT publishers indulge this perception. However, I long ago decided that if English teachers are required to teach dictionary skills, they may reasonably expect lexicographers to assume some of the responsibility for vocabulary teaching. How might this be achieved today and what role might the lexicographer play? I have tried to provide two practical answers to these questions. The first of these is the hybrid that I have already mentioned.

3. Road Map

I would like now to review the various types of resource books for vocabulary development. I will then describe the method used for developing Bagrut Lexicon and detail some of its salient features. This first section of my presentation will be followed by a preview of a second lexicon-cum-vocabulary development workbook which is currently in the final stages of editing and which deals with multi-word expressions (MWEs) or so-called ‘lexical chunks’.

4. Resources for Vocabulary Development

4.1 Many learner’s dictionaries contain usage notes either within the entry or in marked off boxes. The notes usually focus on synonyms or the collocational idiosyncrasies of a particular lemma. Thus they supply supplementary information about the headword, but not practice in its use.

4.2 The Oxford Wordfinder takes a different approach. It is a meaning-to-word dictionary whose purpose is to facilitate language production. The entries do not contain practice exercises, but the OUP website provides free photocopiable worksheets. It is a pity that these
are not mentioned in the book’s front matter because they somewhat compensate for the absence of practice exercises between its covers. The *Longman Language Activator* is said to be similar. While it is gratifying that these two major dictionary publishers have taken up the challenge of vocabulary development, it is unrealistic to expect impecunious students and high-school pupils to purchase both a dictionary and a meaning-to-word lexicon.

4.3 A more useful kind of learning tool can be found in various dictionary-skills manuals linked to specific dictionaries but published separately. One such handbook is Bloch’s *Use Your Dictionary*, which was “tailored specifically to the special features, conventions and entry arrangements of the ADVANCED LEARNER’S DICTIONARY OF CURRENT ENGLISH, written by A.S. Hornby …” (p.7.). It contains a wealth of drills delving into every aspect of the entry structure with examples drawn from the whole range of entries in the dictionary. While the purpose of the book is to explain how dictionaries work, the drills afford excellent opportunities for incidental learning and can be seen as prototypes of the type of exercise that promotes vocabulary acquisition. A drawback is the expense of purchasing the manual in addition to the dictionary it goes with.

4.4 From a teacher’s point of view, none of the three foregoing types is entirely satisfactory. Teacher-designed exercises are an alternative. Their disadvantage is that they are extremely time-consuming to produce and the effort is only worthwhile if all members of a class possess the same dictionary.

4.5 A more deliberate approach to integrating lexicography with pedagogy can be found in Phillips’ *Vocabulary Dictionary and Workbook*. It is a lexicon of 2,856 words for academic purposes together with sets of word games for vocabulary practice. The truly innovative feature of the book is its verso-recto page setup (see Figure 1). The entries are divided into alphabetic groups, called ‘chapters’, consisting of 12 lemmas and each double-page spread is devoted to a single chapter, with the lexicon always on the verso and the word games always on the recto. There are two benefits to this setup. Firstly, it sends a clear message to the learner that merely looking up a word is not sufficient to guarantee retention of its meaning. Secondly, by juxtaposing the two entities on the same folio, it simplifies and encourages repeated look-ups during practice. However, the games themselves disappoint as there are only 12 test items for the 12 lemmas. This reduces the element of risk to the point that the games do not challenge the intellect.

5. Development of *Bagrut Lexicon*

5.1 Lack of appropriate materials
Returning now to *Bagrut Lexicon*, this project was stimulated by the twin observations that learner’s dictionaries do not contain vocabulary practice exercises and that few ELT materials are explicitly devoted to vocabulary development. The majority are textbooks or books concentrating on grammar, reading comprehension, writing or listening. Of all the ELT books in a certain pedagogical library that I visited, only 2% were found to be expressly dedicated to the teaching of words. How then are learners expected to expand their vocabulary? And which words are to be taught?

5.2 Methodology
To answer the latter question, a corpus was constructed using *TextSTAT*, a freely downloadable application. The corpus consisted of 58 homogeneous matriculation-level
reading comprehension texts, and yielded 21,000 tokens and 5,000 types. In the process of constructing the corpus, a downward trend was noted in the incidence of novel types per additional text (see Figure 2). This seems to indicate that matriculation reading texts constitute a particular genre and that the vocabulary of the genre tends to be repetitive. The 5,000 types are therefore the most relevant candidates for a learner's lexicon.

![Novel Types per Additional Texts](image)

Figure 2

All function words, proper nouns and numbers were deleted from the list of 5,000 candidates. Poorly dispersed words and cognates were also removed. The remaining words were manually lemmatised and POS tagged. In the next stage of selection, words listed in the lexical syllabus for primary schools (State of Israel, 1988) were also discarded, thus leaving a final list of 2,000 words.

These 2,000 words became the headword list for the lexicon, which was compiled using TLex Dictionary Compilation Software. The lexicon was designed along the lines of a semi-bilingual learner's dictionary. Specifically, each entry consists of a headword, grammatical information, definition, usage example and a translation equivalent. The usage examples were adapted from the 58-text corpus or various corpora freely available on the Internet. Multiple senses were not provided unless they were evidenced in the original corpus. Other features include phonemic pronunciation guides, usage labels, phrasal verbs with separate headword status and alternative spellings. The result is a unidirectional specialised lexicon for decoding L2 that is fully integrated with a vocabulary acquisition workbook.

5.3 The innovation

There is nothing novel in this well-established format. However, Bagrut Lexicon does break new ground by coupling the semi-bilingual lexicon with Phillips’ verso-recto page setup and by providing practice exercises that genuinely challenge intermediate-to-advanced learners. Meaningful practice is established by three means. Firstly, the lexicon divides entries into alphabetic groups of 20 lemmas and provides 30 practice items per group. This increases the risk of a false answer and thereby obliges the user to return to the lexicon. Secondly, the 30 practice items are sub-divided into six graded exercise types. The fifth type is particularly difficult as it requires 5 choices from 20 possibilities. In the editing stage, the exercises were confirmed to have unambiguous answers.

6. Literature review: vocabulary development

We come now to the second innovation that I am here to present, and I shall begin with a brief review of some of the relevant literature. I mentioned earlier that, in the past, materials writers and coursebook publishers were in two minds about how to teach vocabulary. Today
there is wide agreement on the centrality of vocabulary development to language acquisition. This agreement has found expression in some coursebooks and in 2003 Ranalli gave Longman’s *Cutting Edge* a three-star rating, the second-highest, (4.4, p.33) for actually including a 64-page mini-dictionary within its covers and for providing training in dictionary skills. He remarks that the mini-dictionary “is similar in format to the dictionaries produced by the Longman publishing house, but the headwords consist only of the items appearing somewhere in the students’ book or the class cassette.” To my mind, this is high praise as it highlights the publisher’s commitment to actively promoting vocabulary development.

On the other hand, teacher trainers seem to be less committed, as can be seen from Philadelphia University’s syllabus for a course in ELT methods (2013), which includes the audio-lingual and communicative approaches, amongst others, but omits the lexical approach. Given that the lexical approach predates that syllabus by at least fifteen years, one wonders whether its exclusion from the syllabus is due to a lack of ready-made teaching materials.

Traditionally, ‘vocabulary’ was understood as a set of “single and fully autonomous lexical units” (Criado, 2009). Today, however, it is recognised that “much of lexis consists of sequences of words which operate as single units” (Schmitt, 2000). This raises the practical problem of identifying those sequences. One solution was offered by Martinez and Schmitt in their 2012 journal article, *A Phrasal Expressions List*. The authors were able to produce “a list of the 505 most frequent non-transparent multiword expressions in English”. Without implying anything as to the validity of their methodology, it must be noted that the list tends to lump polysemous MWEs into a single lexical unit, and that more than a few of those items are restricted to British usage. Nevertheless, it does provide a solid foundation for constructing a database of prefabricated MWE entries suitable for integration into pedagogical materials.

7. Status of the MWE Database

7.1 Entry structure

Compilation of the entries for this database was recently completed. Unlike *Bagrut Lexicon*, the database was designed along the lines of a monolingual learner's dictionary. Specifically, each entry consists of a headword, grammatical information, a definition and three to six usage examples. Headwords with like senses are cross-referenced and the usage examples are adapted from or inspired by the British National Corpus. The database additionally provides full inflection information, guidance with syntax and synonyms wherever they apply. Frequency information is also available and the design language can be changed at will. Although translation equivalents are not provided, the *TLex* software allows for adding them in virtually any language, thus making the database fit for the needs of any source-language group. A typical entry appears in Figure 3 below:

<table>
<thead>
<tr>
<th>MIGHT AS WELL</th>
<th>[Also: might just as well, may as well, may just as well]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammatical function:</td>
<td>adverbial</td>
</tr>
<tr>
<td>How to use this word:</td>
<td>This expression is always followed by a verb in the base form.</td>
</tr>
<tr>
<td>Type of expression:</td>
<td>phrasal idiom</td>
</tr>
<tr>
<td>Sense 1</td>
<td>Definition: with equal convenience</td>
</tr>
<tr>
<td>Examples:</td>
<td>• If I have to wait anyway, I <strong>might as well</strong> spend the time reading.</td>
</tr>
<tr>
<td></td>
<td>• [Said at a supermarket checkout] You only have a few items. You <strong>might as well</strong> go ahead of me. I've got a trolley-full.</td>
</tr>
<tr>
<td>Sense 2</td>
<td>Definition: with equal truth; with equal validity</td>
</tr>
<tr>
<td>Examples:</td>
<td>• Astrologers <strong>might just as well</strong> claim that the earth is flat.</td>
</tr>
<tr>
<td></td>
<td>• The translation was so free that there <strong>might as well</strong> be no original text.</td>
</tr>
</tbody>
</table>
7.2 Searching

It is envisaged that the database would be incorporated into an electronic dictionary and that students would look up MWEs in one of three ways. Firstly, straightforward look-ups can be performed on either the full canonical form of an MWE or on any one of its component words. In the latter case, several results are likely to be presented. Secondly, a search can be performed from a synonym to an MWE. The result of a synonym to MWE search for the word ‘occur’ appears in figure 4 below.

<table>
<thead>
<tr>
<th>COME ABOUT</th>
<th>COME ABOUT</th>
<th>COME ABOUT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>In other words:</strong></td>
<td>comes about, coming about, came about</td>
<td>comes about, coming about, came about</td>
</tr>
<tr>
<td><strong>Grammaratical function:</strong></td>
<td>verbal</td>
<td>verbal</td>
</tr>
<tr>
<td><strong>Type of expression:</strong></td>
<td>two-word verb</td>
<td>two-word verb</td>
</tr>
<tr>
<td><strong>Definition:</strong></td>
<td>happen in a series of unplanned stages</td>
<td>happen in a series of unplanned stages</td>
</tr>
<tr>
<td><strong>Examples:</strong></td>
<td>• Rising sea levels have come about as a result of global warming.</td>
<td>• Rising sea levels have come about as a result of global warming.</td>
</tr>
<tr>
<td></td>
<td>• It is important to recognize how fear comes about and how to cope with it.</td>
<td>• It is important to recognize how fear comes about and how to cope with it.</td>
</tr>
<tr>
<td></td>
<td>• Change will not come about by chance, but we can bring it about by our own efforts.</td>
<td>• Change will not come about by chance, but we can bring it about by our own efforts.</td>
</tr>
<tr>
<td><strong>TAKE PLACE</strong></td>
<td><strong>TAKE PLACE</strong></td>
<td><strong>TAKE PLACE</strong></td>
</tr>
<tr>
<td><strong>In other words:</strong></td>
<td>occur, take place</td>
<td>occur, take place</td>
</tr>
<tr>
<td><strong>Grammaratical function:</strong></td>
<td>verbal</td>
<td>verbal</td>
</tr>
<tr>
<td><strong>Type of expression:</strong></td>
<td>two-word verb</td>
<td>two-word verb</td>
</tr>
<tr>
<td><strong>Definition:</strong></td>
<td>happen</td>
<td>happen</td>
</tr>
<tr>
<td><strong>Examples:</strong></td>
<td>• Since 1968 many changes have taken place.</td>
<td>• Since 1968 many changes have taken place.</td>
</tr>
<tr>
<td></td>
<td>• The next meeting will take place on Monday, 9th April.</td>
<td>• The next meeting will take place on Monday, 9th April.</td>
</tr>
<tr>
<td></td>
<td>• The festival takes place once a year, usually in October.</td>
<td>• The festival takes place once a year, usually in October.</td>
</tr>
</tbody>
</table>

8. Future work on the MWE database

8.1 Two further developments are envisaged for this project. Firstly, as suggested by McAlpine (2001), it may be possible to look up MWEs using a synonym and a hyponym as nested search criteria. To facilitate this, an additional field to be populated from a hierarchy of superordinate topics can easily be added to the XML database. The difficulty with this suggestion is in determining at which level in the hierarchy a given MWE headword ought to be indexed. For example, is the appropriate hyponym for the headword about to ‘when?’ or the higher level ‘time and place’, as suggested by the Oxford Wordfinder (p.xix.). The benefit of implementing this suggestion is the assistance it might provide the learner in performing writing assignments.

8.2 The second possible development is to supplement the 505 headwords with sets of vocabulary development exercises, as in Bagrut Lexicon. Lewis (1997) provides thirteen examples of suitable exercise types arising from the lexical approach. Figure v below illustrates how the MWE database might be used to construct exercises on the pattern of Lewis’ Type Nine. All of the MWEs that the exercise calls for appear in the database. The difficulty with this suggestion is in ensuring that all the headwords are treated in at least one exercise.
“Complete the following by adding one word. Only one word is possible in each case. Make sure you know the equivalent expression in your own language.” (Lewis, p.265)

1. She became a widow a month ago and is still finding it difficult to come to … with her husband’s death.
2. We managed to get under cover … as it began to rain.
3. For the … of greater safety, please keep your seatbelt fastened throughout the flight.
4. The bus shelter collapsed under the weight of the snow, but in any … it was not designed to be permanent.
5. All the important branches of physics are now seen as engineering subjects in … own right and are often taught without very much of the underlying physics.

Figure 5

Figure vi below illustrates how the MWE database might be used to construct exercises on the pattern of Type D in Bagrut Lexicon. The advantage of this exercise type is that it is synchronised with the alphabetical order of the entries in the database, thus allowing the publisher to utilize a verso-recto setup.

D. Match the two parts of the sentence.

<table>
<thead>
<tr>
<th>PART I</th>
<th>PART II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Short of calling the doctor,</td>
<td>a were shown to have structural defects.</td>
</tr>
<tr>
<td>2 There are many short term</td>
<td>b I had started my PhD.</td>
</tr>
<tr>
<td>3 Many of the old buildings</td>
<td>c and we just couldn’t shut her up.</td>
</tr>
<tr>
<td>4 She kept talking and talking</td>
<td>d advantages to be had.</td>
</tr>
<tr>
<td>5 I lost sight of the original reason</td>
<td>e I didn’t know what to do.</td>
</tr>
</tbody>
</table>

Figure 6

Finally, it would also be feasible to emulate the verso-recto setup of Bagrut Lexicon in an electronic coursebook. A simple and inexpensive and method for doing so would be as follows:

- Compose the exercises
- Tag all of the MWEs that each exercise calls for
- Export each set of tagged MWE entries to separate RTF files
- Hyperlink each exercise to the set of MWE entries required for its completion

A more comprehensive method would be as follows:

- Compose the exercises
- Attach a monovalent code to each set of exercises
- Attach a monovalent code to each exercise item
- Attach a monovalent code to each lexicon entry
- Export the MWE lexicon to an ODBC database
- For each set of exercises – tag and export the relevant entries to an HTML file
- Build a relational database that links exercise sets, exercise items, and entries in the lexicon
- Build a search facility that enables the user to find all exercises for a given entry
- Build a search facility that enables the user to find all entries for a given exercise

9. Conclusion
This brings us back to our starting point. I wish I could conclude this presentation by saying "Our research revealed that materials writers and coursebook publishers continue to produce language programmes that do not adequately support vocabulary development.” But as you have already seen, I have simply presented two eclectic solutions to a concrete problem. I would rather leave you with the impression that you have just heard an appeal for combining the methods of lexicographers with the insights of teachers. I believe that such partnerships can contribute powerfully to the success of foreign language learners and enhance the status of both professions.

Thank you for your attention. I will be happy to answer any questions now.

References
Nofal, Khalil. 2013. Course Syllabus, Method TFL. Philadelphia University, Faculty of Arts, Department of English.
A New Trend in Lexicography: Bringing Semantic Prosody into Learner’s Dictionary

Junsong Wang
Northwestern Polytechnical University
wjs198012@163.com
Weixia Cui
Xi’an International Studies University

Abstract
Semantic prosody is commonly understood as a form of meaning which is established through the proximity of a consistent series of collocates. Its primary function is the expression of the attitude of the speaker or writer towards some pragmatic situation. Undoubtedly, semantic prosody can provide learners with pragmatic information on how to use the words properly in a given context and should be incorporated into lexicon design. Most of the dictionaries for EFL Learners, however, are either ill-marked or even untouched with regard to semantic prosody since this unique phenomenon tends to remain hidden from the lexicographer’s naked eye. This paper aims to find a systematic way to bring prosodic information into the learner’s dictionary. For this purpose, three ways have been introduced to deal with prosodic information, i.e. displaying with usage label, illustrating with definition & example sentence and setting collocation column. Other key issues, such as how to deal with the register and probability features of semantic prosody in dictionary compilation, are also addressed in this paper. It is hoped that with the prosodic information provided in the dictionary, the users can have a better understanding of English lexicon.

Keywords: semantic prosody, learner’s dictionary, pragmatic information, lexicon design

Introduction
The concept of semantic prosody was introduced to the public in 1990s, and now it has become one of the most important concepts in corpus linguistics. Over the past twenty years, it has attracted the attention of scholars both within corpus linguistics and beyond. With great advances in theory and research methods, many attempts have been made to explore the potential applications of semantic prosody. So far, it has proven to be of great use in many fields, such as language research, language education, and, of course, dictionary compilation. However, “the revealing of semantic prosody in both English and Chinese dictionary-making is far from ideal” (Yu Ping-fang & Jing-chun Cai 2009:25). And the reasons behind this can be summarized as follows: First of all, semantic prosody, which tends to remain inherently hidden to the lexicographer’s naked eyes, are not explicit enough to be retrieved reliably through introspection (Partington 1998). Secondly, some dictionary-writers are not well equipped with up-to-date linguistic theories, which inevitably hinders their lexicographic performance. And finally, there is a lack of a systematic way of dealing with prosodic information in dictionaries. This paper focuses on the last point and aims to find a systematic and efficient way of bringing semantic prosody into learner’s dictionary.

What is semantic prosody?
The term semantic prosody was first introduced to the public by Louw (1993) but John Sinclair’s idea (1991:112) who found that “many uses of words and phrases show a tendency to occur in a certain semantic environment. For example, the word Happen is associated with unpleasant things——accidents and the like.” Inspired by Firth’s notion of phonological prosody, Sinclair coined the term and developed it in his later work (e.g. Sinclair 2004). Though the term has been widely used, semantic prosody is a wide concept which has been approached in many different ways. For example, Louw (1993:157) defines semantic prosody
as the “consistent aura of meaning with which a form is imbued by its collocates” or “a form of meaning which is established through the proximity of a consistent series of collocates” (2000:57). Partington (1998:68) defines semantic prosody as “the spreading of connotational coloring beyond single word boundaries”. Hunston and Thompson’s definition of semantic prosody is “the speaker or writer’s attitude or stance toward, viewpoint or feeling about, the entities and propositions that he or she is talking about” (2000:5).

Given the various perspectives taken towards semantic prosody, semantic prosody as a concept has been assigned a broad range of characteristics. Firstly, semantic prosody mainly functions to express speaker’s/writer’s attitude or evaluation. In most cases semantic prosody is realized in the form of positive (or favorable), neutral and negative (or unfavorable) connotations (Stubbs 1995). Secondly, semantic prosody tends to remain hidden from human perception and intuition. It is much less evident to the naked eye and can be retrieved reliably with the aid of corpus and computerized language-analyzing software. Another characteristic of semantic prosody is that it can be associated with grammatical principles. For example, as Louw (1993) observed, build up tends to reveal a positive semantic prosody when it is used transitively. Used intransitively, however, build up shows a negative prosody. Finally, semantic prosody is more of an attribute of a phrase than that of an individual word.

**Significance of bringing semantic prosody into lexicon design**
Semantic prosody provides learners with pragmatic information on how to use the words properly and should be incorporated into lexicon design. In recent years, the important role of semantic prosody in dictionaries has been widely recognized and many attempts have been made to bring prosodic information into lexicon design. However, most of the dictionaries for EFL Learners are either ill-marked or even untouched with regard to semantic prosody for various reasons. Unable to obtain prosodic information from dictionaries, learners are more prone to making the avoidable mistakes in language communication. For example, Wang Haihua & Tong-shun Wang (2005) found China’s English learners tend to overuse the positive collocates of the English word cause while the native speakers prefer to use the word in a negative environment. Professor Wei Nai-xing’s survey (2011:142-145) also showed that a considerable number of unusual collocations occur in Chinese learner’s English writing which go against the collocation norms and lead to prosodic clash. Obviously, the Chinese English learners do need guidance to improve their consciousness of semantic prosody. Thus, it is an urgent task for the lexicographers to provide prosodic information in the learner’s dictionaries, which may help the users have an easier access to the understandings of English lexicon.

**Providing prosodic information in learner’s dictionaries**
To provide information about the prosodic behavior of entry words in learner’s dictionaries, the following three ways are considered to be worthy of concern:

1. **Displaying with usage label**
A usage label is “the marking of a word or phrase as typical or appropriate in a particular context or language variety”. Usage labels are widely used in both general and specified dictionaries to provide specific information, such as currency or period, e.g. arch (aic); formality or register, e.g. inf(ormal); regionality or dialect, e.g. Am(ericacan), York(shire); technicality or subject filed, e.g. bot(anical); textuality or genre, e.g. poet(ic). (Hartmann & James 2000:15).

Semantic prosody provides pragmatic information of the lexical items and it can be presented by using labels as well. Firstly, position of the labels. To catch the user’s attention, the labels can be placed following the headwords whatever it is a word or a phrase. Secondly, description of semantic prosody. The dichotomy of semantic prosody (positive/favorable VS negative/unfavorable) proposed by Stubbs is brief but inadequate to reveal the specific information the words convey. Therefore, a combined label, which includes both the general
and specific information, may be a better choice. For instance, the phrase true feeling frequently co-occur with expressions like will never reveal, less open about showing, guilty about expressing, and it can be easily concluded that the general semantic prosody that is negative or unfavorable but further analysis shows us more about the speaker’s/writer’s attitude or stance, i.e. unwilling or reluctant to express the true feeling. Thus, the usage label for the entry True feeling can be marked as [Negative: reluctance]. Lastly, as for the form of the usage label, it is suggested that the full form should be used for users could get confused by some abbreviations. And to highlight the prosodic information, the usage labels may be bolded, italicized or colorized but the inconsistency between labels in the same dictionary should be avoided.

Compared to other ways of marking prosodic information, labeling has its own merits of being direct, brief and eye-catching. However, its limitations lies in the fact that prosodic labels could hardly provide sufficient contextual information and some users could not distinguish them from other pragmatic labels, such as laudatory and pejorative.

2. Illustrating with definition and example sentence
Prosodic information can also be indicated by using both definitions and examples, the former offers a description of the semantic prosody and the latter provides evidence and illustrations. A typical example (see Fig.1) quoted from COBUILD dictionary is given as follows:

Rife
ADJ If you say that something, usually something bad, is rife in a place or that the place is rife with it, you mean that it is very common. [v-link ADJ]
e.g. Speculation is rife that he will be fired.

Figure1 Entry Rife from COBUILD

In the above example, the word rife’s prosodic information is given in the definition and proved with an example sentence. The definition is given in the form of a complex sentence, in which the main clause defines the word rife in terms of meaning while the subordinate clause offers a detailed explanation of the specific situation where the word is used. As depicted in this case, the meaning of the word rife is common and it can be only used as v-link adjective and habitually co-occurs with something bad. Obviously, providing prosodic information in the definition is a good choice for it enables the learners to have a better understanding of where and how the word is used. But on the other hand, the number of the example sentences provided in the dictionary is usually small that it cannot meet the needs of the learners to internalize the prosodic knowledge about the word. Therefore, it is advisable to increase the number of example sentences and if possible display in the form of KWIC (key word in context). See Fig. 2

His early history and the record of his films seem rife with the required anxieties.
The two Prize winners say Nixon's presidency was rife with corruption.
Like much of the Middle East, Iran has been rife with rebellion against an oppressive regime.
She noticed that the entire affair was rife with oversimplification.
Beckham’s five years in America have been rife with drama, injuries and trade rumors.
Parts of the area are close to the Afghan border long rife with armed smugglers.

Figure 2 Concordances of phrase rife with

If the example sentences are given in this way, the learners would find it easier to master the usage of the word, particularly the prosodic information. Looking up and down, left and right, the learners may notice almost all the collocates, such as anxieties, corruption, rebellion, injuries, smugglers, to the right side of the phrase rife with are usually something bad,
unwanted or unpleasant. By observing the example sentences, the learners may have a deeper understanding of the information given in the definition.

3. Setting collocation column
As is known to all, semantic prosody is largely dependent on the collocates of the lexical item. To add prosodic information into learner’s dictionary, the importance of collocates should be given top priority.

Setting collocation column is a third way of providing prosodic information in learner’s dictionary. In collocation column, the collocates of the lexical item are not only listed as many as possible but also classified in an orderly manner, which is followed by example sentences and usage notes. All of these can be taken advantage by the lexicographer to deal with the prosodic information. The Macmillan Collocations Dictionary for Learners of English (hereafter MCD) serves as a good example in providing prosodic information by means of collocation. In MCD, the various collocates within each entry (or sub-entry) are grouped in the first instance according to the lexico-grammatical structure formed by the collocational phrase, such as adj+N and N+n. And the collocates within lexico-grammatical categories are often grouped into semantic sets. For example, at Cause V+n, there are five sets, labeled respectively as physical damage, physical harm, confusion or disorder, mental harm, and disagreement. And in each set, there is a list of high frequency collocates which are set to be boldfaced. See Figure 3.

Cause v
Make something happen, usually something bad
● V+n physical damage damage, devastation, harm, problem, trouble
The damage caused by the earthquake of 1989 is still not restored.
►Physical harm to people discomfort, hardship, injury, pain, suffering
Any operation is bound to cause some discomfort afterwards.
►Confusion or disorder chaos, confusion, disruption, disturbance, havoc, mayhem
A terrific storm blew, causing havoc on land and sea.
►Mental harm alarm, annoyance, concern, consternation, distress, embarrassment
Noise nuisance can cause severe distress to people who are subjected to it.
►Disagreement controversy, disagreement
The president’s speech caused controversy among some of his supporters.

This presentation of collocates is clear, friendly and offers better accessibility than other dictionaries to the lexical item’s prosodic information. In the first place, the lexico-grammatical structure between the headword and collocates is worthy of note since the semantic prosody is closely associated with the grammatical relationship. The semantic prosody of the same word may vary under different grammatical relationships. In the second place, the semantic set (also called semantic preference) designed in MCD is also innovative and advisable because it is the basis on which the semantic prosody is concluded. And lastly, a long list of collocates with common features may directly lead the learners to judge in general what the semantic prosody of the lexical item tends to be.

Other issues
To bring semantic prosody into learner’s dictionary, there are still some other issues to be addressed. First of all, semantic prosody is completely a new term to most of the users and it should be defined or given illustration in the foreword or front page of the dictionary so as to help the users to establish awareness of semantic prosody. Secondly, semantic prosody of the lexical item may be affected by register. Partington (2004:153) indicates that “it is highly
likely that the quality and strength of the prosody of a good many items will differ from genre to genre or from domain to domain”. For instance, the adjective lavish seems to have a disapproving connotation of excessive wastefulness in newspaper prose whereas in other fields, such as the arts and entertainment, it was generally a neutral-to-good word, e.g. one of the most lavish and entertaining ballets, supplemented with lavish and numerous illustrations. And these differences should be manifested in a particular way, such as using labels or user notes. Thirdly, there is no absolute positive or negative prosody and some items have stronger positive or negative prosody than others. Positive prosody doesn’t mean all the collocates the key item attracts have positive connotation and the same is true with negative prosody. On some occasions, the speaker or writer will on purpose use unusual collocations to create a sense of irony by means of prosodic clash. So it is of great importance to let the users know the probability of semantic prosody. Last but not least, it is unnecessary to provide prosodic information for all the lexical items because most of them have mixed semantic prosody, i.e. the key item attracts both positive and negative collocates. And in providing prosodic information, it is suggested that larger units, such as compound words or phrases, should be handled first because their semantic preference and functional features are relatively evident and easy to follow compared to smaller unit, such as word.

Conclusion

With the aid of corpus and computerized language analyzing software, a great number of lexicographers have come to know the concept of semantic prosody and some of them have tried to put prosodic information into the practice of dictionary compilation. In this paper, three different ways, i.e. displaying with usage label, illustrating with definition & example sentence and setting collocation column, have been proposed as options of dealing with semantic prosody in learner’s dictionary. Nevertheless, it must be admitted that each way has its own merits and demerits, and lexicographers may flexibly choose any one or a combination of them in consideration of the type of target users and design purpose of the dictionary.

Acknowledgments

The authors are grateful to Professor Tian Jianguo for his useful comments on the improvement of this paper. We would like to acknowledge support from Shaanxi Provincial Social Science Fund under Grant No. 13K111. We would also like to acknowledge support from Northwestern Polytechnical University Basic Research Fund under Grant 3102014RW0015 and Xi’an International Studies University Basic Research Fund under Grant 11XWB01.

References

Abstract
As it is targeted at English learners all around the world, an English learners’ dictionary should generally be assumed to be oriented internationally (Cowie 1999), rather than conforming to the norms of a particular English variety and culture. This requirement is further reinforced by the increasing globalisation of English in recent years. In this study, we examine how the eighth edition of the Oxford Advanced Learners’ Dictionary of Current English (OALD8) explicitly and implicitly deals with English varieties and culture-specific representations in its entries, and what progress it has made in this respect over one of its earlier editions (OALD3). We found that OALD8 apparently represents global English varieties in headword selection and in pronunciation, etc. Yet, dictionary-making is contingent upon social practices (Benson 2001), and Anglo-centric stances or ideologies (Kachru & Kahane 1995) in the presentation of materials still abound in OALD8. For instance, when it defines “swallow” as “a small bird […] that spends winter in Africa […]”, OALD8 offers a definition implicitly from the Anglo-centric perspective. We discuss the implications of this and similar findings for pedagogical lexicography, and provide recommendations on how the contents of entries may be improved in future editions.

Keywords: globalisation, ideology, Anglocentrism, OALD

1 Introduction – Defining Anglocentrism in Dictionaries

Being targeted at English learners all around the world, English learners’ dictionaries should generally be assumed to be oriented internationally in their coverage (Cowie 1999), rather than explicitly or inadvertently conforming to, or propagating, the norms of a particular English variety and culture. However, as Cowie (1999), Benson (2001: 34), and others, have frequently pointed out, the creation of (learner) dictionaries can hardly be seen as being free of such ethnocentric bias.

Hornscheidt (2008: 107) states that “Modern monolingual dictionaries […] are regarded as an important (re)source for the construction, manifestation, and naturalization of public attitudes within society”, and sets up a fairly detailed “research agenda” (ibid.) involving a series of questions to be used for critically oriented lexicographic work, albeit not focussing on learner dictionaries, but monolingual ones in general, and in the context of the identification of racism or colonialism. Some of these, such as “Do all people or groups of people have the same chance to identify themselves with word entries in the dictionary?” (ibid.: 115) may also be deemed highly relevant for our present study, although our own agenda is somewhat more neutral in trying to identify aspects of usefulness in, or the general necessity for, entries of OALD8 for learners.

Focussing on ethnic and religious issues, Al-Ajmi & Al-Otaibi (2006) investigate the coverage of terms relating to Islam and the Arab world in three influential dictionaries of English, one of them a learner’s dictionary, the Longman Dictionary of Contemporary English (3rd ed.). Despite establishing a potential ‘taxonomy’ of shortcomings, loosely similar
to ours, identified in the design of the materials under investigation, they conclude that, overall, the dictionaries sighted by them “provide a comparatively more objective image of Islam and the Arab world than what people usually see and read in the mass media and many works of literature, despite the fact that the compilers of these dictionaries are influenced by their social environments and views of other cultures with obviously inadequate access to the original Islamic sources” (Al-Ajmi & Al-Otaibi 2006: 9).

The purposes of the two above-mentioned articles, despite being of relevance to our topic, are more specialised and also, to some extent, more narrowly focused, so at this point it is important to define our own research objectives a little more closely. In a similar vein to Al-Ajmi & Al-Otaibi, in this paper, we want to identify and exemplify a set of criteria that exhibit a certain, unnecessary, cultural bias, and may therefore make entries in the OALD less useful for learners than its compilers may have assumed in creating the list of headwords and corresponding entries. More specifically, we here focus on entries that are in various ways characterised by Anglo-centric views. However, as our conceptualisation of Anglo-centrism may differ somewhat from that of other researchers, we should first try to provide a broad answer to the question “What is Anglocentrism?”.

To do this, it is perhaps useful to begin with some existing definitions in established dictionaries themselves. Beginning with OALD8 itself, we find that it is not defined there at all, and neither is it in a number of other dictionaries we consulted. One of the few traditional dictionaries that does in fact contain a relevant definition, albeit for the corresponding adjective *anglocentric*, is the *Concise Oxford Dictionary* (COD11), which defines it as “centred on or considered in terms of England or Britain”. This, though, is somewhat narrower than our notion of the term, but we finally found an already more adequate definition on Wiktionary:

> The practice of viewing the world from English or Anglo-American perspective, with an implied belief, either consciously or subconsciously, in the preeminence of English or Anglo-American culture.

This definition at least already also incorporates the American perspective missing from the one in COD11, but, as will soon become obvious from our discussion, we here use the term in an even wider sense, including additional assumptions about (necessary) cultural background and linguistic knowledge of dictionary users.

## 2 Methodology

The methodology we employed follows, as far as access to electronic versions of the data allowed us, a corpus-based approach. This revolved, in the first instance, around the extraction of a headword list of 81,742 entries from the electronic version of OALD8, which was then investigated in a number of different ways.

The second step involved trying to narrow down the list of potential candidate entries from the headword list to be investigated for Anglo-centricity. To do so, we worked on the assumption that most Anglo-centric concepts would appear neither in lists of more general English vocabulary, as part of an academic core, nor as part of the most frequent vocabulary in the most comprehensive and carefully sampled corpus of British English to date, the British National Corpus (BNC). We therefore proceeded to create a ‘subtraction list’ from a number of relevant lexical resources, the New General Service List (NGSL; Browne 2014), the New Academic Word List (NAWL; Browne et al. n.d.), as well as a list of types occurring at least 10 * per million words in the BNC. Comparing this list to the OALD headword list and automatically extracting entries only occurring in the latter resulted in a reduced list of 50,196
entries which could then be filtered in various ways and compared to the entries in the electronic version of OALD8.

The same subtraction mechanism was also used to compare the two different editions, this time using an electronic version of the OALD3 that was originally distributed by the Oxford Text Archive (OTA), but has since unfortunately disappeared from their catalogue.

The analysis itself involved filtering the lists for acronyms and proper names on the one hand, and for words containing common (learned) suffixes, on the other, and then investigating the relevant entries. To be able to do the latter more efficiently, we also used wildcards with suffixes in the electronic version of OALD8.

3 Findings and Discussion

Through our research, we found issues in the coverage of OALD8 that may affect foreign learners on a number of different levels. Some of these, as previously referred to, are similar to the categories identified by Al-Ajmi & Al-Otaibi (2006), although their categorisation scheme seems to make some distinctions that appear rather too fine-grained to be applied directly.

3.1 The ‘Missionary Perspective’

We shall begin our discussion by looking at what one may term a ‘missionary perspective’, i.e. a(n) deliberate or inadvertent coverage of terms that are predominantly related to what may be perceived as the promotion of Anglo-American culture, to the extent of neglecting the inclusion of entries related to other cultures. This is, in effect, indirectly related to Hornscheidt’s question stated earlier about whether all users are able to identify themselves with the entries in a dictionary, at least in the sense that an imbalance in the coverage skewed towards Anglo-centric terms may have the converse effect of reducing the number of non-Anglo-centric entries, thereby overemphasising Anglo-American cultural aspects that might in fact not always be of relevance to dictionary users.

We can see such an imbalance very clearly in a number of different features that can easily be quantified. First of all, in OALD8, references to various Inner Circle entities and varieties are still all-too prevalent. More specifically, we find that the occurrences of the prefix {Anglo}- or Inner Circle-related proper names, acronyms and adjectives show up in 11.51% of all entries (see Error! Reference source not found.), while references to Asian/African equivalents do so in only 0.85% (see Error! Reference source not found.).

<table>
<thead>
<tr>
<th>Term</th>
<th># of entries</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>anglo(-)</td>
<td>64</td>
<td>0.078%</td>
</tr>
<tr>
<td>British</td>
<td>2233</td>
<td>2.732%</td>
</tr>
<tr>
<td>Britain</td>
<td>2318</td>
<td>2.836%</td>
</tr>
<tr>
<td>UK</td>
<td>40</td>
<td>0.049%</td>
</tr>
<tr>
<td>American</td>
<td>1356</td>
<td>1.659%</td>
</tr>
<tr>
<td>America</td>
<td>407</td>
<td>0.498%</td>
</tr>
<tr>
<td>US</td>
<td>699</td>
<td>0.855%</td>
</tr>
<tr>
<td>Australian</td>
<td>54</td>
<td>0.066%</td>
</tr>
<tr>
<td>Australia</td>
<td>102</td>
<td>0.125%</td>
</tr>
<tr>
<td>Canadian</td>
<td>71</td>
<td>0.087%</td>
</tr>
<tr>
<td>Term</td>
<td># of entries</td>
<td>%</td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------</td>
<td>-------</td>
</tr>
<tr>
<td>Canada</td>
<td>191</td>
<td>0.234%</td>
</tr>
<tr>
<td>Irish</td>
<td>154</td>
<td>0.188%</td>
</tr>
<tr>
<td>Ireland</td>
<td>386</td>
<td>0.472%</td>
</tr>
<tr>
<td>New Zealand</td>
<td>48</td>
<td>0.059%</td>
</tr>
<tr>
<td>Kiwi (person)</td>
<td>1</td>
<td>0.001%</td>
</tr>
<tr>
<td>Kiwi (fruit)</td>
<td>1</td>
<td>0.001%</td>
</tr>
<tr>
<td>Kiwi (bird)</td>
<td>1</td>
<td>0.001%</td>
</tr>
<tr>
<td>Scottish</td>
<td>272</td>
<td>0.333%</td>
</tr>
<tr>
<td>Scotland</td>
<td>509</td>
<td>0.623%</td>
</tr>
<tr>
<td>Scots/Scotsmen</td>
<td>37</td>
<td>0.045%</td>
</tr>
<tr>
<td>Welsh</td>
<td>116</td>
<td>0.142%</td>
</tr>
<tr>
<td>Wales</td>
<td>347</td>
<td>0.425%</td>
</tr>
<tr>
<td>Total</td>
<td>9407</td>
<td>11.508%</td>
</tr>
</tbody>
</table>

Table 2 Frequencies of terms pertaining to Asian/African countries

<table>
<thead>
<tr>
<th>Term</th>
<th># of entries</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia</td>
<td>119</td>
<td>0.146%</td>
</tr>
<tr>
<td>Asian</td>
<td>71</td>
<td>0.087%</td>
</tr>
<tr>
<td>Africa</td>
<td>154</td>
<td>0.188%</td>
</tr>
<tr>
<td>African</td>
<td>83</td>
<td>0.102%</td>
</tr>
<tr>
<td>India</td>
<td>50</td>
<td>0.061%</td>
</tr>
<tr>
<td>Indian</td>
<td>38</td>
<td>0.046%</td>
</tr>
<tr>
<td>China</td>
<td>26</td>
<td>0.032%</td>
</tr>
<tr>
<td>Chinese</td>
<td>58</td>
<td>0.071%</td>
</tr>
<tr>
<td>Japan</td>
<td>23</td>
<td>0.028%</td>
</tr>
<tr>
<td>Japanese</td>
<td>44</td>
<td>0.054%</td>
</tr>
<tr>
<td>Korea</td>
<td>7</td>
<td>0.009%</td>
</tr>
<tr>
<td>Korean</td>
<td>2</td>
<td>0.002%</td>
</tr>
<tr>
<td>Thailand</td>
<td>1</td>
<td>0.001%</td>
</tr>
<tr>
<td>Thai</td>
<td>1</td>
<td>0.001%</td>
</tr>
<tr>
<td>Indonesia</td>
<td>5</td>
<td>0.006%</td>
</tr>
<tr>
<td>Indonesian</td>
<td>4</td>
<td>0.005%</td>
</tr>
<tr>
<td>Malaysia</td>
<td>3</td>
<td>0.004%</td>
</tr>
<tr>
<td>Malay</td>
<td>2</td>
<td>0.002%</td>
</tr>
<tr>
<td>total</td>
<td>691</td>
<td>0.845%</td>
</tr>
</tbody>
</table>

Apparently contradicting the general assumption that British dictionaries are less encyclopaedic in nature than their American counterparts (cf. Béjoint 2000: 48-51), the electronic version of OALD8 contains a substantial number of encyclopaedic entries relating to Anglo-American entities & concepts, e.g.

- 26 entries alone pertaining to names of native Indian tribes in the US
- around 200 acronyms related to Anglo-American ‘institutions’, such as for AA: “Automobile Association (a British organization which provides services for car
owners)” or AAA: “American Automobile Association (an American organization which provides services for car owners)”

As a matter of fact, the number of proper names/concepts has increased strongly from version 3 to 8, with 6,287 such terms added, with only 2,936, presumably outdated ones, having been deleted. This still represents an overall rise of encyclopaedic entries amounting to 3,351 new headwords (constituting 4.1% of all entries), which not only appears to indicate an increasing ‘Americanisation’, but where many of them, such as All Things Bright and Beautiful, America’s Most Wanted, BSkyB, Dalek, Disney World, Gatorade, Lewis and Clark, Mount Rushmore, Ode to the West Wind, Tuskegee University, etc., refer to (sometimes literally) products of Anglo-American culture that should by no means be considered essential vocabulary for learners. There may of course sometimes be a need to include such ‘cultural facts’ in a dictionary, but essentially only if they do enter into important idiomatic or opaque constructions that could otherwise not be made transparent to the user. This, however, does not appear to apply to the selection of terms listed above.

A certain element of Anglo- or perhaps Western-centric ‘linguistic imperialism’ or cultural bias can also be detected in the frequent references to natives, native-like and native speakers. Thus, although the frequent references to Native American/Canadian in entries related to indigenous populations are rather neutral per se, the definition of native as “connected with the people who originally lived in a country before other people, especially white people, came there” still arguably bears inappropriate connotations of colonialism and racial superiority. Something similar applies to OALD8’s definitions of native speaker as “a person who speaks a language as their first language and has not learned it as a foreign language” and idiomatic as “containing expressions that are natural to a native speaker of a language”. While the first one is increasingly being questioned in discussions revolving around the globalisation of English, especially in contexts where the only ‘qualification’ deemed necessary for English teachers internationally all too often used to be that of being a native speaker, the second one is positively discriminatory, implying that idiomaticity is something only native speakers can ever hope to achieve.

3.2 Misleading Definitions Based on Anglo-Centric Perspective

We next turn to a discussion of entries where part of the definition should either be seen as incorrect/incomplete or at the very least run counter to the expectations a ‘non-Western’ or non-Anglo-Saxon audience may have regarding them. Below is a non-exhaustive sample list of such entries that exhibit various features:

1) Alsatian: a large dog, often trained to help the police, to guard buildings or (especially in the US) to help blind people find their way

2) swallow (n): a small bird with long pointed wings and a tail with two points, that spends the winter in Africa but flies to northern countries for the summer one swallow doesn’t make a summer (saying) you must not take too seriously a small sign that sth is happening or will happen in the future, because the situation could change

3) wine: 1 [uncountable, countable] an alcoholic drink made from the juice of grapes that has been left to ferment. There are many different kinds of wine; 2 [uncountable, countable] an alcoholic drink made from plants or fruits other than grapes

4) cheese: cheese! what you ask sb to say before you take their photograph

Number 1) is exemplary in its Anglo-centricity as it focusses particularly on the American context, but fails to recognise that, after all, Alsatian is only a different term for the type of dog originally referred to as ‘German shepherd’, which clearly marks the origin of the species and the fact that Alsatians were probably trained for various purposes, including the

---

1 ... in a non-Darwinian sense ;-)
ones in definition, in Europe, long before they were introduced to the US. Thus, according to Wikipedia:

“The first guide dog training schools were established in Germany during World War I to enhance the mobility of returning veterans who were blinded in combat, but interest in guide dogs outside of Germany did not become widespread until Dorothy Harrison Eustis, an American dog breeder living in Switzerland, wrote a first-hand account about a guide dog training school in Potsdam, Germany, that was published in The Saturday Evening Post in 1927.” (Wikipedia, Guide dog)

2) is inadequate, not only from a cultural, but also geographical perspective, as not all swallows do in fact migrate, especially not if they live in a more tropical climate, but even if they do, they may not be migrating to and from the locations provided in the definition. However, the notion that swallows from colder climates in Northern Europe do migrate is of course relevant from the idiomatic perspective referred to above, as can be seen in the idiom part following the definition in 2), where the definition from above is a pre-requisite for understanding the idiom in the first place. Nevertheless, to avoid any potential confusion on the part of the learner, the definition should probably be expanded, so that the contrast between migratory and non-migratory behaviour would potentially facilitate comprehension of the idiom.

Example 3) potentially confusing for Chinese dictionary users, as no reference is made to the strength of wine as a distinguishing factor, as for them, ‘wine’ generally refers to stronger types of alcohol unless specifically qualified as ‘grape wine’ (葡萄酒; pútáojiǔ). Therefore, the definition should be enhanced to make these differences in understanding more explicit.

The final example illustrates a kind of pseudo-explanation that may be culturally inadequate or incomplete because here an essential part of the definition is in fact missing in that only the ‘what’ is explained, but not the ‘why’. In other words, the definition lacks an explanation as to why people should say this particular word, i.e. in order make them spread their lips to mimic a smile during the rhyme of the syllable, apart from the fact that the word before should more accurately be replaced by while. Interestingly enough, though, the expression has also been transferred to Chinese as the phonologically similar 茄子 (qiézi; aubergine/eggplant) in analogy.

3.3 Circular, Pseudo-Encyclopaedic vs. Detailed Anglo-Centric Definitions

Another phenomenon observable in some definitions again relates to proper names or concepts, only that this time, we can not only see Anglo-centric features alone in the choice of entries and definitions, but also a stark contrast in the nature of Anglo-centric as opposed to non-Anglo-centric definitions. Here, the Anglo-centric definitions first of all tend to (predominantly) be more detailed than their counterparts, as well as clearly encyclopaedic, e.g. in:

1) Aberdeen: a city and port in north-east Scotland. It is an important fishing port, and was the main centre of the North Sea gas and oil industries. Many of its buildings are made of granite, a hard grey stone, and for this reason it is often called the ‘granite city’.

2) Aberdonian: a person from Aberdeen in Scotland

3) Appomattox Court House: a former small community in the US state of Virginia where the American Civil War ended. General Robert E Lee of the Confederate States surrendered on 9 April 1865 to General Ulysses S Grant in a private house. The area is now a national historical park

4) Beacon Hill: an old, fashionable area of Boston in the US, where many rich families and politicians live
Examples 1), 2) and 3) are highly encyclopaedic, containing a level of detail that marks a clear emphasis on illustrating socio-political aspects. 2), in contrast, appears positively laconic in its brevity, but can of course easily be cross-referenced to 1).

In comparison, the names of e.g. cities, countries, or their inhabitants outside the Anglo-centric domain are patently non-encyclopaedic through a distinct lack of information that frequently tends to render them circular, as e.g. exemplified through the definitions of Slovakia and Slovakian as “a country in central Europe” and “(a person) from Slovakia”, respectively, where no further attempt is made to clarify the exact location of the country as “central Europe” covers a rather large area. As a matter of fact, without the addendum “a person” in brackets in the adjective, the meaning could equally well relate to a concept that the adjective is derived from. The same essentially goes for a large number of similar adjectival entries that end in the suffixes -{ian}, -{ean}, and -{ese}. To resolve the difficulties presented through this, one would at least expect a cross-reference to a map in the printed dictionary or a hyperlink to one in the electronic edition.

3.4 Insufficiently Explained Greek or Latinate Words/Roots

A difficulty of a rather different type presents itself to the (modern) learner through entries pertaining to mythology, religion, general learned vocabulary items, or such that have Greek and Latinate roots. Here, we encounter certain assumptions on the part of the dictionary makers that should either no longer be seen as universally applicable or may potentially already have been causing issues for particular non-Western learner groups in the past.

Let us begin by looking at two examples that belong to the first category, the entries for Augean stables, which is defined as “(in ancient Greek stories) the very large stables which Hercules cleaned in a day by making a river flow through them”, and that for Herculean, explained as “needing a lot of strength, determination or effort”. Now, while such entries as the first one may still be relevant for a student of literature who may encounter frequent references to Greek mythology in reading the classics from the Renaissance or Romantic period, this cannot be seen as essential vocabulary, even for the advanced learner the dictionary is aimed at, while the latter entry may still have more general currency. To compound the issue, though, both entries essentially rely on an understanding of the name Hercules that has no associated entry, so that the definitions basically remain opaque and their usefulness to the learner is thus highly questionable.

Similar issues of transparency apply to definitions of words with Latin or Greek roots, although of course here the usefulness of the terms is generally far less questionable. This is because a large proportion of foreign words from both languages, whether borrowed or neo-classical, is in fact used in English in everyday-life situations, even by speakers with a lower educational background to refer to medical conditions, etc. A classic example here would be pneumonia, for which (at least to our knowledge) there is no non-specialist equivalent, as is for instance the case in the German Lungenentzündung (literally ‘inflammation of the lungs’). Thus, a medical term like amenorrhoea, represented as “a condition in which an adult woman does not menstruate (= there is no flow of blood from her womb every month)”, certainly constitutes an essential entry in a dictionary, but essentially requires following two hyperlinks/cross-references for learners without a suitable language background in both Latin and Greek, as even the ‘word origin’ information “early 19th cent.: from a- ‘without’ + menorrhoea.” does little to alleviate this problem.

The same status, however, should not necessarily be accorded to other learned entries such as terpsichorean, simply defined as “relating to dancing” with a ‘word origin’ addendum of “early 19th cent.: from Terpsichore (used in the 18th cent. to denote a female dancer or the art of dance) + -an.”, which is, despite the fact that it actually occurs three times in the BNC, such a rare and specialist term that it probably does not deserve inclusion in a general-purpose
learner dictionary, especially when much better encyclopaedic, and etymologically also better-founded information about its root can be found online these days:

In Greek mythology, Terpsichore (/tərˈsɪkɔrə/; Τερψιχόρη) "delight in dancing" was one of the nine Muses and goddess of dance and chorus. [...] Her name comes from the Greek words τέρπω ("delight") and χορός ("dance"). (Wikipedia: Terpsichore)

What the Wikipedia entry illustrates quite clearly is how an appropriate amount of etymological information could in fact really allow the user to achieve a proper degree of understanding of the word, especially if it were additionally cross-referenced to other entries containing the same root, such as choreography, chorea the user can follow to deepen their awareness of the degree of productivity and scope of semantic relations of the root.

This lack of direct and sufficiently detailed etymological information, despite frequent links to 'word origin' information and the addition of 127 suffixes ('combining forms') between versions 3 and 8 may constitute a genuine obstacle to the learner who wants to understand such entries properly, rather than simply developing a superficial understanding that may not allow them form mental appropriate associations with a term.

While the above issues may, on the surface, have little connection with notions of Anglocentricity, we can argue that this is in fact rooted in outdated notions regarding the level of classical education a learner may have received, coupled with the above fact that learned words are in fact so prevalent in English. However, even if perhaps about 30 years ago, an in-depth knowledge of Latin and/or Greek language and mythology, as well as other cultural factors relevant to the understanding of the learned part of the English vocabulary may, at least for Western-educated learners, still have been the norm, this may in fact rarely ever have applied to other learners, and even does so to a much lesser extent to younger native and non-speakers of the language from various language backgrounds where such a classical education model may no longer have any currency.

4 Conclusion

What we set out to do in this paper was to investigate to what extent the vocabulary in OALD8 is Anglo-centric, as well as demonstrate the effects that such a potential Anglo-centricity may have on the learners using this dictionary. As our discussion above has hopefully illustrated, there is indeed a substantial number of entries related to encyclopaedic Anglo-centric information present in this edition, and the relevance of at least some of these to learners may be highly questionable. Equally questionable is whether the occasionally relatively large space dedicated to them at the expense of non-Anglo-centric entries is in any way justifiable. Conversely, we have shown that, even if relevant non-Anglo-centric concepts are included in the dictionary to ‘balance’ the Anglo-centric ones, their definitions are frequently too brief to be useful, often affected by circularity, or may be culturally misleading and/or incomplete.

Last, but not least, we have tried to point out how issues related to established background expectations about a certain (Western-)educated readership may affect the overall intelligibility and usefulness of parts of the learned vocabulary in OALD8, claiming that these notions are no longer, or may indeed never have been, applicable to the target audience of the dictionary, and that ability to understand many such entries may successively be turning into an issue for Western-educated users, too.

Based on the above observations, we would like to propose to implement the following recommendations in future editions of the OALD, as well as perhaps other learner’s dictionaries that might be affected by the same issues. As a first step, it may be advisable to
reduce the amount of unnecessary encyclopaedic information related to Anglo-centric concepts and entities. Learners can easily find such information on Wikipedia, where it can also constantly be updated if necessary, and space is not an issue, so that far more detailed explanations can be provided than is practical in a dictionary, especially in a print version. Doing so would allow the compilers of the OALD to focus on improving the remaining essential entries in at least two of ways. First, the relevant entries could be improved by adding more etymological information directly where and when applicable, which would potentially enhance the user’s awareness of morphological structure and productivity. Secondly, focussing on fewer items would perhaps also make it possible to remove at least some of the Anglo-centric bias in order to adopt more global perspectives that are more appropriate for an international and modern readership.

References
Browne, Charles. 2014. A New General Service List: The Better Mousetrap We’ve Been Looking for? Vocabulary Learning and Instruction, Volume 3(1). 1-10. doi: http://dx.doi.org/10.7820/vli.v03.1.browne
Pragmatic Information in Learner’s Dictionaries: Redefined

Anmin Wang
Macquarie University
anmin.wang@students.mq.edu.au

Abstract
The importance of pragmatic knowledge to foreign learners of a language has long been acknowledged. Learner’s dictionaries may be of help to these learners in acquiring such knowledge via pragmatic information offered in them. The definition of such information by different researchers from various perspectives explains why its presentation varies in lexicographical practice. To offer pragmatic information systematically, it needs to be redefined, which calls for not viewing pragmatics and lexicography as two fields on a par with each other, but delimiting the former within the latter, with conventionalized stabilized speaker meaning as their point of convergence. This will specify the portion of pragmatic meaning capable of being treated in a learner’s dictionary—lexico-pragmatic meaning, laying the foundation for presenting such meaning as pragmatic information systematically.

Keywords: pragmatic information, redefine, conventionalized, speaker meaning, lexico-pragmatic meaning

Introduction
In learning a foreign language, learners can commit grammatical as well as pragmatic errors. A “considerable proportion of them [learner errors] are caused by the learner’s lack of pragmatic knowledge” (Yang 2007: 147). Thus it may follow that such information should be included in a learner’s dictionary. However, even if the importance of such information to foreign language learners has been well acknowledged, generally speaking, the inclusion of such information still leaves a lot to be desired. English-Chinese (henceforth abbreviated as E-C) dictionaries for beginning learners of Chinese as a foreign language (shortened into CFL learners) is no exception.

Even if “there are pragmatic components and effects in the language under description…in dictionaries”(Zgusta 1988: 246), collectively known as pragmatic information in lexicographical field, it is a pity that “[i]t is not even certain what …pragmatic information in the context of … lexicography is”(Kawamura 2014: 6). The divergence in presenting such information among various dictionaries may have stemmed from researchers’ disagreement on what pragmatic information is. The different perspectives to define it, including the pragmatic one, help explain this. Meanwhile, even those defining pragmatic information from the perspective of pragmatics also displays variation, which results from people’s different understanding of what pragmatics is. Thus before reviewing definitions of pragmatic information from various perspectives, this article considers what pragmatics is. Taking Ariel’s (2010) view on pragmatics and scrutinizing different definitions of pragmatic information prepare the author to delimit the territory of pragmatics in the context of lexicography, with conventionalized speaker meaning as the point of their convergence. Thus, lexico-pragmatics, a term coined by the author, studies lexico-pragmatic meaning of three characteristics. Pragmatic information is the presentation of such meaning in a dictionary via appropriate lexicographical means on every structural level in dictionaries.
1 What is pragmatics

What pragmatics is may not be a question that has a unanimous answer among even pragmatists themselves. Though numerous topics fall under the same big tent of pragmatics, generally speaking, “[W]ith contemporary pragmatics, there are two main schools: Anglo-American and Continental” (LoCastro 2012: 7), which focus on different research targets.

Pragmatists of the former school take “pragmatics as a core area that overlaps with semantics within formal linguistics, along with syntax, morphology, phonology, and semantics” (LoCastro 2012: 7). The topics appealing to pragmatists from Anglo-American school were first systematically listed out by Levinson (1983: 14) in Pragmatics, which comprise deixis, conventional implicature, presupposition, speech act, conversational structure and so on. Pragmatists within this group seem to have the consensus that pragmatics has definite units for study or analysis, covering “the pragmatic properties of specific words or linguistic structures” (He & Chen 2004: 7) and those beyond these words and structures. Listing these topics together may leave the impression that they are on a par with each other in the field of pragmatics, which actually is not the case. Some of them have thus far been favored by researchers, such as speech act, while other ones may have been comparatively orphaned or marginalized, like conventional implicature. However, it ought to be noted that “P[r]agmatic analysis cannot afford to be restricted to some predetermined subject matter, no matter how liberal the choice is” (Ariel 2010: 19). In addition, whether dictionaries can and to what extent they can present pragmatic information associated with some of the topics, especially something like particularized conversational implicature, remains a question.

Meanwhile, the Continental school, with Verschueren as one of its representatives, holds the view that the study objects of pragmatics — “linguistic phenomena to be studied from the point of view of their usage can be situated at any level of structure or may pertain to any type of form meaning relationship” (Verschueren 1999: 3). This view broadens as well as blurs the boundary of pragmatics, since the “definition does not introduce a strict boundary between pragmatics and some other areas in the field of linguistics” (ibid.). Thus such a definition may lend little help to setting up a systematic framework for integrating pragmatic information into learner’s dictionaries as well. Of course, this is not to deny that this view of pragmatics does not hold water, but to say that the over-broad concept of pragmatics is not conducive to offering pragmatic information systematically in dictionaries.

Therefore, what pragmatics is may need an alternative perspective. To integrate it into lexicography, it can neither be restricted to fixed topics nor be expanded unlimitedly. Pragmatics “calls for the identification of the pragmatic aspects relevant to any language use, and teasing these apart from grammatical aspects” (Ariel 2010: 19). Then, according to her (ibid.: 249), pragmatics ought to focus on “inferences which are based on the linguistic strings expressed when contextual assumptions are taken into account. Then what is left should be studied by grammarians rather than pragmatists. By taking this view, she appears to argue from hearer’s perspective in a communication. Her criterion can be applied unanimously to determine whether something belongs to pragmatics, thus contributing to incorporating pragmatic information in a systematic way into learner’s dictionaries.

2. Definitions of pragmatic information: A review

Up to now, compared with research on grammatical information in dictionaries, especially learner’s ones, that on pragmatic information is rather scarce as well as sketchy. With a preface entitled Pragmatics and the Dictionary in Longman Dictionary of Contemporary English (the 2nd edition), Leech and Thomas introduced the concept of pragmatic information into learner’s dictionaries in 1987. Since then, only a few journal
articles, certain book sections, unpublished doctoral theses and one monography on lexicography have discussed it from various perspectives: communicative, contextual, socio-cultural and pragmatic.

2.1 Communicative perspective
Researchers like Sharpe and Apreseyan look into pragmatic information from the communicative perspective. Sharpe (1989: 315) defines it as “knowing which lexical items target language speakers would use to achieve a certain communicative task”. The definition, aside from being too general to guide the provision of pragmatic information in learner’s dictionaries, also limits such information to the lexical level in a language, overlooking the fact that pragmatic information can be presented on the level beyond, such as sentential or textual level in a dictionary. Apreseyan (1988 cited in Burkhanov 2003: 103) holds that pragmatic specifications [i.e. pragmatic information] in lexicography concentrates on conveying speakers’ attitude to reality, the message and/or the interlocutor, which is encoded in linguistic signs as units of language-system. This definition has mentioned speaker’s attitude, one important component of pragmatic information. However, it forms just part of such information. Thus it can be said that the communicative perspective, though realizing the presence of pragmatic information on the lexical level in a language or capturing certain components of pragmatic information like speakers’ attitude, fails to come up with a satisfactory definition for it in the lexicographical context.

2.2 Contextual perspective
Some researchers examine pragmatic information from a contextual perspective, breaking such information into various components which mostly coincide with usages in dictionaries. Thus it is not surprising that labelling becomes the main means to present it. Hagasi (1981) views pragmatic information as different components of language use. The first includes temporal plus spatial use of language and language users’ social status, which is partly echoed by Kipfer (1984: 41), who defines it as variation in language, temporal and spatial features as well as the relationship between the interlocutors. The second involves their age, gender, and occupation; the third is made up of audience, register and so on. Svensén (1993: 6), Huang and Chen (2001: 48–49), and Li and Zhou (2001: 138) more or less take a similar view to that of Hagasi’s. Though factors like language users’ gender, age, social status, their relationship may constitute part of pragmatic information, it is questionable whether temporal, spatial or other variations also fall into the category of such information.

To researchers like Landau, pragmatic information covers some of contextual factors mentioned above, but is not limited to them. Aside from agreeing that such information includes temporal and regional variation, language users’ social status, he expands it to be concept almost the same as usage, covering other factors like technical or specialized terminology, restricted or taboo sexual and scatological usage, insult, slang, style or register. (Landau 2001: 217-272) Though certain usage, like insult, may convey the speaker attitude or feelings, thus can be taken as pragmatic information, usage itself is not the same as pragmatic information.

Thus it deserves to be mentioned that even if some contextual factors make up important components of pragmatic information, not all of them do. Meanwhile, pragmatic information should not be equated with usages. Furthermore, one “critical weakness of this view is that it confines pragmatic information to the lexical level” (Yang 2007: 148) in the microstructure of dictionaries as well. Meanwhile, even if labeling can be an important means to present such information in a dictionary, some of it may well require other forms of presentation. Thus, this perspective cannot offer a framework for integrating pragmatic information systematically into learner’s dictionaries.
2.3 Socio-Cultural Perspective
Some researchers examine pragmatic information from the socio-cultural perspective. Nuccorini (1993) defines pragmatics in the scope of lexicography in Pragmatics in Learner’s Dictionaries: “[I]n this framework, pragmatics is restricted just to an incomplete set of commonly shared, culture-specific, social conventions and linguistic assumptions, more or less clearly stated and exemplified”. He equals such information to socio-cultural conventions and linguistic assumptions, which may not cover all the pragmatic information. Hartmann and James (2000: 111) view such information as something “on the sociocultural rules of speaking”, covering paralinguistic features like tone, choosing words in terms of politeness and formality conventions, etc. (ibid.). Though choosing words out of politeness and formality may be part of pragmatic information, just as Yang (2007: 149) points out, this definition is “too broad to be useful as a guide to what pragmatic information should be contained in the dictionary”. Therefore, this perspective, like three others above, still fails to accomplish the mission of defining pragmatic information adequately.

2.4 Pragmatic perspective
Some researchers approach pragmatic information by integrating pragmatics proper with lexicography, viewing the two fields on a par with each other. Qian (1995), Yang (2004, 2005, 2007) all belong to this group. Qian (1995) makes mention of three types of pragmatic information: referential, textual and socio-pragmatic, which is associated with deictic terms, discourse markers, speaker’s attitude and intention plus the relationship between the speaker and the hearer, etc. respectively. Even if certain components mentioned above, like speaker’s attitude plus the communicator’s relationship, do make up an important part of pragmatic information, his definition still can’t offer a unified framework for presenting such information. Yang (2004, 2005, 2007) attempts to define such information by integrating pragmatic meaning into lexicographic meaning. Her notion of pragmatic meaning comes from Levinson (1983): conventional implicature, presupposition, felicity condition, generalized conversational implicature, particularized conversational implicature and inferences based on conversational structure, to which she adds pragmatic meaning arising from politeness. Though it remains a question to what extent a print dictionary can present certain kinds of pragmatic meaning, like particularized implicature, she is right to point out that pragmatic information is not restricted to lexical level. Yang and Qian appear to take pragmatics as something made up of fixed canonical topics, such as speech acts theory, politeness, deixis, which represent the Anglo-American pragmatic traditions, and consider all the linguistic aspects associated with these topics entitle them to a position in learner’s dictionaries as pragmatic information, which is not necessarily true. In this aspect, her view echoes that of Cowie’s (1984). The latter argues that pragmatic specialization, which is similar to pragmatic information, associated with expressions like “would you mind” should be provided for foreign learners of English, since they can perform particular speech acts.

Kawamura (2014) represents another attempt to define pragmatic information by combining pragmatics with lexicography. She (2014: 40) confines pragmatic information to “those discourse or pragmatic functions and pragmatic biases, a learner’s ignorance of which could cause a serious problem, especially in those cases where [speaker]’s utterances might sound rude or offensive”. This definition does help narrow down the scope of pragmatic information in a learner’s dictionary. However, it remains to be seen how such restriction will help establish the systematic framework for offering pragmatic information in a learner’s dictionary, since it is still not clear what such functions and biases exactly are. Besides, it is doubtful that they can stand for all pragmatic information.
To sum up, all the efforts to define pragmatic information from diverse perspectives above, except Yang’s, have failed to provide a unified framework for integrating such information into learner’s dictionaries. Nevertheless, Yang does not seek to restrict the scope of pragmatic meaning that can be treated in dictionaries. On the contrary, she tries to bring all the components of pragmatic meaning into lexicographical meaning, which may not be workable for a dictionary. Thus to define pragmatic information, the relationship between pragmatics and lexicography needs to be re-examined. They should not be taken as two fields enjoying equal status, but the former delimited in the latter.

3 Pragmatics in the context of lexicography: Redefined as lexico-pragmatics

It can be argued that attempts to define pragmatic information from perspectives other than pragmatics may not be able to produce a framework for offering such information systematically in learner’s dictionaries. Thus taking a communicative, contextual, socio-cultural view can not help achieve such a goal. To attain it, the relationship between pragmatics and lexicography needs to be re-examined, thus requiring us to delimit pragmatics within lexicography and picking out a point of convergence to combine them.

3.1 Delimiting pragmatics within lexicography

To define pragmatic information adequately, instead of viewing pragmatics as something enjoying similar status as lexicography, the scope of pragmatics needs to be delineated in the context of lexicography. Previous researchers of pragmatic information, like Qian (1995), Yang (2004, 2005, 2007), seem to regard these two fields as equals and try to merge them together. However, to reveal their relationship in the context of lexicography, the fact that the two fields are not on a par with each other should be the foundation. Kawamura (2014) is correct in attempting to limit the scope of pragmatics within lexicography by restricting pragmatic meaning in a dictionary. She (2014: 34) believes pragmatic information concerns “the part of pragmatic meaning that is [can be] accommodated in the dictionaries” and they specify “what portion of pragmatic meaning should be included”(ibid.). However, she only restricts pragmatic meaning to that conveyed by pragmatic functions and biases, which may not be complete.

To present a true picture of the relationship between these two fields, “[i]t should be emphasized that … the term ‘pragmatics’ [in lexicography] primarily refers not to a branch of linguistic inquiry, but to linguistic phenomena [originally italicized] which should be studied within the scope of that discipline.” (Burkhanov 2003: 103) On this basis, it can be argued that pragmatics within the scope of lexicography constitutes something that can be entitled lexico-pragmatics, which concerns pragmatic information in learner’s dictionaries and will be detailed below by discussing the point of convergence of the two fields.

3.2 Meaning: Where pragmatics and lexicography converge

Aside from the aspect mentioned above, redefining the relationship between pragmatics and lexicography needs a point where the two fields converge: meaning. It is something appealing to and investigated by both pragmatists and lexicographers, though obviously in different ways.

Lexicographers mainly deal with the meaning of words and word strings in a language system. Their central task is to deal with lexical meaning. “It is, thus, in explaining, describing and defining the meaning of words that the major function of dictionaries is considered to lie, and on which they are judged.” (Jackson 2001: 86) Or as researchers like (Zgusta 1971: 21, Zhang, 2002: 3) point out, to lexicographers, word meaning is their focal issue. They “try to obtain knowledge of the lexical meaning of a word [words] as a part of the system of language”(Zgusta, 1971: 24) and “the goal of the lexicographer’s work is [to define] their abstract value in the [language] system”(ibid.: 26). In other words, lexicographers deal with systematic meaning or stabilized meaning of lexical
items within a language system. Understanding lexical meaning in such a way entails the need to include pragmatic meaning in a dictionary.

Pragmatics, on the other hand, mostly generally speaking, deals with meaning in use in socio-cultural settings. Though well aware of the dynamic nature of pragmatic meaning as a type of meaning negotiation between interlocutors in specific contexts, I take it here as speaker meaning conveyed by the linguistic strings expressed where average hearers may be able to infer in the usual contextual settings. “Considering the character of dictionary meaning, dictionaries can and should only deal with most fixed parts of pragmatic meaning” (Kawamura 2014: 37). Therefore, it naturally follows that a learner’s dictionary can not include all the speaker meanings as pragmatic meaning, but it can often help foreign learners of a language “by giving examples of typical speaker-meanings” (Leech & Thomas 1987: F12) via appropriate lexicographical means.

To conclude, the two previous subsections: pragmatics is delimited in lexicography as linguistic phenomena investigated in lexicography; pragmatics within lexicography is incapable of dealing with all the probable speaker meanings, plus the perspective that pragmatics deals with inferences based on linguistic strings that are expressed with contextual information considered in section 1, pave the way for the study of pragmatic information in dictionaries. Let us call it lexico-pragmatics: a branch of lexicographical study handling conventionalized speaker meaning conveyed by linguistic strings via appropriate lexicographical means on various structural levels of dictionaries. It deals with lexico-pragmatic meaning. The relationship between pragmatics and lexicography is illustrated in the diagram below.

![Figure 1 The diagram of lexico-pragmatic meaning](image)

4. Lexico-pragmatic meaning expounded

Three essential features help define lexico-pragmatic meaning, a concept mentioned above. The features of such meaning will be detailed in the following paragraphs. The first one is that it must be conventionalized in socio-cultural settings. The second one is that it must be stabilized, conveyed by linguistic strings as part of language system rather than their occasional applications. The last one is that it must be speaker meaning, including his or her attitudes, feelings, etc., conveyed by such strings.
4.1 Conventionalized
Being conventionalized depicts the very nature of lexico-pragmatic meaning, contrasting it with pragmatic meaning proper, which can be understood as meaning in use in socio-cultural settings, as well as narrowing down its own territory. Such meaning produced by linguistic strings or expressions, when applied in communication, is usually “instantly recognized as such, and no series of steps need to be posited for the hearer to be able to work out the intended illocutionary force of the utterance.” (Nattinger & DeCarrico 1992: 49) In other words, average hearers are able to immediately infer or recognize the conventionalized speaker meaning under usual contextual settings. The pragmatic meaning generated by linguistic strings or expressions ad hoc is beyond the capacity of lexico-pragmatists.

4.2 Stabilized
Stability of meaning, on which definitions of meaning in dictionaries are based, also applies to lexico-pragmatic meaning. Lexicographers must “derive the generalized description, or formulation, of the lexical [lexico-pragmatic] meaning of which…applications…are actual manifestations”(Zgusta 1971: 26). If one application of such meaning is occasional, adopted by only some language user temporarily to achieve certain pragmatic effects, then such meaning won’t be included in a learner’s dictionary, since it is still not part of the system of language. Lexico-pragmatists are attracted by how speaker “meaning that lexical[linguistic] items display in various contexts is to be described” (Van der Eijk et al. 1995: 1-2) as belonging to language system, rather than their occasional applications in specific contexts, which is usually the interest of pragmatists.

4.3 Speaker Meaning
Lexico-pragmatic meaning must be speaker meaning conveyed by linguistic strings, including his or her attitudes, feelings, etc., usually beyond the lexicographical meaning of such strings. Such meaning is thus associated with or conveyed by these strings but not encoded in their lexicographical meaning as revealed by their definitions. Take for example, wènti shì 问题是 “The problem is” in Chinese is often taken as an expression to mark the speaker is going to introduce something negative or that (partially) contradicts what others have expressed. However, such meaning is not incorporated into the definitions of the two individual words making up this expression. Speaker-meaning like this constitutes lexico-pragmatic meaning in learner’s dictionaries. Thus pragmatic information in such dictionaries is the presentation of lexico-pragmatic meaning via various lexicographical means on different structural levels in dictionaries.

Conclusion
This article, among other things, has established the following points. Firstly, it has provided an up-to-date review of the definitions of pragmatic information from various perspectives. In addition, inspired by Ariel’s definition of pragmatics, I have redefined the relationship between pragmatics and lexicography, with the former delimited within the latter to generate lexico-pragmatics, which studies lexico-pragmatic meaning. Last but not the least, three core features of lexico-pragmatic meaning are explained, which set the criteria for presenting it as pragmatic information in learner’s dictionaries in a systematic way.

Meanwhile, it ought to be pointed out that the present attempt still leaves much space for future exploration. First, with the further research on pragmatics, future researchers may approach pragmatic information from a different perspective. Second, the present paper is only a theoretical exploration, which must be followed up by investigating pragmatic information associated with a certain foreign language in learner’s dictionaries, and seeking out appropriate lexicographical means to present such information. Though it is not stated openly, the present paper is limited to examining pragmatic information in print.
dictionaries. The ways to present such information in learner’s dictionaries in other media are worthy of exploring as well.

References
Li, Ming and Jing Hua, Zhou. 2001. 双语词典编纂 Shuangyu Cidian Bianzuan (On Compiling Bilingual Dictionaries. Shanghai: Shanghai Foreign Language Education Press.


Crowdsourcing Microdata for Cost-Effective and Reliable Lexicography

Martin Benjamin
École Polytechnique Fédérale de Lausanne, Switzerland
martin.benjamin@epfl.ch

Abstract
Lexicography has long faced the challenge of having too few specialists to document too many words in too many languages with too many linguistic features. Great dictionaries are invariably the product of many person-years of labor, whether the lifetime work of an individual or the lengthy collaboration of a team. Is it possible to use public contributions to vastly reduce the time and cost of producing a dictionary while ensuring high quality? Crowdsourcing, often seen as the solution for large-scale data acquisition or analysis, is fraught with problems in the context of lexicography. Language is not binary, so there may be no one right answer to say that a word “means” a particular definition, or that a word in one language “is” the same as a particular translation term. People may misinterpret instructions or misread terms or make typographical or conceptual errors. Some crowd members intentionally add bad data. Without a payment system, incentives for participation are slim; micro-payments introduce the incentive to maximize income over quality.

Our project introduces a public interface that breaks lexicographic data collection into targeted microtasks, within a stimulating game environment on Facebook, phones, and the web. Players earn points for answers that win consensus. Validation is achieved by redundancy, while malicious users are detected through persistent deviations. Data can be collected for any language, in an integrated multilingual framework focused on the serial production of monolingual dictionaries linked at the concept level. Questions are sequential, first eliciting a lemma, then a definition, then other information, according to a prioritized concept list. The method can also be used to merge existing data sets. Intensive trials are currently underway in Vietnamese, with the inclusion of additional Asian languages an explicit objective.

Keywords: crowdsourcing, gamification, distributed lexicography, multilingual, quality assurance

1. Introduction.

Kamusi GOLD is the Global Online Living Dictionary, with the goal of producing a multilingual dictionary with comprehensive data for “every word in every language”. Given the size of each language, the large amount of information and nuance that can be associated with each term, the number of languages and variants, and the near impossibility of extensively engaging informants for many smaller languages, the goal of “every word” is axiomatically out of reach. However, “every word” sets a target for system design: a structure that can accommodate the full range of linguistic data associated with each term (Benjamin 2014 b and c), and a set of procedures that can elicit and process data with speakers of any language (Benjamin 2014 a, Benjamin and Radetzky 2014 a and b). In this paper, we describe the public interface that breaks lexicographic data collection into targeted microtasks. Tasks build upon each other, including the creation of new data and the validation of others’ contributions. Questions are presented within stimulating game environments (McGonigal 2011, Hamari and Koivisto 2013) on Facebook, mobile devices, and the web. The ultimate goal is a resource that can give people and technology services a wide assortment of useful information for any written or spoken term. Each language should have a monolingual dictionary of all the concepts it produces, with definitions in its own language, extended information such as word forms, and links to close equivalent ways of expressing the idea in other languages. Our premise is that most of the data we are seeking cannot or will not be produced in a codified form by relying on traditional methods of field or desk lexicography. Through well-regulated public engagement, the methods described in this paper can collect
data systematically for numerous languages that would otherwise be out of reach for detailed lexicographical investigation.

2. Challenges addressed

2.1 Great dictionaries take time. Individuals often compile dictionaries over decades, and teams may not be much faster. Acquiring and arranging lexicographical data is a difficult task that requires serious oversight from knowledgeable individuals. Rare is the case where lexicography pays the rent, and rare is the case where a language enthusiast can see a serious dictionary through to completion; because of limitations of time and person power, great dictionaries for most languages have never been written. The Kamusi Project seeks to change that, through a combination of systems for language experts to compile dictionaries using a central lexicographical resource, computational techniques for harvesting existing data, and the methods discussed in this paper for users to contribute to the growth of resources for their own languages. The crowd will still need many person-years to achieve quality results, but the metered time can be condensed through an expanded number of participants versus traditional lexicography (Brabham 2011). With careful control of data acquisition and review, we contend that public engagement will result in quality lexicographical data for many languages that otherwise would not or could not exist.

2.2. Crowdsourcing offers the potential of rapid data collection at low cost, but it is also fraught with dangers (Saxton et al. 2013). Crowdsourcing with micropayments through services such as Mechanical Turk introduces a host of incentives for participants to cheat (Kittur et al. 2008, Hirth et al. 2013) that we would not want to expend resources in combatting, were our budget even to allow. Absent a financial incentive, many people with great intentions may nevertheless provide poor information. They may have a faulty understanding of a question, or an inadequate understanding of the parameters of a satisfactory answer. They may have poor spelling or poor grammar. Additionally, users might provide wrong information for malicious reasons. Many crowdsourcing projects surmount such problems by greatly limiting the tasks that contributors can perform (Hsueh et al. 2009), with limited goals such as identifying the features of photos. Other projects ferret out bad information by relying on users to find problems, whether the Wiki method of trusting that mistakes will be found and fixed, or Amazon asking users if they find a review helpful. The methods we describe are designed to learn from each contributor to the maximum of their capabilities, while constraining the information to the format needed for consistent data, and ensuring that bad data does not get published to the project.

2.3. In the context of language, crowdsourcing has peculiar advantages as well as posing particular problems. People are passionate about their languages in a way they are not about their reviews on Trip Advisor or their votes on Slashdot. Their passion may become dogmatic, to the extent that one speaker may try to overwrite contributions from someone who speaks a variant form. Even answers that are essentially the same may adhere to different orthographies, which the methods discussed herein do not address other than to pass conflict to a separate stage of manual review. Yet, important variations in language can be revealed by the crowd in a way not accessible to most lexicography projects. Both the numbers of potential participants and the ability for people to contribute along the entire geographic range across which a language might vary open public lexicography to a never-before-seen breadth of speaker input. For most crowd projects, participants can offer knowledgeable opinions, whether confirmation that an image from a telescope has the characteristics of a galaxy, or a fully-researched article on Wikipedia (Schenk and Guittard 2009). With language, however, a native speaker is an inherently authoritative, if problematic, source. The challenge for the project is to transfer this authoritative knowledge to a codified, reliable form.
3. Methods for data elicitation and validation

This section describes the systems we have designed for eliciting valid lexical data from members of the public. The system is under continuous development, with frequent modifications, so the description at the time of writing might not exactly correspond to the implementation at the time of reading. Each of the modes has the same general facets: tasks should be simple to understand, simple to complete, and acquire validity through consensus. Some of the modes can be activated as-is for any language, while other modes need to be configured for the particular needs of a language (e.g., eliciting the dual form of nouns in Arabic). All modes would ideally have their controls and explanations properly localized for their own language, though this will require fairly extensive direct interaction with speakers to implement.

3.1 Translation Terms. Our first mode is geared to producing a set of parallel concepts across languages, in order to establish a baseline vocabulary for each. If we do not already have digitized data available for a language, which would be treated in the Alignment mode, this will be the first set of information that enters the system for the language. Although the project aspires to go well beyond crude mot à mot translations, this mode gives a foundation on which more complex data can be assembled.

In translation mode, the participant is shown a word and definition in English, and asked for the word they would use to express that concept in their language. Importantly, we do not ask questions such as, “What is ‘pen’ in your language”, but rather, “Pen: a handheld writing implement with ink” or “Pen: an enclosure for animals”. The mode serves terms from an imperfect priority list based on factors such as English corpus frequency (Benjamin 2013). When Term X has multiple senses, we present the sense that appears first arbitrarily in our dataset, then move through senses of the next ten terms on the priority list, then cycle back to the next unqueried sense of Term X, so that the participant is not inundated with being asked about multiple senses of the same key term in rapid succession. When someone skips a question, we anonymously store that information so that we can learn which senses to prioritize and which to demote; e.g., all languages might have a way to express “run” as opposed to “walk”, but not a “run” in baseball. When a threshold number of people submit the same term, we consider the item valid, pass it to the next mode, and publish it on Kamusi as a preliminary entry.

3.2. Synonyms. Although Translation mode can elicit the most popular term for a given concept, it also has the potential to bring in ambiguous information that needs a different mode for review. Were 10 Swahili speakers asked to translate “car”, six might propose “motokaa”, three might suggest “gari”, and one could input “motocaa”. Both “motokaa” and “gari” are valid answers, while “motocaa” is an error. In Synonyms mode, we are able to compare a validated term for which we have a definition with another term for which we have reason to suspect equivalence, either from original submissions via Translation mode or from the Open Multilingual Wordnet (http://compling.hss.ntu.edu.sg/omw/). We show the validated term and its definition, and ask whether the comparison term is another term with the same meaning, an alternate spelling of the same term, or an error. In the example, “gari” would quickly reach consensus and “motocaa” eliminated.

3.3. Word forms. Kamusi approaches morphological forms from two directions: cataloguing each form, or producing them algorithmically. When a part of speech has a small number of forms within a language, such as the five inflections of an English verb (see, sees, saw, seen, seeing), setting up procedures to elicit the forms is straightforward. In cases where the task is unambiguous, we can present the crowd with the lemma and its definition and ask for the desired subsidiary form, e.g., what is the plural for “car”, a vehicle with four wheels?
This task has particular dangers. First, non-specialists are unlikely to be able to answer a question such as, “What is the past participle for ‘see’”? Our current solution is to set up the Forms mode on a case-by-case basis, in consultation with language specialists; if we cannot find a way to phrase a question to get the proper form from the public, we will defer that item until we are able to assign the task to a trusted user. Second, we cannot assume that all senses of a lemma have the same inflected forms, e.g., the plural of the insect “louse” is “lice”, while the plural of a human “louse” is “louses”. Our solution is to ask for Forms for only a single sense, and then ask for confirmation that additional senses conform to the same spelling pattern – a low-threshold microtask. Third, alternate spellings are possible, such as “spelled” and “spelt”. To solve this, we place competing spellings in the same type of verify/reject loop that we designed for Synonyms.

Certain parts of speech have too many forms to tackle with iterative public input. The twenty or so forms that an adjective could have in many Bantu languages reaches the outer edge of what can be sought in Forms mode. Verb conjugation tables in Romance languages, which can hover near 100 forms, are best filled in by trusted users; a future tool will populate such tables. For agglutinative languages with highly predictable forms, such as the nearly one billion possibilities for the Kinyarwanda verb, the only solution is to work with specialists on algorithms to parse the extended forms.

3.4. Definitions. “Dictionaries” for most languages are generally compilations of bilingual word pairs and selected extended information, with the assumption that readers have an a priori grasp on the underlying concept. For an almost random example, dictionaries for many languages pair a term in that language to English “bull”, with the typical first guess that the entry references a male cow. However, most good English monolingual dictionaries list more than a dozen other senses for “bull” as well. Our premise is that each term should be defined in its own language, creating monolingual references that link to the expressions of similar concepts across languages. Members of the public can provide these definitions, but writing a definition is a difficult task. The Definition mode is designed to work toward agreed definitions that could pass the muster of a trained lexicographer, with the acknowledgement that initial data might be less than perfect.

The starting point for the Definitions mode is the less-than-perfect set of ~200,000 English terms from the Princeton WordNet (PWN: http://wordnet.princeton.edu/). In the English version, we present an English term, such as “elevator car”, and the PWN working definition, in this case, “where passengers ride up and down”, and ask whether the given definition is good as is or if the reader can write a better definition. The reader also has the option to skip any entry, or to choose a definition submitted by an earlier participant. Within game play, a player receives substantial points for writing the winning definition, and one point for upvoting the eventual winning entry. Once a definition has passed the consensus threshold, it is published in Kamusi. However, Kamusi entries will have the option for readers to challenge the published definition, putting it back into consideration within the game setting.

For other languages, the Definition mode can be activated once a translation equivalent has been validated. We present the term, and show its English equivalent and definition (or PWN working definition). Showing the English definition unfortunately tends to steer the result toward the English concept rather than the nuance of the indigenous term, but is necessary for two reasons. First, it is absolutely essential that to align the correct senses across languages – to grab the right “bull” by the horns. Second, since most contributors do not arrive knowing how to write a good dictionary entry, the English definition will provide guidance that will give them a launching point, with the real potential that subsequent readers will have the incentive to improve weak definitions.

3.5. Examples. For languages that have a digital footprint, the written corpus can provide a trove of material for demonstrating the meanings conveyed in writing. The sense of a term
cannot, however, be gleaned automatically, to a level of confidence one would inscribe in a
dictionary. Expending enormous computational firepower to distinguish a raging bull from a
bull market from a pile of bull would still result in only a statistical guess about whether a
particular sentence pertains to a particular sense. The Examples mode combs through corpus
sources, finds sentences that match one of the forms of a term, and presents several of those
sentences to reviewers. The participants’ task is to select the lines that they feel exemplify a
sense.

In Examples mode, a sentence must reach a high threshold to be accepted, but a lower one to
be rejected. Once a few reviewers look at a sentence and decide it is not helpful, project time
is better spent returning to the corpus and selecting another candidate for evaluation (Vaish et
al. 2014). Finding a new example is cheap, whereas over-evaluating the same poor sentence
costs time and goodwill. When set as a game, players receive points for all sentences where
they are on the consensus side, either for selecting a good sentence, or for not selecting ones
that others also agree are inadequate.

The initial version of Examples uses Twitter to harvest English text and the Helsinki Corpus
of Swahili (HCS) for Swahili. We display the relevant term and definition, and for Swahili
also the English translation and definition if available, and ask the user to click on all the
results that are “excellent” examples of the sense in question. When we have three good
examples for a sense, we publish them and remove the sense from rotation. For English, we
use the Twitter API to find a number of English tweets that contain the term of interest to us,
filtering for offensive terms. Tweets are messy text, with slang, misspellings, shortenings,
hashtags, call signs, links, and languages mixing, so a large proportion of English tweets are
unsuitable. Furthermore, experience teaches us that terms with many senses are too diffuse for
this method. We therefore throttle this mode to terms with three or fewer senses in our
database. A future generation of Examples will allow users to match candidate sentences to
the list of available definitions.

For Swahili examples, we query HCS, which contains a large body of text from Swahili
newspapers and books. HCS has parsed and tagged every term for its lemmatic form, which
usually corresponds to the headword field in the Kamusi database. We thus already know that
a sentence with a form such as “alipokitupa” is a candidate to exemplify “tupa”, leaving our
users the task of deciding if it exemplifies a particular sense of “tupa”.

Future development of the Examples mode will build on both Twitter and the experience of
mining HCS. The Indigenous Tweets project identifies people who usually tweet in a minority
language. We will use that information to restrict queries for a language to its known
producers. We will also apply the technique to more conventional text corpora for other
languages, when the data is freely available.

3.6. Alignment. The biggest lexicographical barrier to uniting existing data in pursuit of a
universal multilingual dictionary is the task of aligning concepts across languages. There is no
computational way to know which English sense to choose for a term matched to “bull” from
Vietnamese to English, for example, nor from Indonesian to English, and using the English
spelling connection to link Vietnamese to Indonesian would multiply the risks of error. In
Alignment mode, we compare an individual sense from a dataset we are importing to the
senses of the term it is said to match in a language that is already in the Kamusi system. The
task for the user is to select the correct match, if any. When a threshold number of users
agrees, we consider the relation confirmed. If results are ambiguous, we flag the item for
manual review by a trusted user.

Future work will resolve how we treat items for which no match yet exists in our dataset.
When the consensus is that an item does not have an existing sense match in Kamusi, we
queue the pair for future consideration by a trusted user for the two languages being
compared. A proper entry will need to be created for the corresponding language, which cannot be achieved through crowd microtasking.

3.7. Equivalence. Whether a bilingual pair is confirmed through the Translation mode or Alignment mode, we wish to know the degree of equivalence between the terms. Concepts might be functionally identical, e.g. a male cow will have parallel referents across virtually every language group that raises cattle. Concepts might have substantial similarities, but also notable differences, such that English “bull” as in bluster might be matched to something that has a more malicious connotation in another language, but is nevertheless the closest translation equivalent. Or, terms might be created in one language in the dictionary purely for the purpose of explaining a concept indigenous to another language, such that an explanatory phrase created in Mongolian to encapsulate a “bull” as an optimistic investor. In Equivalence mode, we show a bilingual pair and ask if the terms are parallel, similar, or an explanation on one side of an idea indigenous to the other. When a term passes consensus, the equivalence relation is published to the entries on both sides.

Future work will create a “Difference” mode for users to provide an explanation of the divergence between concepts marked “similar”. This mode will operate along the lines of Definitions, with users able to write their own explanation or vote on explanations provided by others. In addition, the Difference mode will be employed for items shown as synonyms within a language, to distinguish, e.g., the difference between a boat and a ship.

4. Reasoning for method design

The tools discussed in section 3 were devised around a set of principles, some tested and some subject to testing. The methods remain experimental, and will be adjusted as the results become manifest. Notably, the methods discussed herein are not suitable for eliciting terms for which we do not already have a defined concept in a reference language, though future games could be built around processes such as the Rapid Word Collection technique (Moe 2007). In this section, we lay out the logic of the current instantiation.

We begin by noting that “ground truth” does not exist for most lexicographic tasks, in the form of digitized data that is comprehensive and indisputable. Datasets have not been compiled with the rigor we seek for languages that have some documentation, and are not available to science for most. Where data exists, it can never be considered definitive – knowing that a place to hold miscreants captive is a jail does not exclude the possibility that it is a prison, a penitentiary, or in Australia a gaol. However, we can find answers that contain truth, and we can determine when information is clearly wrong. Without getting mired in well-trodden linguistic theory, language is about communicating between people. If one person uses a word and the other person understands what is meant, communication has occurred. If a group of people all understand the same thing from the same word in the same context, that term can be considered part of their language. Roughly stated, a language consists of the set of expressions that are generally mutually agreed upon by its speakers, with the possibility of variations among regions or groups. All English speakers will agree about the basic sense of “water”, however differently pronounced, for example, while Americans will be content to let South Africans have “robots” instead of “stoplights”, and landlubbers accept that they will adopt an existing specialized boating vocabulary should they ever wish to sail.

Linguistic “truth” is pliable, but we posit that language groups can reach agreements that generally work. We deduce consensus when a number of participants repeat information, or when information provided by one is agreed acceptable by others. In many cases, the iterative process of seeking and validating information from a language’s speakers may result over time in data that is qualitatively better than can be gathered by a solo lexicographer.
In modes such as Translation, no user sees what has been proposed by others. We seek a participant’s first reaction, and if other people respond the same way, we can be confident that a common basis for communication has been located. That is, if we ask for a term for a timepiece that is worn on the wrist, and most people answer “watch”, then we have achieved a truth that many speakers of the language share that understanding of that term.

Modes such as Synonyms and Examples limit people to either/or answers. Questions are based on review of imported data or submissions from other users that can be resolved as right or wrong. Here, our premise is that people can usually agree about whether a data element is factual; “wristwatch” is a legitimate alternate term for a timepiece that is worn on the wrist, while “wratch” is not. If voting results are ambiguous, though, we acquire interesting information: we are alerted to the existence of conflicting opinions (Chklovski and Mihalcea 2003, Zhou and Li 2010, Aroyo and Welty 2013). These conflicts reveal non-binary facets of language, such as regional variation. When Boolean questions yield ambiguous answers, the results are queued for human review.

Definitions mode and the upcoming Difference mode rely on a mix of competitiveness and ego. When played as a game, the person who writes a winning definition receives many points, while people who vote for a winning definition each receive a single point. There is therefore substantial incentive to write a good definition if one has not already been proposed, but a disincentive to waste time challenging quality work. Moreover, the writer of a definition is given authorship credit online, but any definition can be challenged; readers thus have the incentive to submit improved definitions and thereby receive authorship credit. Of course, many participants will help improve definitions for the purely altruistic reason of contributing to the knowledge base, as often happens with Wikipedia, but we anticipate that motivating people to contribute for fun and recognition is a faster route to data collection that will nonetheless tend toward excellence over time.

The number of people required for an item to pass a consensus threshold is not firm. For starters, different modes have different considerations; more people must agree on the translation of an English noun than must verify that the translation is also a noun in their language. Within a mode, we can more quickly accept agreement that information is bad than that it is good, since “good” rejected data should someday reemerge, whereas publishing bad information undermines the confidence people can place in our results. It is impossible to say mathematically when validity is established; if one person proposes something, we know it is not yet valid, but what if two other people agree with the answer? Seven? 45? If five agree and four disagree, we should be less confident than seven agreeing and two disagreeing, but how much less, and how long do we keep asking? If four agree and five disagree, do we eliminate the answer? We are still tweaking the algorithm as we learn from experience, with the premise that answers that achieve unanimous support require a lower threshold than answers that generate conflict, and answers that generate too much noise should be passed to a trusted user for manual resolution. Thresholds may be adjusted for languages with few participants, if we can determine the credibility of the individual players, lest game data never acquire enough votes to enter the system. Users also gain trust by consistently being on the consensus side, with their votes acquiring increasing weight. Finally, the threshold can be overridden for a particular trusted user, such as a field linguist, who uses the system as a controlled elicitation tool, rather than as a game.

5. Future work

Additional modes are planned for future crowdsourced microtasking. Among the tasks will be geotagging of usage sightings to build vocabulary clouds for the study of dialect, user contributions of pronunciations geotagged to their location of origin, and cataloguing of user-uploaded photos to illustrate terms and confirm labels across languages. For the present, the chief task is refining the processes based on user experience, and rolling out new languages in
a manner that does not overwhelm our processing capacity, at the level of both management and machine.

6. Conclusions

This paper has spelled out the design and reasoning behind the crowdsourcing features the Kamusi Project has introduce to acquire lexicographical data across numerous languages. Elsewhere we have discussed the underlying structure of the multilingual data, the multiplicity of approaches to data collection, and the specific use of gamification for harvesting social energy in the cause of lexicography. Our underlying premises are that much language knowledge can be affirmed by the consensus of its speakers about what elements achieve a communicative function, and that we can find that consensus by asking well-crafted questions to multiple speakers as targeted microtasks. However, we also find that some crowd results reveal ambiguity; this ambiguity produces valuable information about which data should be channeled for further analysis by crowd methods or designated language specialists. Our distributed approach is intended to elicit data for languages that would not otherwise receive attention from lexicographers, and also to elicit more extensive data than many existing lexicographic projects are able to acquire through expensive and time-consuming processes of data collection in the field or through techniques such as corpus analysis. Crowdsourcing is not meant to be the exclusive method by which data is gathered and validated within the Kamusi system, but it provides an essential path to a wide range of lexical data that would not be available within a realistic timeframe via other methods.

References:


Comparison of Dictionary Use and Corpus Use:
Different Effects on Learning L2 Phrases

Yoshiho Satake
Surugadai University, Japan
satake.yoshiho@surugadai.ac.jp

Abstract
Although the effectiveness of the use of dictionaries (Bishop 1998) and the effectiveness of the use of corpora (Flowerdew 2010) in language learning have been stated, it remains unclear about whether dictionaries and corpora have different effects on language learning. While Satake (2014) states that dictionary users look up and memorize more collocations and corpus users output more collocations, more empirical research is needed to judge how dictionaries and corpora can contribute to language learning. This study investigates whether a dictionary and a corpus have different learning effects when learners do different kinds of tasks. The research questions are as follows: (1) Do a dictionary and a corpus have different effects on different kinds of tasks? (2) Do a dictionary and a corpus have different effects to learners at different levels of proficiency? (3) Do dictionary users and corpus users access and process different information? The two group (dictionary users vs. corpus users) pre-post design was used to analyse the effects of dictionary use and corpus use. At a Japanese university, approximately 30 intermediate and 30 upper-intermediate English learners looked up the target phrases and did different tasks by consulting a dictionary or BNC. The results indicated the following: (1) Dictionary users were better at gap-filling tasks and output fewer phrases. (2) Corpus users, especially upper-intermediate learners were better at the tasks of semantic prosody and analogy. They output more phrases. The findings suggest that considering different learning effects of dictionaries and corpora should lead to more effective vocabulary teaching in classrooms.

Keywords: dictionaries, corpora, data-driven learning, inductive inference, British National Corpus

1. Introduction

Although the effectiveness of the use of dictionaries (e.g., Bishop 1998) and corpora (e.g., Flowerdew 2010) in language learning has been stated, it remains unclear whether dictionaries and corpora have different effects on language learning. While Satake (2014a) says that dictionaries promote learners’ accuracy and corpora promote learners’ fluency, more empirical research is needed to judge how dictionaries and corpora can contribute to language learning. Moreover, further studies are needed regarding whether different tasks and levels have different effects on language learning when learners use dictionaries or corpora.

2. Literature

Regarding the effectiveness of the use of dictionaries, Bishop (1998) states that language learners use dictionaries not only for meanings but also for various information, such as gender, spellings, synonyms, verb, noun, and adjective forms, registers, and pronunciation. Regarding the effectiveness of the use of corpora, Flowerdew (2010) writes that corpora show learners phraseological patterning involving collocations, colligations, and semantic preferences and prosodies.

Concerning the comparison between the effect of corpus and dictionary use on learning languages, Satake (2014a) says that dictionary users look up and memorize more collocations, whereas corpus users output more collocations.
While the strength of both dictionary and corpus use have been stated, more empirical research needs to be conducted to judge the different effects dictionaries and corpora have on language learning.

3. Research questions

This study investigates whether a dictionary and a corpus have different learning effects when learners perform different kinds of tasks. The research questions are as follows:
1. Does the use of a dictionary and a corpus have different effects on different kinds of tasks?
2. Does the use of a dictionary and a corpus have different effects for learners at different levels of proficiency?
3. Do dictionary and corpus users access and process different information?
These questions are investigated by comparing the results of dictionary users with the results of corpus users.

4. Method

A two-group (dictionary users vs. corpus users) pre-post design was used to analyze the effects of dictionary and corpus use.

4.1 Participants

Two classes of Japanese undergraduate students (approximately 60 students) at a private university in Tokyo participated in the study. One class of students who majored in history were intermediate English learners and had reached level B1 in the Common European Framework of Reference for Languages (CEFR); whereas, the other class of students who majored in English and American literature were upper intermediate English learners and had reached level B2 in the CEFR.

4.2 The dictionary

The majority of Japanese learners of English mainly use English–Japanese dictionaries; therefore, the paper-based Shogakukan Progressive English–Japanese dictionary (hereafter, Progressive) (2012) was used for this research because the dictionary is targeted to high school and university students and adult learners and is thus suitable for the participants. Although Japanese university students use more electronic dictionaries than paper dictionaries, this paper dictionary was used to make an appropriate comparison between the effects of dictionary and corpus use because some electronic dictionaries have corpus-like functions, such as extracting co-occurrence expressions. The author copied the pages with the target words and gave the copies to students when they were to be used.

4.3 The corpus

The British National Corpus (BNC) was used because it is a large, balanced corpus that functions as a normative model for learners of English. The BNC is an approximately 100-million-word corpus of British written and spoken English from the latter half of the 20th century. To enable students to use the BNC easily, the Intelligent Tools for Creating and Analysing Electronic Text Corpora for Humanities Research (IntelliText) was introduced. IntelliText was developed by the Centre for Translation Studies at the University of Leeds (Wilson, Hartley, Sharoff, & Stephenson 2010). This publicly available online corpus query system allows users to search corpora, including the BNC. The interface is comparatively user-friendly. Before performing the tasks, the students were given an approximately 20-minute instruction session on how to use IntelliText.

4.4 Tasks
The students were given 15 minutes to complete three types of tasks by consulting the Progressive dictionary or the BNC. First, they read four sentences with preposition errors and underlined search words. An example is shown in (1):

(1) Mary is not interested in dance. She is indifferent from dance.

The students searched for the underlined words in the Progressive dictionary or the BNC, read the example sentences or the concordance lines of the underlined search words, and performed error correction. They kept a record of the example phrases or sentences they consulted. The students were given five minutes for this task. This task was used to examine whether a dictionary or a corpus would help the learners correct their errors by using an enumerative induction, which is a generalization based on reasoning from particular examples to all examples.

Second, the students searched for four target words, such as “utterly” and “outbreak,” in the Progressive dictionary or the BNC to determine whether the words were used with any semantic associations. They kept a record of the example phrases or sentences they consulted. They were given five minutes for this task. This task was used to examine whether a dictionary or a corpus would help the learners recognize semantic prosody, “a form of meaning which is established through the proximity of a consistent series of collocates” (Louw 2000: 57). To be more specific, semantic prosody relates to the way in which some words carry positive or negative associations because of certain frequent combinations of words.

Third, the students searched for two target phrase frames, such as verb-into-noun (i.e., any verbs can take the place of a “verb” and any nouns can take the place of a “noun”), in the Progressive dictionary or the BNC to determine whether the verbs and nouns used for each word frame had something in common. They kept a record of the example phrases or sentences they consulted. The students were given five minutes for this task. This task was used to examine whether a dictionary or a corpus would help the learners recognize the shared characteristics certain phrase frames have.

These three types of tasks were used to examine whether a dictionary and a corpus have different effects on different kinds of tasks. In addition, it was useful for the students to perform the tasks because it is difficult for learners to use English prepositions (e.g., Bitchener et al. 2005, Murata & Isahara, 2004). Furthermore, learning semantic prosody can help learners improve their vocabulary knowledge (O’Keeffe et al. 2007), and learning phrase frames can help them create various phrases easily because the category that builds sentences may consist of a very limited number of phrase frames (Nattinger & DeCarrico 1992).

4.5 Tests
Before and after the tasks, students were given 10 minutes to take a pre-test and a post-test without consulting a dictionary or a corpus. By comparing the results of the pre-test and the post-test, the author examined whether the above tasks were effective for learning the target words, phrases, and frames.

First, the students read four sentences with blanks, which they were supposed to fill in with the appropriate prepositions. An example is shown in (2).

(2) He likes to be alone and is indifferent (          ) the feelings of others.

The students learned the target phrases with prepositions when they completed the first task. The students were given two minutes to complete four sentences. Second, the
students were told to write as many phrases as possible using the four underlined search words in the first tasks. They were given one minute for each search word, for a total of four minutes. Third, the students answered four multiple-choice questions. An example is shown in (3).

(3) She felt utterly ( ).
① stupid
② normal
③ great

The questions were related to the semantic prosody task. The students were given two minutes to answer the questions.

Fourth, the students were told to translate four phrases into Japanese. The phrases used the two target phrase frames, such as verb-into-noun, they learned when they performed the third task. An example is shown in (4).

(4) come into force

The students were given two minutes to translate them into Japanese.

4.6 Procedures

Two classes of students were given 10 minutes to take the pre-test. They were then given 15 minutes to perform the three tasks described above. Each class was divided into two groups: one group of students consulted the Progressive dictionary and the other group consulted the BNC while they completed the tasks. After the task stage, the students were given 10 minutes to take the post-test, which was the same as the pre-test. The next week, the same procedure was repeated for the different set of target words, phrases, and frames. Furthermore, the resource they used was different from the one they had used the previous week: students who consulted the Progressive dictionary the first week consulted the BNC the second week and vice versa. The two resources were exchanged to minimize the influence of the different levels of English skills the two groups in each class might have had. Additionally, from a pedagogical viewpoint, the exchange had an advantage because the students could learn the difference between searching a dictionary and searching a corpus.

The author compared the results from the pre-test and post-test to examine whether the tasks were effective for learning the target words, phrases, and frames. The author also compared the results of intermediate students and the results of upper-intermediate students to investigate whether a dictionary and a corpus have different effects for learners with different levels of English skills. Several statistical tests were used to determine whether a significant difference existed between the results from the pre-test and post-test.

5. Results

To examine whether a significant difference existed between the results from the pre-tests and the post-tests, the author ran two statistical tests. To decide which statistical test should be used, the author conducted Bartlett’s test to examine whether variances were equal across the results of the two kinds of tests. At less than 5%, variances were not equal across the results from the pre-tests and the post-tests for both classes. Therefore, the author conducted the Friedman test. At less than 1%, both classes showed a significant difference between the results from the pre-tests and the post-tests. For the first, the third, and the fourth questions, the author conducted McNemar’s test to examine whether a significant difference existed between the number of students who gave wrong answers on the pre-tests but correct answers on the post-tests and the number of students who gave correct answers on the pre-tests but wrong answers on the post-tests. If the first number was significantly greater than the second number, it can be said that the tasks were effective for learning the target words, phrases, and frames.
5.1 Enumerative induction tasks
In all groups, the number of students who gave wrong answers on the pre-tests but correct answers on the post-tests and the number of students who gave correct answers on the pre-tests but wrong answers on the post-tests were significantly different at less than 1%. Therefore, we can say that the first number was significantly greater than the second number (Table 1). In both classes, the p-value for dictionary users was lower than the p-value for corpus users. Therefore, it can be said that the enumerative induction tasks were effective for both dictionary users and corpus users to learn the target phrases, but they were more effective for dictionary users than for corpus users. All the target phrases were in the Progressive dictionary.

Table 1 Enumerative induction tasks: Pre-test and post-test results

<table>
<thead>
<tr>
<th></th>
<th>Intermediate dictionary users (n=28)</th>
<th>Intermediate corpus users (n=27)</th>
<th>Upper-intermediate dictionary users (n=28)</th>
<th>Upper-intermediate corpus users (n=28)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posttest</td>
<td>Posttest</td>
<td>Posttest</td>
<td>Posttest</td>
<td>Posttest</td>
</tr>
<tr>
<td>correct</td>
<td>38</td>
<td>51</td>
<td>51</td>
<td>52</td>
</tr>
<tr>
<td>wrong</td>
<td>10</td>
<td>9</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>McNemar’s test p-value</td>
<td>7.658e-08</td>
<td>0.002055</td>
<td>7.648e-06</td>
<td>0.0003466</td>
</tr>
</tbody>
</table>

5.2 Output of the target phrases
In all groups, the students output more phrases using the target words in the post-tests than in the pre-tests (Table 2). In both classes, corpus users output more phrases using the target words than did the dictionary users. Accordingly, we can say that the enumerative induction tasks were effective for both the dictionary users and corpus users to promote the output of phrases using the target words, but they were more effective for corpus users than for dictionary users. Considering the difference in English skills, the tasks were more effective for upper-intermediate corpus users than for intermediate corpus users.

Table 2 Number of phrases output by students using the target words

<table>
<thead>
<tr>
<th></th>
<th>Intermediate dictionary users (n=28)</th>
<th>Intermediate corpus users (n=27)</th>
<th>Upper-intermediate dictionary users (n=28)</th>
<th>Upper-intermediate corpus users (n=28)</th>
</tr>
</thead>
<tbody>
<tr>
<td>pretests</td>
<td>66</td>
<td>64</td>
<td>76</td>
<td>69</td>
</tr>
<tr>
<td>posttests</td>
<td>117 (up 77.3%)</td>
<td>133 (up 107.8%)</td>
<td>118 (up 55.3%)</td>
<td>149 (up 115.9%)</td>
</tr>
</tbody>
</table>

5.3 Semantic-prosody tasks
In the groups of corpus users, the number of students who gave wrong answers on the pre-tests but correct answers on the post-tests and the number of students who gave correct answers on the pre-tests but wrong answers on the post-tests were significantly different at less than 5%. Therefore, it can be said that the first number was significantly greater than the second number (Table 3). In the groups of dictionary users, however, the first and second numbers were not significantly different. Therefore, we can say that the semantic prosody tasks were effective for corpus users and were not effective for dictionary users. Considering the difference in English skills, the tasks were more effective for upper-
intermediate corpus users than for intermediate corpus users. Not all the target phrases were in the copies of the Progressive dictionary.

Table 3 Semantic prosody tasks: Pre-test and post-test results

<table>
<thead>
<tr>
<th></th>
<th>Intermediate dictionary users (n=28)</th>
<th>Intermediate corpus users (n=27)</th>
<th>Upper-intermediate dictionary users (n=28)</th>
<th>Upper-intermediate corpus users (n=28)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Posttest</td>
<td>Posttest</td>
<td>Posttest</td>
<td>Posttest</td>
</tr>
<tr>
<td></td>
<td>correct</td>
<td>wrong</td>
<td>correct</td>
<td>wrong</td>
</tr>
<tr>
<td>Pre test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correct</td>
<td>40</td>
<td>8</td>
<td>41</td>
<td>8</td>
</tr>
<tr>
<td>wrong</td>
<td>11</td>
<td>39</td>
<td>20</td>
<td>29</td>
</tr>
<tr>
<td>McNemar’s test p-value</td>
<td>0.6464</td>
<td>0.03764</td>
<td>0.4725</td>
<td>0.0001264</td>
</tr>
</tbody>
</table>

5.4 Phrase frame tasks
In the groups of corpus users, the number of students who gave wrong answers on the pre-tests but correct answers on the post-tests and the number of students who gave correct answers on the pre-tests but wrong answers on the post-tests were significantly different at less than 5%. Therefore, it can be said that the first number was significantly greater than the second number (Table 4). In the groups of dictionary users, however, the first and second numbers were not significantly different. Therefore, we can say that the phrase frames tasks were effective for corpus users and were not effective for dictionary users. Considering the difference in English skills, the tasks were more effective for upper-intermediate corpus users than intermediate corpus users. Not all the target phrases were in the copies of the Progressive dictionary.

Table 4 Phrase frame tasks: pre-test and post-test results

<table>
<thead>
<tr>
<th></th>
<th>Intermediate dictionary users (n=28)</th>
<th>Intermediate corpus users (n=27)</th>
<th>Upper-intermediate dictionary users (n=28)</th>
<th>Upper-intermediate corpus users (n=28)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Posttest</td>
<td>Posttest</td>
<td>Posttest</td>
<td>Posttest</td>
</tr>
<tr>
<td></td>
<td>correct</td>
<td>wrong</td>
<td>correct</td>
<td>wrong</td>
</tr>
<tr>
<td>Pre test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correct</td>
<td>10</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>wrong</td>
<td>5</td>
<td>83</td>
<td>8</td>
<td>84</td>
</tr>
<tr>
<td>McNemar’s test p-value</td>
<td>0.07364</td>
<td>0.01333</td>
<td>0.4795</td>
<td>2.668e-06</td>
</tr>
</tbody>
</table>

6. Discussion and conclusion
The results show that for learning the target phrases, the enumerative induction tasks whose target phrases were in the copies of the Progressive dictionary were more effective for dictionary users than for corpus users. The semantic prosody tasks and phrase frame tasks whose target phrases were not in the copies of the Progressive dictionary were more effective for corpus users than dictionary users. We can say that (1) it was more effective to use a dictionary when the target phrases were in the dictionary and (2) it was more effective
to use a corpus when the target phrases were not in the dictionary. The first finding occurs because consulting a dictionary is easier than consulting a corpus as the students generally used dictionaries and did not have previous experience using a corpus, considering that both the dictionary and corpus contained the target phrases for the enumerative induction tasks. The second finding occurs because the corpus contained information that the dictionary did not contain. Furthermore, we can say that (3) upper-intermediate students could more effectively use a corpus to find the information for tasks with target phrases that were not in the dictionary. This finding occurs because using a corpus to acquire the needed information required higher English skills as it is difficult for students to interpret concordance lines (Boulton 2009, Lavid 2007, Liu & Jiang 2009, Varley 2009). Furthermore, it can be said that (4) a corpus could promote more students’ output of phrases using the target words than a dictionary. This matches the results from two previous studies by Satake (2014a, 2014b), possibly because more exposure to the target words, phrases, and frames helped students improve their productive knowledge.

As dictionaries and corpora have different advantages, teachers should make an informed choice when selecting learning resources to enhance the learning effect as students look up vocabulary in classrooms. Furthermore, as dictionary users learned the target phrases better if these phrases were present in the dictionary, learners could improve their vocabulary if dictionaries contained detailed information regarding headwords, such as semantic prosody or phrase frames, which is not always included.

References
House M.D. Corpus Analysis: A Linguistic Intervention of Contemporary American English

Locky LAW
The Hong Kong Polytechnic University
13902936r@connect.polyu.hk / Lx3h@yahoo.com

Abstract
Television drama, despite its enormous popularity across the globe, has rarely received attention from the linguistics field. The dearth of research into television drama dialogue is further exposed by the thriving contributions from various other fields such as philosophy, psychology, cultural studies and media studies. This paper seeks to promote research interest in this unique mediated text by selecting the renowned medical dramedy House M.D. as research subject and comparing its 927,922-word House M.D. pure dialogue corpus (HMDC) to both the 450-million-word Corpus of Contemporary American English (COCA) and its 95-million-word spoken subcorpus (COCA Spoken) using an adaptation of Bednerak’s (2011) ranked frequency list method. Using WordSmith Tools in the calculation of n-gram (n = 1, 2, 3) at the words/clusters level, the findings indicate that HMDC is more interpersonal than COCA and has a closer resemblance to COCA Spoken than to COCA. HMDC also contains 3.4 times more negativity than COCA Spoken and 2.8 times more than COCA. As such, viewers are presented with English far more interpersonal, as well as involving significantly more disagreement than one will encounter in the real world. This study not only shows similarities and differences between House M.D. and contemporary American English, but also provides a preview of the huge potential in television drama-related research.

Keywords: TV drama, House M.D., corpus linguistics, ngram, COCA, ranked frequency list

Introduction
Television drama and film have long played an influential role in the entertainment industry. In this modern era, they have become an essential part of the living style for the global population (Heritage 2013; Cox 2013). Apart from their popularity, these two forms of performance entertainment share resemblance in many aspects, including casting, acting, cinematography, post-production editing and screenwriting. As far as film is concerned, these aspects have provided various topics for linguistic research such as intonation, gestures, facial expressions and visual images (Wojcik 2004; Petric 1993; Condon and Ogston 1966; Toolan 2001; Kress and van Leeuwen 2006 [1996]; Dan-dan 2011). Yet the history of linguistic studies has always shown an inclination towards literary texts (poetry, traditional drama and film-scripts), and thus the language in television drama has not been taken seriously in the linguistics field (Bednarek 2010; Pennycook 2007; Androustopoulos 2012), so much to a point that it is marked “equally as popular” as it is “devalued” (Bignell and Lacey 2005:3). This linguistic negligence, or repulsion, is further exposed by the thriving contributions of research on television drama dialogue from various other fields (Bednarek 2010) such as philosophy (Jacoby and Irwin 2008), psychology (Cascio and Martin 2011; Jamieson 2011; Whitbourne 2012; Clyman 2009), cultural studies (Cover 2004; Chua 2008; Song 2010), media studies (Chua 2008; Munt 2006) and medical humanities (Goodier and Arrington 2007), to name a few. With over 3000 non-linguistic journal articles on television studies between 1995 and 2004 (Allen 2004; Bednarek 2010), the push for television drama-related research from the linguistic department, in summary, has been far less than adequate.

Realizing “the urgent need … for a treatment of fictional cinema and television from various linguistic perspectives” (Piazza, Bednarek and Rossi 2011:2), several scholars have contributed in the investigation of dialogues of some of the world’s best television dramas,
namely *The West Wing* (Chamber 2003), *Sex and The City* (Bubel 2006), *Dawson’s Creek* (Quaglio 2008), *Friends* (Quaglio 2009) and *Gilmore Girls* (Bednarek 2010). By adopting a range of approaches from critical discourse analysis, corpus linguistics, to corpus stylistics, their efforts have successfully yielded significant linguistic insights. These attempts have demonstrated that television drama is not only indeed a plentiful resource waiting to be explored and exploited, but also a unique form of “mediated” text rich in language and in culture (Richardson 2010:177) – “the true heir to great literature” in literary agent Steven Axelrod’s words (Lavery 2012) – which deserves the same level of attention as its cinematic counterpart. This paper aims to add a push of opposing force to such film/television drama developmental disproportion by drawing comparisons between a corpus created using revised fan scripts of a renowned medical detective television dramedy *House M.D.*, the Corpus of Contemporary American English (COCA) and its spoken subcorpus.

There are 3 main sections in this paper. The first section outlines the background of *House M.D.*, provides key reasons for supporting the choice of television drama and lists the research questions. The next section describes the reference corpus and the construction of HMDC as well as the methods involved. In the third section, data is analysed at the words/clusters level and discussed quantitatively. Finally, the last section concludes by unveiling some thematic characteristics of the TV drama captured by corpus linguistic analysis.

**Background and research questions**

In order to understand the choice of data for this study, it is important to know the background of television drama. *House M.D.* is an American television medical ‘dramedy’ stretching eight seasons with a total of 177 episodes aired on the FOX Network from 16th November 2004 to 21st May 2012 (Wikia n.d.), created by Primetime Emmy Awards Outstanding Writing for a Drama Series winner David Shore and executively produced by famous television writers including film director of *Valkyrie and X-Men* Bryan Singer and twice winner of the Golden Globe Best Performance by an Actor in a Television Series – Drama, Hugh Laurie (IMDb n.d.). For each season of *House M.D.*, character arcs were first mapped out by a team of writers before individual writers created their respective episodes (Wild 2005a), with diagnoses and accuracy of medical cases consulted by actual medical advisers (Gonzalez 2009; Oldenburg 2005; Woznicki 2005). It is a highly popular and successful television drama which has attracted viewers worldwide. In 2008, it was one of the top-ten rated shows in the United States as well as the most watched television program in the world (AFP 2009). By 2011, it had been viewed by a spectacular 81.8 million in 66 countries (The Telegraph 2011), placing Hugh Laurie’s name on the Guinness Book of Records since 2011 as the world’s Most-Watched (Leading) Man On Television and the 2nd on Forbes’s list of the Highest-Paid TV Actors in 2012 (Pomerantz 2012) at $400,000 (£247,230) per episode (Guinness World Records 2011). As of May 2015, it has received an 8.9 / 10 rating from 255,714 users on IMDb.com (IMDb n.d.). Bednarek (2010) argues that popularity of television and programmes alone is worthy of study due to its significant impact on our daily lives and societies. Bignell and Lacey (2005:6) echo this argument, “it is television’s very familiarity, and its conventional focus upon the familiar, the present time and the everyday, that opens up alternative formal and stylistic possibilities.”

A series of research questions were asked and addressed specifically with respect to *House M.D.*, the television dramedy and COCA:

How do dialogues in *House M.D.* differ/resemble contemporary American English?  
Does the language use in *House M.D.* mimic more towards the spoken or written form contemporary American English and thus which corpus then has a closer resemblance with *House M.D.?*  
What can be unveiled about *House M.D.* through a corpus linguistic approach?
To answer these questions, COCA – the reference corpus of American English and its subcorpus along with a corpus from *House M.D.* will be used.

**Data and methodology**

The Reference Corpus: COCA

As “the largest freely-available corpus of English, and the only large and balanced corpus of American English” (Davies 2008), COCA contains more than 450 million words in 189,431 texts equally divided in 5 genres: spoken, fiction, popular magazines, newspapers and academic journals, including 20 million words each year from 1990-2012 with the most recent addition of texts (Apr 2011 - Jun 2012) completed in June 2012 (Davies 2008). The spoken part of COCA (hereafter referred to as COCA Spoken) contains 95 million words [95,385,672] of transcripts of unscripted conversation from more than 150 different TV and radio programs such as *All Things Considered* (NPR), *Newshour* (PBS), *Good Morning America* (ABC), *Today Show* (NBC), *60 Minutes* (CBS), *Hannity and Colmes* (Fox), *Jerry Springer*, etc (Davies 2008). COCA Spoken is arguably an authentic representation of actual spoken conversation given its data is about 95% unscripted with “overwhelming” amount of discourse markers (Davies 2008; 2014).

The Data: House M.D. Corpus (HMDC) Construction

![Figure 1 Original fan script of House M.D. Season 1 Episode 1](image)

The construction of the 927,922-word *House M.D.* Corpus (hereafter referred to as HMDC) involves the data collection of *House M.D.* fan scripts of every episode from the internet (therefore not the original screenwriters’ scripts) as shown in Figure 1, followed by a removal of all non-dialogue elements such as fade-ins, scene headings, action sequences, scene transitions, mood brackets, parentheticals, commercial tags and character name tags, resulting in ‘pure’ dialogues stored as txt-format in 177 individual files (one file per episode) which form a raw, unscripted and unannotated version of HMDC, as shown in Figure 2.
Figure 2 'Cleaned' script of House M.D. Season 1 Episode 1

To improve accuracy of the transcribed dialogues in the HMDC, every line has been manually checked against the actual lines performed by the actors in the television series after watching all episodes at least 8 times (till Oct 2014), and partial corpus checks were performed repeatedly throughout the entire duration of the research project whenever possible and necessary. Spell checks, in particular, were greatly assisted by Google and internet resources. Although the achievement of a 100% accurate corpus remains highly unlikely, it is of my belief that such longitudinal effort has helped to minimise negative effect to analytical outcomes caused by corporal impurities.

Methodology

To answer the research questions, an empirical study of ngrams from HMDC, COCA and COCA Spoken was performed. Ngrams were selected because of its sensitivity in text-types discrimination (Stubbs and Barth 2003; Bednarek 2011; O’Keeffe, McCarthy, and Carter 2007; Biber 2009).

Ngrams data such as the non-case sensitive 2-grams, 3-grams, 4-grams and 5-grams of the (approximately) 1,000,000 most frequent n-grams are available for free download (Davies 2011) whereas ngrams of the COCA Spoken, 1-gram of COCA and HMDC must be calculated from the raw text data using Wordsmith Tools 6 (Scott 2014). The duration of each ngram calculation ranged from 20 minutes to 20 hours and increased significantly as the tokens of ngram increased.

Ranked frequency lists, which are lists of ngrams/clusters ranked in terms of frequency (Bednarek 2011), will be used to show the degree of similarities and differences of their occurrences while noting the fact that “a similar frequency of [an ngram] does not necessarily indicate that [it] is used with similar functions in two corpora” (Bednarek 2011:15).

Using rank differences, the following three comparisons were made:

- HMDC 1-gram, 2-gram and 3-gram (hereafter referred to as 1-to-3-grams) with respect to COCA’s 1-to-3-grams and vice versa,
- HMDC 1-to-3-gram with respect to COCA Spoken’s 1-to-3-grams and vice versa,
- Negativity / positivity of 3-grams HMDC with COCA and COCA Spoken, and vice versa.
## Findings and discussion

HMDC VS COCA Ngrams

### Table 1 HMDC VS COCA 1-gram

<table>
<thead>
<tr>
<th>HMDC 1-gram</th>
<th>COCA 1-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word</strong></td>
<td><strong>Freq.</strong></td>
</tr>
<tr>
<td>YOU</td>
<td>36,360</td>
</tr>
<tr>
<td>I</td>
<td>29,182</td>
</tr>
<tr>
<td>THE</td>
<td>27,437</td>
</tr>
<tr>
<td>IS</td>
<td>23,616</td>
</tr>
<tr>
<td>TO</td>
<td>21,979</td>
</tr>
<tr>
<td>A</td>
<td>26,201</td>
</tr>
<tr>
<td>IT</td>
<td>17,918</td>
</tr>
<tr>
<td>NT</td>
<td>14,485</td>
</tr>
<tr>
<td>THAT</td>
<td>11,959</td>
</tr>
<tr>
<td>AND</td>
<td>11,327</td>
</tr>
<tr>
<td>DO</td>
<td>9,519</td>
</tr>
<tr>
<td>OF</td>
<td>9,814</td>
</tr>
<tr>
<td>IS</td>
<td>9,211</td>
</tr>
<tr>
<td>HE</td>
<td>9,103</td>
</tr>
<tr>
<td>WE</td>
<td>8,307</td>
</tr>
<tr>
<td>IN</td>
<td>8,431</td>
</tr>
<tr>
<td>NOT</td>
<td>7,579</td>
</tr>
<tr>
<td>RE</td>
<td>7,251</td>
</tr>
<tr>
<td>WHAT</td>
<td>7,162</td>
</tr>
<tr>
<td>YOUR</td>
<td>7,079</td>
</tr>
<tr>
<td>HAVE</td>
<td>6,563</td>
</tr>
<tr>
<td>SHE</td>
<td>6,338</td>
</tr>
<tr>
<td>FOR</td>
<td>6,332</td>
</tr>
<tr>
<td>ME</td>
<td>6,208</td>
</tr>
<tr>
<td>WAB</td>
<td>6,159</td>
</tr>
<tr>
<td>M</td>
<td>5,942</td>
</tr>
<tr>
<td>BE</td>
<td>5,926</td>
</tr>
<tr>
<td>NO</td>
<td>5,907</td>
</tr>
<tr>
<td>HER</td>
<td>5,777</td>
</tr>
<tr>
<td>ON</td>
<td>5,615</td>
</tr>
<tr>
<td>THIS</td>
<td>5,520</td>
</tr>
<tr>
<td>MY</td>
<td>5,435</td>
</tr>
<tr>
<td>WITH</td>
<td>5,089</td>
</tr>
</tbody>
</table>

Table 1 shows the frequency and percentage of HMDC and COCA 1-gram sorted according to frequency. This table is then translated into Table 2 to show ranked frequency, both sorted with respect to (wrt) HMDC and COCA.
Comparing the top twenty 1-gram of HMDC and COCA in Table 2, there are thirteen out of twenty 1-grams ranked among top 20 on both lists (cells highlighted in yellow), implying that HMDC does often contain some of the most frequent words in American English. Though there are some differences in ranking for certain words, particularly You, and I (among the shared thirteen 1-grams), N’t, Do, We, Not, ’re, What and Your (double digit rank difference) being more prominent in HMDC than in COCA, the total rank differences (267 and 261) are similar, implying a minor degree of difference between these two corpora. At this point, early result could reflect that House M.D. appears to be more interpersonal and 1st and 2nd person singularly focused than the norm in general written and spoken American English, but higher token ngrams must be considered.

<table>
<thead>
<tr>
<th>Ngram</th>
<th>Rank wrt HMDC</th>
<th>Rank wrt COCA</th>
<th>Rank Difference</th>
<th>Rank wrt COCA</th>
<th>Rank wrt HMDC</th>
<th>Rank Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>You</td>
<td>1</td>
<td>14</td>
<td>13</td>
<td>The</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>I</td>
<td>2</td>
<td>10</td>
<td>8</td>
<td>And</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>The</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>Of</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>’s</td>
<td>4</td>
<td>9</td>
<td>5</td>
<td>To</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>To</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>A</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>A</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>#</td>
<td>6</td>
<td>49</td>
</tr>
<tr>
<td>It</td>
<td>7</td>
<td>11</td>
<td>4</td>
<td>In</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>N’t</td>
<td>8</td>
<td>31</td>
<td>23</td>
<td>That</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>That</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>’s</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>And</td>
<td>10</td>
<td>2</td>
<td>8</td>
<td>I</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Do</td>
<td>11</td>
<td>36</td>
<td>25</td>
<td>It</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>Of</td>
<td>12</td>
<td>3</td>
<td>9</td>
<td>Is</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>Is</td>
<td>13</td>
<td>12</td>
<td>1</td>
<td>For</td>
<td>13</td>
<td>23</td>
</tr>
<tr>
<td>He</td>
<td>14</td>
<td>16</td>
<td>2</td>
<td>You</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>We</td>
<td>15</td>
<td>27</td>
<td>12</td>
<td>Was</td>
<td>15</td>
<td>25</td>
</tr>
<tr>
<td>In</td>
<td>16</td>
<td>7</td>
<td>9</td>
<td>He</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>Not</td>
<td>17</td>
<td>30</td>
<td>13</td>
<td>On</td>
<td>17</td>
<td>30</td>
</tr>
<tr>
<td>’re</td>
<td>18</td>
<td>79</td>
<td>61</td>
<td>With</td>
<td>18</td>
<td>33</td>
</tr>
<tr>
<td>What</td>
<td>19</td>
<td>39</td>
<td>20</td>
<td>As</td>
<td>19</td>
<td>79</td>
</tr>
<tr>
<td>Your</td>
<td>20</td>
<td>69</td>
<td>49</td>
<td>At</td>
<td>20</td>
<td>66</td>
</tr>
</tbody>
</table>

Total rank difference 267
Total rank difference 261
<table>
<thead>
<tr>
<th>N</th>
<th>Word</th>
<th>Freq.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IT'S</td>
<td>6,720</td>
<td>0.72</td>
</tr>
<tr>
<td>2</td>
<td>I'M</td>
<td>5,857</td>
<td>0.63</td>
</tr>
<tr>
<td>3</td>
<td>YOU'RE</td>
<td>5,100</td>
<td>0.55</td>
</tr>
<tr>
<td>4</td>
<td>DO NOT</td>
<td>4,792</td>
<td>0.61</td>
</tr>
<tr>
<td>5</td>
<td>HE'S</td>
<td>3,000</td>
<td>0.32</td>
</tr>
<tr>
<td>6</td>
<td>SHE'S NOT</td>
<td>2,379</td>
<td>0.26</td>
</tr>
<tr>
<td>7</td>
<td>THAT'S</td>
<td>2,371</td>
<td>0.26</td>
</tr>
<tr>
<td>8</td>
<td>IN THE</td>
<td>2,089</td>
<td>0.22</td>
</tr>
<tr>
<td>9</td>
<td>SHE</td>
<td>2,088</td>
<td>0.22</td>
</tr>
<tr>
<td>10</td>
<td>CAN'T</td>
<td>2,012</td>
<td>0.22</td>
</tr>
<tr>
<td>11</td>
<td>I DO</td>
<td>1,985</td>
<td>0.21</td>
</tr>
<tr>
<td>12</td>
<td>DID NOT</td>
<td>1,984</td>
<td>0.21</td>
</tr>
<tr>
<td>13</td>
<td>S-A</td>
<td>1,722</td>
<td>0.18</td>
</tr>
<tr>
<td>14</td>
<td>YOU DO</td>
<td>1,650</td>
<td>0.17</td>
</tr>
<tr>
<td>15</td>
<td>DOES NOT</td>
<td>1,564</td>
<td>0.16</td>
</tr>
<tr>
<td>16</td>
<td>ARE YOU</td>
<td>1,475</td>
<td>0.15</td>
</tr>
<tr>
<td>17</td>
<td>TO BE</td>
<td>1,455</td>
<td>0.16</td>
</tr>
<tr>
<td>18</td>
<td>THERE'S</td>
<td>1,434</td>
<td>0.15</td>
</tr>
<tr>
<td>19</td>
<td>IF YOU</td>
<td>1,432</td>
<td>0.15</td>
</tr>
<tr>
<td>20</td>
<td>WE'RE</td>
<td>1,484</td>
<td>0.15</td>
</tr>
<tr>
<td>21</td>
<td>GOING TO</td>
<td>1,341</td>
<td>0.13</td>
</tr>
<tr>
<td>22</td>
<td>I WAS</td>
<td>1,289</td>
<td>0.13</td>
</tr>
<tr>
<td>23</td>
<td>WANT TO</td>
<td>1,244</td>
<td>0.13</td>
</tr>
<tr>
<td>24</td>
<td>ON THE</td>
<td>1,232</td>
<td>0.13</td>
</tr>
<tr>
<td>25</td>
<td>OF THE</td>
<td>1,213</td>
<td>0.13</td>
</tr>
<tr>
<td>26</td>
<td>THIS IS</td>
<td>1,187</td>
<td>0.13</td>
</tr>
<tr>
<td>27</td>
<td>WHAT'S</td>
<td>1,174</td>
<td>0.13</td>
</tr>
<tr>
<td>28</td>
<td>IT WAS</td>
<td>1,161</td>
<td>0.12</td>
</tr>
<tr>
<td>29</td>
<td>HAVE TO</td>
<td>1,137</td>
<td>0.12</td>
</tr>
<tr>
<td>30</td>
<td>DO YOU</td>
<td>1,122</td>
<td>0.12</td>
</tr>
<tr>
<td>31</td>
<td>I LL</td>
<td>1,086</td>
<td>0.12</td>
</tr>
<tr>
<td>32</td>
<td>YOU HAVE</td>
<td>1,021</td>
<td>0.11</td>
</tr>
<tr>
<td>33</td>
<td>TO THE</td>
<td>1,020</td>
<td>0.11</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>of the</td>
</tr>
<tr>
<td>2</td>
<td>in the</td>
</tr>
<tr>
<td>3</td>
<td>to the</td>
</tr>
<tr>
<td>4</td>
<td>on the</td>
</tr>
<tr>
<td>5</td>
<td>and the</td>
</tr>
<tr>
<td>6</td>
<td>to be</td>
</tr>
<tr>
<td>7</td>
<td>at the</td>
</tr>
<tr>
<td>8</td>
<td>for the</td>
</tr>
<tr>
<td>9</td>
<td>in a</td>
</tr>
<tr>
<td>10</td>
<td>do n't</td>
</tr>
<tr>
<td>11</td>
<td>with the</td>
</tr>
<tr>
<td>12</td>
<td>from the</td>
</tr>
<tr>
<td>13</td>
<td>it was</td>
</tr>
<tr>
<td>14</td>
<td>of a</td>
</tr>
<tr>
<td>15</td>
<td>that the</td>
</tr>
<tr>
<td>16</td>
<td>as a</td>
</tr>
<tr>
<td>17</td>
<td>is a</td>
</tr>
<tr>
<td>18</td>
<td>going to</td>
</tr>
<tr>
<td>19</td>
<td>by the</td>
</tr>
<tr>
<td>20</td>
<td>and i</td>
</tr>
<tr>
<td>21</td>
<td>it is</td>
</tr>
<tr>
<td>22</td>
<td>with a</td>
</tr>
<tr>
<td>23</td>
<td>i think</td>
</tr>
<tr>
<td>24</td>
<td>for a</td>
</tr>
<tr>
<td>25</td>
<td>he was</td>
</tr>
<tr>
<td>26</td>
<td>one of</td>
</tr>
<tr>
<td>27</td>
<td>you know</td>
</tr>
<tr>
<td>28</td>
<td>is the</td>
</tr>
<tr>
<td>29</td>
<td>did n't</td>
</tr>
<tr>
<td>30</td>
<td>was a</td>
</tr>
<tr>
<td>31</td>
<td>i was</td>
</tr>
<tr>
<td>32</td>
<td>out of</td>
</tr>
<tr>
<td>33</td>
<td>this is</td>
</tr>
</tbody>
</table>
Comparing the top twenty 2-gram of HMDC and COCA in Table 4, only three out of twenty 2-grams ranked among top twenty on both lists, in addition to that, rank difference for some of the top twenty 2-grams in HMDC has increased drastically in COCA, most noticeably for I’m, You’re, ‘s not, ‘s a (see Figure 3), and We’re, showing a significant gap in the frequency of occurrence of these 2-grams. Subsequently, five out of twenty 2-grams namely It’s, That’s, She’s, Can’t and There’s are not found in COCA.

Such major rank discrepancies have arisen from contraction alternatives. For instance, 2-gram I’m ranks 352,014 (frequency 81) in COCA, however, if the alternative form I am (rank 170, frequency 91,303) is considered all together, the total frequency will reach 91,384 and rank 170, thus greatly reducing the rank difference. As convenient and elegant this method may seem, one must choose between being faithful to the given data to maintain individuality of instances and being adaptive to variations of instances. Due to the respect for originality of the corpora, I have decided to keep the data as it is. That being said, when rank difference of an ngram is strikingly unreasonable and the respective contraction alternatives are available in a corpus, rank difference will only be taken as a sign of irregularity rather than persuasive statistical evidence.

Further inspection of the COCA 2-grams finds instances of I’m I (rank 216,511), You’re you (rank 387,273), We’re we (rank 542,449), There’s there (rank 582,027) as well as multiple instances of That’s that/i/you/Mr/it. Evidently, COCA 2-grams by Davis (2011) consider the likes of I—’m and You—’re as separate words and as one word but only consider There—’s
and *That*—*’s as one word and not separate words, resulting in a case of “not found”. This may or may not be a standalone case, and it is not possible to account for such irregularity without knowing the algorithm adopted by the author, but the total rank differences (968,777 excluding “not found” and 1,807) have provided certain signs of confirmation for a divergence of phrases found in HMDC from the COCA – which contains both written and spoken American English, in terms of its abundance of contracted PRON+VERB 2-grams commonly found in spoken English and the lack of PREP-DET 2-grams that “dominated” the top twenty 2-grams of COCA (Stubbs and Barth 2003:70).

<table>
<thead>
<tr>
<th>N</th>
<th>Concordance</th>
<th>Set</th>
<th>Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>*be Listena in the cream cheese. 17 *’s a stupid number. *What *’s Chase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>*clock in the shape of Africa. Adoption *’s a cheat, remember? *There <em>’s no real</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>*the hood. I’m not suggesting *Alie *’s a substitute, but he would make a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td><em>Amber. No. Why not</em> *Amber *’s a stripper name. Find me one</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>*are shutting down. Also ... Amber *’s a racist. You knew they’d get</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>*’s a real thing. Meningitis. Anemia *’s a stretch. Page Dr. James Wilson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>*calendar, I’m a week early. Beard *’s a nice touch. Lets everyone know</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>*give you the space. She can’t. A bet *’s a bet. Yes. And that rule outranks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>*Bill *’s going to eat you because Bill *’s a bear. Are you on something? You</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>*’s the theory here? This girl *’s body *’s a lemon? Faulty manufacturing?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>*tendonitis. Not the boy. No, the boy *’s a jerk, she knows that, and yet she</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>*Marrow transplant to treat Brother *’s a match. Confirm and do, I been</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>*for the most likely suspects. Bug *’s a better idea than blindly running</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>*can’t tell up from down and Cameron *’s a girl - all that mechanical stuff.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>*I talk to you? Sure. Look, Cameron *’s a friend. This whole dating thing –</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>*Which leaves bone cancer. Cancer *’s a long shot. Why? Because</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>*realise we’re on to it. Cancer *’s a hard diagnosis for patients even</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3 HMDC 2-gram *’s a (rank 13 of top 20 in HMDC)
Table 5 HMDC VS COCA 3-grams

<table>
<thead>
<tr>
<th>HMDC 3-gram</th>
<th>COCA 3-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td><strong>Word</strong></td>
</tr>
<tr>
<td>1</td>
<td>I DO NT</td>
</tr>
<tr>
<td>2</td>
<td>IT S NOT</td>
</tr>
<tr>
<td>3</td>
<td>YOU DO NT</td>
</tr>
<tr>
<td>4</td>
<td>I M NOT</td>
</tr>
<tr>
<td>5</td>
<td>IT S A</td>
</tr>
<tr>
<td>6</td>
<td>YOU'RE NOT</td>
</tr>
<tr>
<td>7</td>
<td>I DID NT</td>
</tr>
<tr>
<td>8</td>
<td>I CA NT</td>
</tr>
<tr>
<td>9</td>
<td>DO NT WANT</td>
</tr>
<tr>
<td>10</td>
<td>YOU CA NT</td>
</tr>
<tr>
<td>11</td>
<td>DO NT KNOW</td>
</tr>
<tr>
<td>12</td>
<td>DO NT HAVE</td>
</tr>
<tr>
<td>13</td>
<td>NT WANT TO</td>
</tr>
<tr>
<td>14</td>
<td>YOU WANT TO</td>
</tr>
<tr>
<td>15</td>
<td>THERE S NO</td>
</tr>
<tr>
<td>16</td>
<td>YOU DID NT</td>
</tr>
<tr>
<td>17</td>
<td>HE S NOT</td>
</tr>
<tr>
<td>18</td>
<td>WE DO NT</td>
</tr>
<tr>
<td>19</td>
<td>WHAT ARE YOU</td>
</tr>
<tr>
<td>20</td>
<td>DO NT THINK</td>
</tr>
<tr>
<td>21</td>
<td>I M GONNA</td>
</tr>
<tr>
<td>22</td>
<td>WE NEED TO</td>
</tr>
<tr>
<td>23</td>
<td>RE GOING TO</td>
</tr>
<tr>
<td>24</td>
<td>YOU RE A</td>
</tr>
<tr>
<td>25</td>
<td>IF IT S</td>
</tr>
<tr>
<td>26</td>
<td>THERE S A</td>
</tr>
<tr>
<td>27</td>
<td>IF YOU RE</td>
</tr>
<tr>
<td>28</td>
<td>WHAT DO YOU</td>
</tr>
<tr>
<td>29</td>
<td>S NOT A</td>
</tr>
<tr>
<td>30</td>
<td>YOU RE GONNA</td>
</tr>
<tr>
<td>31</td>
<td>IT S THE</td>
</tr>
<tr>
<td>32</td>
<td>THAT S NOT</td>
</tr>
<tr>
<td>33</td>
<td>WHY ARE YOU</td>
</tr>
<tr>
<td>34</td>
<td>END OF THE</td>
</tr>
<tr>
<td>Ngram</td>
<td>Rank wrt HMDC</td>
</tr>
<tr>
<td>---------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>I do n’t</td>
<td>1</td>
</tr>
<tr>
<td>It’s not</td>
<td>2</td>
</tr>
<tr>
<td>You do n’t</td>
<td>3</td>
</tr>
<tr>
<td>I’m not</td>
<td>4</td>
</tr>
<tr>
<td>It’s a</td>
<td>5</td>
</tr>
<tr>
<td>You’re not</td>
<td>6</td>
</tr>
<tr>
<td>I did n’t</td>
<td>7</td>
</tr>
<tr>
<td>I ca n’t</td>
<td>8</td>
</tr>
<tr>
<td>Do n’t want</td>
<td>9</td>
</tr>
<tr>
<td>You ca n’t</td>
<td>10</td>
</tr>
<tr>
<td>Do n’t know</td>
<td>11</td>
</tr>
<tr>
<td>Do n’t have</td>
<td>12</td>
</tr>
<tr>
<td>n’t want to</td>
<td>13</td>
</tr>
<tr>
<td>You want to</td>
<td>14</td>
</tr>
<tr>
<td>There’s no</td>
<td>15</td>
</tr>
<tr>
<td>You did n’t</td>
<td>16</td>
</tr>
<tr>
<td>He’s not</td>
<td>17</td>
</tr>
<tr>
<td>We do n’t</td>
<td>18</td>
</tr>
<tr>
<td>What are you</td>
<td>19</td>
</tr>
<tr>
<td>Do n’t think</td>
<td>20</td>
</tr>
<tr>
<td>Total rank difference</td>
<td></td>
</tr>
</tbody>
</table>

From Table 6, both lists of top twenty 3-gram of HMDC and COCA share six clusters in common, indicating that HMDC partially resembles COCA in a more spoken aspect, that is eighteen out of the top twenty 3-grams in HMDC involve the use of contractions, except You want to and What are you. These contractions, typically found in spoken English (Quaglio 2008), have shown a positive correlation with the use of negativity (highlighted in orange) along with interpersonal pronouns or auxiliary Do in HMDC, accounting for seventeen of the twenty 3-grams and is more than just 6 instances found in the top twenty 3-grams in COCA. This implies that HMDC appears to contain far more disagreement than COCA and COCA Spoken do.

Five instances of 3-grams are not found including It’s not, I’m not, It’s a, You’re not and There’s no, however, alternative forms of such instances are located, namely It isn’t (rank 532), I am not (rank 638), It is a (rank 64), You are not (rank 2,663), You aren’t (rank 6,009) and There is no (rank 52), however, they do not rank high enough to affect the position of the top twenty 3-grams.
The results obtained thus far have point towards a possibly closer resemblance of HDMC with the spoken English of COCA rather than COCA as a whole as indicated by the total rank differences (33,189 excluding “not found” and 15,110). Further investigation will be carried out in the next section.

HMDC VS COCA Spoken Ngrams

Table 7 HMDC VS COCA Spoken 1-gram

<table>
<thead>
<tr>
<th>HMDC 1-gram</th>
<th>COCA Spoken 1-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td><strong>Word</strong></td>
</tr>
<tr>
<td>1</td>
<td>YOU</td>
</tr>
<tr>
<td>2</td>
<td>THEY</td>
</tr>
<tr>
<td>3</td>
<td>THE</td>
</tr>
<tr>
<td>4</td>
<td>IS</td>
</tr>
<tr>
<td>5</td>
<td>TO</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
</tr>
<tr>
<td>7</td>
<td>IT</td>
</tr>
<tr>
<td>8</td>
<td>NOT</td>
</tr>
<tr>
<td>9</td>
<td>THAT</td>
</tr>
<tr>
<td>10</td>
<td>AND</td>
</tr>
<tr>
<td>11</td>
<td>DO</td>
</tr>
<tr>
<td>12</td>
<td>OF</td>
</tr>
<tr>
<td>13</td>
<td>BE</td>
</tr>
<tr>
<td>14</td>
<td>WHO</td>
</tr>
<tr>
<td>15</td>
<td>IN</td>
</tr>
<tr>
<td>16</td>
<td>THAT</td>
</tr>
<tr>
<td>17</td>
<td>NOT</td>
</tr>
<tr>
<td>18</td>
<td>WHAT</td>
</tr>
<tr>
<td>19</td>
<td>YOUR</td>
</tr>
<tr>
<td>20</td>
<td>HAVE</td>
</tr>
<tr>
<td>21</td>
<td>SHE</td>
</tr>
<tr>
<td>22</td>
<td>FOR</td>
</tr>
<tr>
<td>23</td>
<td>ME</td>
</tr>
<tr>
<td>24</td>
<td>WAS</td>
</tr>
<tr>
<td>25</td>
<td>OF</td>
</tr>
<tr>
<td>26</td>
<td>BE</td>
</tr>
<tr>
<td>27</td>
<td>NO</td>
</tr>
<tr>
<td>28</td>
<td>HERE</td>
</tr>
<tr>
<td>29</td>
<td>ON</td>
</tr>
<tr>
<td>30</td>
<td>THIS</td>
</tr>
<tr>
<td>31</td>
<td>MY</td>
</tr>
<tr>
<td>32</td>
<td>WITH</td>
</tr>
<tr>
<td>33</td>
<td>ABOUT</td>
</tr>
</tbody>
</table>
Comparison of HMDC and COCA Spoken shows the corpora sharing fifteen out of the top twenty most frequently used words, which are two more than the comparison between HMDC and COCA 1-gram at thirteen. The total rank differences of 180 and 168 shown in Table 8 recorded lower values than the reading of 267 and 261 in Table 2, providing early evidence that COCA Spoken might be more suitable than COCA to be used as a reference corpus for the television drama *House M.D.*
Table 9 HMDC VS COCA Spoken 2-gram

<table>
<thead>
<tr>
<th>HMDC 2-gram</th>
<th>COCA Spoken 2-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word</strong></td>
<td><strong>Freq</strong></td>
</tr>
<tr>
<td>IT</td>
<td>6,720</td>
</tr>
<tr>
<td>I</td>
<td>5,857</td>
</tr>
<tr>
<td>YOU RE</td>
<td>5,106</td>
</tr>
<tr>
<td>DO NT</td>
<td>4,792</td>
</tr>
<tr>
<td>HE S</td>
<td>3,006</td>
</tr>
<tr>
<td>S HOY</td>
<td>2,576</td>
</tr>
<tr>
<td>THAT S</td>
<td>2,371</td>
</tr>
<tr>
<td>IN THE</td>
<td>2,089</td>
</tr>
<tr>
<td>SHE S</td>
<td>2,068</td>
</tr>
<tr>
<td>CAN NT</td>
<td>2,012</td>
</tr>
<tr>
<td>I DO</td>
<td>1,966</td>
</tr>
<tr>
<td>DID NT</td>
<td>1,722</td>
</tr>
<tr>
<td>YOU DO</td>
<td>1,558</td>
</tr>
<tr>
<td>DOES NT</td>
<td>1,504</td>
</tr>
<tr>
<td>ARE YOU</td>
<td>1,475</td>
</tr>
<tr>
<td>TO BE</td>
<td>1,465</td>
</tr>
<tr>
<td>THERE S</td>
<td>1,434</td>
</tr>
<tr>
<td>IF YOU</td>
<td>1,412</td>
</tr>
<tr>
<td>WE RE</td>
<td>1,404</td>
</tr>
<tr>
<td>GOING TO</td>
<td>1,341</td>
</tr>
<tr>
<td>I WAS</td>
<td>1,289</td>
</tr>
<tr>
<td>WANT TO</td>
<td>1,244</td>
</tr>
<tr>
<td>ON THE</td>
<td>1,232</td>
</tr>
<tr>
<td>OF THE</td>
<td>1,213</td>
</tr>
<tr>
<td>THIS IS</td>
<td>1,187</td>
</tr>
<tr>
<td>THAT S</td>
<td>1,174</td>
</tr>
<tr>
<td>IT WAS</td>
<td>1,161</td>
</tr>
<tr>
<td>WANT TO</td>
<td>1,137</td>
</tr>
<tr>
<td>HAVE TO</td>
<td>1,122</td>
</tr>
<tr>
<td>DO YOU</td>
<td>1,122</td>
</tr>
<tr>
<td>I LL</td>
<td>1,086</td>
</tr>
<tr>
<td>YOU HAVE</td>
<td>1,021</td>
</tr>
<tr>
<td>TO THE</td>
<td>1,020</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frequency</th>
<th>alphabetical</th>
<th>statistica</th>
<th>Frenemies</th>
<th>notes</th>
<th>Frequency</th>
<th>alphabetical</th>
<th>statistica</th>
<th>Frenemies</th>
<th>notes</th>
</tr>
</thead>
</table>

243
## Table 10 HMDC VS COCA Spoken 2-gram ranked with rank difference

<table>
<thead>
<tr>
<th>Ngram</th>
<th>Rank wrt HMDC</th>
<th>Rank wrt COCA Spoken</th>
<th>Rank Difference</th>
<th>Ngram</th>
<th>Rank wrt HMDC</th>
<th>Rank wrt COCA Spoken</th>
<th>Rank Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>It’s</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>of the</td>
<td>1</td>
<td>25</td>
<td>24</td>
</tr>
<tr>
<td>I’m</td>
<td>2</td>
<td>12</td>
<td>10</td>
<td>in the</td>
<td>2</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>You’ve</td>
<td>3</td>
<td>20</td>
<td>17</td>
<td>It’s</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Don’t</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>I think</td>
<td>4</td>
<td>53</td>
<td>49</td>
</tr>
<tr>
<td>He’s</td>
<td>5</td>
<td>21</td>
<td>16</td>
<td>You know</td>
<td>5</td>
<td>35</td>
<td>30</td>
</tr>
<tr>
<td>’s not</td>
<td>6</td>
<td>69</td>
<td>63</td>
<td>Going to</td>
<td>6</td>
<td>21</td>
<td>15</td>
</tr>
<tr>
<td>That’s</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>That’s</td>
<td>7</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>In the</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td>Do n’t</td>
<td>8</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>She’s</td>
<td>9</td>
<td>138</td>
<td>129</td>
<td>To be</td>
<td>9</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>Ca n’t</td>
<td>10</td>
<td>86</td>
<td>76</td>
<td>And I</td>
<td>10</td>
<td>43</td>
<td>33</td>
</tr>
<tr>
<td>I do</td>
<td>11</td>
<td>18</td>
<td>7</td>
<td>On the</td>
<td>11</td>
<td>24</td>
<td>13</td>
</tr>
<tr>
<td>Did n’t</td>
<td>12</td>
<td>47</td>
<td>35</td>
<td>I’m</td>
<td>12</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>’s a</td>
<td>13</td>
<td>22</td>
<td>9</td>
<td>To the</td>
<td>13</td>
<td>33</td>
<td>20</td>
</tr>
<tr>
<td>You do</td>
<td>14</td>
<td>97</td>
<td>83</td>
<td>And the</td>
<td>14</td>
<td>86</td>
<td>72</td>
</tr>
<tr>
<td>Does n’t</td>
<td>15</td>
<td>122</td>
<td>107</td>
<td>This is</td>
<td>15</td>
<td>26</td>
<td>11</td>
</tr>
<tr>
<td>Are you</td>
<td>16</td>
<td>151</td>
<td>135</td>
<td>It was</td>
<td>16</td>
<td>28</td>
<td>12</td>
</tr>
<tr>
<td>To be</td>
<td>17</td>
<td>9</td>
<td>8</td>
<td>We’re</td>
<td>17</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>There’s</td>
<td>18</td>
<td>found</td>
<td>#VALUE!</td>
<td>I do</td>
<td>18</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>If you</td>
<td>19</td>
<td>39</td>
<td>20</td>
<td>For the</td>
<td>19</td>
<td>54</td>
<td>35</td>
</tr>
<tr>
<td>We’re</td>
<td>20</td>
<td>17</td>
<td>3</td>
<td>You’re</td>
<td>20</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>Total rank difference</td>
<td>730</td>
<td>Total rank difference</td>
<td>371</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Similar to the comparison between HMDC and COCA 1-gram in Table 4, the 2-gram There’s is treated as 1-gram in COCA and COCA Spoken\(^2\), thus producing no results in Table 10, but with the exception of this case, nineteen of the top twenty 2-grams appeared in HMDC can be found in COCA Spoken as compared to just fifteen in COCA. Nine out of the top twenty 2-grams in HMDC can be found in COCA Spoken as compared to a bare three in COCA, which is a decent increase of 3 folds. The largest rank differences in Table 10 are 135 and 72 with total rank differences of 730 and 371 as oppose to 352,012 and 420 in Table 4 with total rank differences of 968,777 and 1,807. The statistical results presented are now strong evidence reflecting the relatively close relation between HMDC and COCA Spoken than between HMDC and COCA. To ensure the adequacy of proof for this claim, a final comparison of 3-gram between HMDC and COCA Spoken is as follows:

\(^2\) Here is the list of There’s found as 1-gram in COCA Spoken: There’s there (rank 191,885), There’s I (rank 344,659), That’s there’s (rank 375,122), well there’s (rank 462,557), think there’s (rank 558,399), but there’s (rank 569,218), there’s you (rank 601,544), mean there’s (rank 634,390), there’s it (rank 652,912), and there’s (rank 663,066), now there’s (rank 769,311), there’s that (rank 791,738), if there’s (rank 842,453), know there’s (rank 965,620), so there’s (rank 1,008,952), yeah there’s (rank 1,035,314), There’s the (rank 1,200,464).
Table 11 HMDC VS COCA Spoken 3-gram

<table>
<thead>
<tr>
<th>HMDC 3-gram</th>
<th>COCA Spoken 3-gram</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Word</td>
</tr>
<tr>
<td>1</td>
<td>I DO N'T</td>
</tr>
<tr>
<td>2</td>
<td>IT S NOT</td>
</tr>
<tr>
<td>3</td>
<td>YOU DO N'T</td>
</tr>
<tr>
<td>4</td>
<td>I M NOT</td>
</tr>
<tr>
<td>5</td>
<td>IT S A</td>
</tr>
<tr>
<td>6</td>
<td>YOU RE NOT</td>
</tr>
<tr>
<td>7</td>
<td>I DID N'T</td>
</tr>
<tr>
<td>8</td>
<td>I CA N'T</td>
</tr>
<tr>
<td>9</td>
<td>DO N'T WANT</td>
</tr>
<tr>
<td>10</td>
<td>YOU CA N'T</td>
</tr>
<tr>
<td>11</td>
<td>DO N'T KNOW</td>
</tr>
<tr>
<td>12</td>
<td>DO N'T HAVE</td>
</tr>
<tr>
<td>13</td>
<td>N'T WANT TO</td>
</tr>
<tr>
<td>14</td>
<td>YOU WANT TO</td>
</tr>
<tr>
<td>15</td>
<td>THERE S NO</td>
</tr>
<tr>
<td>16</td>
<td>YOU DID N'T</td>
</tr>
<tr>
<td>17</td>
<td>HE S NOT</td>
</tr>
<tr>
<td>18</td>
<td>WE DO N'T</td>
</tr>
<tr>
<td>19</td>
<td>WHAT ARE YOU</td>
</tr>
<tr>
<td>20</td>
<td>DO N'T THINK</td>
</tr>
<tr>
<td>21</td>
<td>I M GONNA</td>
</tr>
<tr>
<td>22</td>
<td>WE NEED TO</td>
</tr>
<tr>
<td>23</td>
<td>RE GOING TO</td>
</tr>
<tr>
<td>24</td>
<td>YOU RE A</td>
</tr>
<tr>
<td>25</td>
<td>IF IT S</td>
</tr>
<tr>
<td>26</td>
<td>THERE S A</td>
</tr>
<tr>
<td>27</td>
<td>IF YOU RE</td>
</tr>
<tr>
<td>28</td>
<td>WHAT DO YOU</td>
</tr>
<tr>
<td>29</td>
<td>S NOT A</td>
</tr>
<tr>
<td>30</td>
<td>YOU RE GONNA</td>
</tr>
<tr>
<td>31</td>
<td>IT S THE</td>
</tr>
<tr>
<td>32</td>
<td>THAT S NOT</td>
</tr>
<tr>
<td>33</td>
<td>WHY ARE YOU</td>
</tr>
<tr>
<td>34</td>
<td>THE WHITE HOUSE</td>
</tr>
</tbody>
</table>
Finally, Table 12 shows a sharing of six clusters in the top twenty 3-grams of HMDC and COCA Spoken, matching the count in Table 6 when comparing HMDC with COCA. However, the most important difference between Table 6 and Table 12 is that the former consists of five unfound clusters on the left side of the table whereas the latter presented with a full list with relatively low total rank difference. As on the right side of the tables, despite yielding a total rank difference of 24,702 in Table 12 which is larger than the Table 6 counterpart at 15,110, various acronyms of the 3-gram *The United States* (*the U.S. / the US / The States / USA / America*) and contraction alternatives of *We are going* (*We’re going / We’re gonna/ We are gonna*) do exist, thus possibly contributing to the largeness of rank difference.

Lastly, Table 12 records 5 instances of negativity (highlighted in orange) found in the top twenty 3-grams in COCA Spoken which is one less than that in COCA. Such decrease in the frequency of negativity in COCA Spoken with respect to COCA is a result of an increase in the positivity in spoken American English. Therefore considering the top twenty 3-grams, COCA Spoken contains 5% more positivity than COCA, while HMDC contains 3.4 times more negativity than COCA Spoken and 2.8 times more than COCA. In a way, *House M.D.* has brutally intervened in viewers’ perception of the norm of American English.
Conclusion

All in all, this study has quantitatively demonstrated the strengths and weaknesses of using 1-to-3-grams rank difference in comparing HMDC with COCA and COCA Spoken, addressed potential methodological issue of contraction alternatives and acronyms affecting ngram ranking and rank difference, discussed how the language used in House M.D. is related to contemporary spoken American English, and finally shown how House M.D. can be identified as a dramedy far more interpersonal, 1<sup>st</sup> and 2<sup>nd</sup> person-addressed and disagreeing than one would encounter in the real world. Judging by the results obtained from this simple analysis, further studies along the line of television drama are worthy researchers’ attention, interest and devotion. The future of television drama in linguistics is and deserved to be bright, and there is certainly a wealth of exotic knowledge waiting to be discovered.

References

AFP. 2009. “‘House’ is world's most popular TV show.” Google. Retrieved April 10, 2013 (http://www.google.com/hostednews/afp/article/ALeqM5gGRhjVWTeAVMws-iEDRJOY3IDH7g).


Building an English-Filipino Tourism Corpus and Lexicon for an
ASEAN Language Translation System

Charmaine S. Ponay
University of Santo Tomas
csponay@ust-ics.mygbiz.com
Charibeth K. Cheng
De La Salle University
chari.cheng@delasalle.ph

Abstract
A parallel corpus is a valuable resource in natural language processing and computational linguistics. It is used in machine translation, lexicography, statistical language analysis, cross language information retrieval, among others. This study aimed to build a parallel corpus of Philippine tourism data and evaluated the corpus using a statistical machine translation system. The corpus is to be used by the Philippine component of the ASEAN-MT project. ASEAN-MT is a network-based ASEAN language public translation service. It aims to automatically translate among ASEAN languages and English, with English as the pivotal language.

The source texts were in English and were gathered from four websites about Philippine tourism. The Filipino translation was done manually. Translations were by sentence. Guidelines to translating were followed like, multiple translations must be avoided and the commonest word order must be followed. Aside from the translations, the corpus is also annotated with the named entities, such as people’s names, group names, company names, currency units, temporal entities, language names, locations, products, and artistic creations.

The parallel corpora has 21,491 sentences with 370,910 English words and 416,290 Filipino words. Each sentence was from varying lengths, from one word to approximately 30 words per sentence. The corpora are categorized into Festivals and Events, Provincial Profile, Tourist Attractions and General Information. These categories were based on the source websites and the information of the text.

A lexicon was created by manual extraction of named-entities from the different Philippine tourism websites. Most of the named-entities for the lexicon were retrieved from lists provided by the different Philippine tourism websites also. The named-entities were manually translated and classified.

The corpora was assessed based on the following factors – (1) its overall BLEU when used on a statistical machine translation system, (2) BLEU score of the sub-corpora per translator, (3) BLEU per category, and (4) BLEU score of the sub-corpora per translator per category.

We found out that the number of function words, named-entities and numbers, ambiguity and alignment are the main factors which affect the quality of the machine translation. Function words add to the number of words in the translated sentence which causes misalignment while named-entities and numbers causes “penalty” to the BLEU score because they were not translated.

Keywords: Statistical machine translation systems, corpus building, tourism, BLEU score and named-entity

1 http://aseanmt.org
1. Introduction
In natural language processing and computational linguistics a parallel corpus is a valuable resource. It is used in machine translation, lexicography, statistical language analysis, cross language information retrieval, among others.

This study aimed to build a parallel corpus of Philippine tourism data and evaluate the corpus using a statistical machine translation system. The corpus is to be used by the Philippine component of the ASEAN-MT\(^4\) project.

The ASEAN-MT is a network-based ASEAN language public translation service. It is an automatic machine translation system that focuses on cross-lingual communication. It aims to automatically translate among ASEAN languages and English, with English as the pivotal language.

ASEAN-MT makes use of a statistical approach to machine translation which relies on the extraction of a bilingual dictionary and translation rules from a large volume of bitext data or training data and the selection of the most probable translation by statistically disambiguating structural ambiguity. The machine learns how to translate words by observing a large amount of examples and assuming constraints from them. The more translation examples are, the more accurate the translation becomes.

The translation of natural language using Statistical machine translation (SMT) treats the process as a machine learning problem. By examining many samples of human-produced translation, SMT algorithms automatically learn how to translate.

2. The Corpus
The process of corpus creation for this project was done manually with the help of existing tools like a language modeller, a web crawler and a machine translation system.

Since most information from Philippine tourism websites were in English, human parallel translation was done to achieve the bilingual corpus. For the quality of bilingual text provided by humans bears direct effects to a statistical machine translation system’s accuracy, because the consistency and regularity of the translation examples facilitate the learning process, thus producing higher quality of translation. (Bookwan, 2013)


Most of the data came from the website, http://www.visitmyphilippines.com. The presentation of information of this website (http://www.visitmyphilippines.com) is the basis of the categorization. From this website, the Philippine tourism data was categorized into the different regions, followed by descriptions of the regional or provincial profile, tourist attractions, festivals and events.

The corpus is a parallel of English and Filipino, composing of 370,910 English words and 416,290 Filipino words with 21491 sentences. Each sentence were from varying lengths, from one word to approximately 30 words per sentence. The corpus was distributed in to the following, Festivals and Events, Provincial Profile, Tourist Attractions and General Information.

\(^4\) http://aseanmt.org
Table 1 shows the distribution of the number of sentences per category. It shows that it has more data on Tourist Attractions as compared with the other categories. This is due to the reason that the travel websites, from which the corpus came from, contain mostly information on the different tourist attractions that a particular place or region would offer.

<table>
<thead>
<tr>
<th>Categories</th>
<th>No. of Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Festivals and Events</td>
<td>2040</td>
</tr>
<tr>
<td>2 Provincial Profile</td>
<td>5048</td>
</tr>
<tr>
<td>3 Tourist Attractions</td>
<td>11383</td>
</tr>
<tr>
<td>4 General Information</td>
<td>3020</td>
</tr>
</tbody>
</table>

Figure 1 Percentage of sentences distribution per category

There are 4 categories, Festivals and Events, Provincial Profile, Tourist Attractions, and General Information. Each category is described below:

Festivals and Events contains information on the names of feasts and events per region or province, dates or tentative dates when an event or feast is held and brief descriptions of the activities which happens during an event or a festival. The format usually starts with a heading of the name of the festival or event, followed by the dates, then the description. For example:

Metro Manila Film Festival (ika-3 linggo ng Disyembre)
Maringal na parada ng mga karosa sa kahabaan ng Roxas Boulevard ng iba-ibang mga movie entries sa taunang Metro Manila Film Festival.

Provincial Profile contains information about the province or the region like its history, when it was founded, how it was founded and some important persons in the history or government of the province or region, the topography, population, climate, and a brief description of the province.

For example:

Lanao del Norte - Lupain ng Kagandahan at Karangyaan
KLIMA
Nasa labas ng typhoon belt ang lalawigan ng Lanao del Norteat mayroong non-seasonal climate. Nagkakaroon ng pag-ulan sa buong taon. Nararanasan ang pinakamabigat na pag-ulan mula Mayo hanggang Disyembre may average na 7 inches kada buwan, habang ang mas kaunting presipitasyon mula Enero hanggang Abril na may average na 3 inches kada buwan. 60inches ang average yearly rainfall at 80 degrees sa Fahrenheit scale ang average na temperatura.
Lenggwahe / Dialekto
Ang panumbasan ng Kristiyano-Muslim ay 60:40 na mas pabor sa mga Kristiyano. Malawak ang winiwika ang Cebuano, Ingles at Filipino ng mga Kristiyano, at Maranao naman para sa Muslim.

Tourist Attractions contains information on the different tourist spots per province or region. It usually have the name of the tourist attraction as a heading followed by a description of the place. For example,

Katedral ng Imus

General Attraction contains information about the Philippines in general, like, the climate, the language used, currencies, suggested clothing, business hours, airport tax, time difference, some rules like where to smoke, customs, brief descriptions of popular tourist spots and food, everyday greetings and question words along with the translations in Filipino, modes of transportation, and activities. For Example:

Ang unang kalahati ng taon, mula Enero hanggang Mayo, ang pinaka magandang panahon upang bisitahin ang bansa. Malamig ang Nobyembre hanggang Pebrero, habang mainit at tuyo ang Marso hanggang Mayo./

The corpus is specific to Philippine tourism domain. The corpus, therefore, has several names of places, persons, organizations, dates etc. So, to find out the representativeness of the corpus, we counted the number of named-entities per category. Table 2 shows that percent of named-entities in proportion to the size of the lexicon per category. It shows that approximately 40% of the category Provincial Profile was composed of named-entities.

Table 2 Percent of named-entities in proportion to the size of the lexicon per category.

<table>
<thead>
<tr>
<th>Category</th>
<th>% of Named-entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Tourist Attraction</td>
<td>23%</td>
</tr>
<tr>
<td>2 Provincial Profile</td>
<td>39%</td>
</tr>
<tr>
<td>3 Festivals and Events</td>
<td>23%</td>
</tr>
<tr>
<td>4 General Information</td>
<td>14%</td>
</tr>
</tbody>
</table>

The corpus was also manually annotated based on the guidelines set by NECTEC. Annotation was done on named-entities, like persons, artistic and sports group, airline companies, credit card companies, organizations, currency units, temporal entities, languages, locations, restaurants, hotels, roads and streets, transportation, food, brands, artistic creations and tourist attractions. The annotations, however, were not validated by anyone.

3. Translation and Annotation Process
3.1 The Translation Process
All of the tourism information came from Philippine tourism websites and were all in English, therefore, manual translations were done. The corpus was translated by 4 human translators. It was a parallel translation, every sentence in English has its equivalent translation in Filipino.
Table 3 is an example translation with annotations, the process of annotation will be discussed in the later section.

Table 3 Sample parallel translation with annotations

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>The ascending highway offers a breathtaking sight of the world famous [ATTR]Taal Volcano[ATTR], a crater within an island within a lake.</td>
<td>Nagbibigay ang papataas na highway ng kamangha-manghang tanawin ng kilala sa buong mundong [ATTR] Bulkan ng [LOC]Taal[LOC][ATTR], isang bunganga sa loob ng isang isla sa loob ng isang lawa.</td>
</tr>
<tr>
<td>Coconut groves, pineapple farms, flower gardens and agricultural crops dot the ridges of [ATTR]Tagaytay[ATTR].</td>
<td>Ang mga puno ng niyog, taniman ng pinya, hardin ng mga bulaklak at agrikultural na pananim ang pumatapibot sa galugod ng [ATTR] Tagaytay [ATTR].</td>
</tr>
<tr>
<td>Varied types of accommodations and numerous restaurants give the visiting tourist the pleasures of meditation, relaxation, leisure and a unique scenic landscape that defines [ATTR]Tagaytay City[ATTR] as a leading destination.</td>
<td>Iba-iba uri ng matitirhan at ilang kainan ang nagbibigay sa bumbisitang mga turista ng pagkakataon para sa meditasyon, pagsasaya, pagpapahinga, pagsasaya at isang naibang scenikong landscape na nagtatangi sa lungsod ng [ATTR] Tagaytay [ATTR] bilang isang punatahan.</td>
</tr>
</tbody>
</table>

The corpus was separated into files according to categories and were randomly divided among four translators. Table 4 shows the distribution of the categories per translator.

Table 4 Distribution of the total number of sentences per category per translator

<table>
<thead>
<tr>
<th>Categories</th>
<th>Translator1</th>
<th>Translator2</th>
<th>Translator3</th>
<th>Translator4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Festivals and Events</td>
<td>192</td>
<td>488</td>
<td>273</td>
<td>81</td>
</tr>
<tr>
<td>2 Provincial Profile</td>
<td>742</td>
<td>2072</td>
<td>2679</td>
<td>1012</td>
</tr>
<tr>
<td>3 Tourist Attractions</td>
<td>4956</td>
<td>2957</td>
<td>1263</td>
<td>1158</td>
</tr>
<tr>
<td>4 General Information</td>
<td>55</td>
<td>1034</td>
<td>202</td>
<td></td>
</tr>
</tbody>
</table>

3.2. The Named-entity Lexicon and the Annotation Process

The named-entity lexicon was created by manual extraction of data from the different Philippine tourism websites. However, most of the named-entities for the lexicon were retrieved from lists provided by the different Philippine tourism websites also. The named-entities were manually translated, classified and manually annotated based on the guidelines stated by NECTEC.

The following named-entities were identified and annotated:

- PERSONS, the title of persons were also annotated.
  [ARTG] Eraserheads [ARTG]
  [ALG] Cebu Pacific [ALG]
  [CCG] Citibank Mastercard [CCG]
ORGANIZATIONS, this includes names of companies, private and government organizations, universities and banks, except airline companies and credit card companies and brand names of products
[ORG] Siliman University[\ORG]
CURRENCY UNITS, if it contains a numeric entity, it was tagged also
[CUR] Peso [\CUR]
TEMPORAL ENTITIES, includes date and clock time expressions, months and days of week, anything related with time, like tomorrow morning, this afternoon
[DTM]ika-[NUM]3[\NUM] siglo[\DTM]
LANGUAGES
[LG] English [\LG]
GEO-POLITICAL LOCATIONS, these are countries, cities, provinces, area names, territories but does not include train stations, airport names, road names, restaurants and companies
[LOC] Manila [\LOC]
RESTAURANTS
[RST] Gerry’s Grill [\RST]
HOTELS, if it contains a location name, it should be tagged too
[HOT] [LOC] Manila [\LOC] Hotel [\HOT]
ROADS AND STREETS
[ROAD] EDSA [\ROAD]
TRANSPORTATION, which were names of bus and train stations, bus tops/stations/depots, train lines, airports
[TRA] Victory Liner [\TRA]
FOODS, if there is a location name inside the food name, the location name was annotated also
[FOOD] [LOC] Vigan [\LOC] Longanissa [\FOOD]
BRANDS, which includes car brands/models, medicines, drugs, except food brands
[BRN] Biogesic [\BRN]
ARTISTIC CREATIONS, any artistic creation, song names, book names, movies, TV programs, other composition names and titles, newspaper and magazine names, and event names; if there was an artistic group or band or a person before the creation, it was annotated also
[OTR][\ARTG] Eraserhead [\ARTG]’s Ang Huling El Bimbo [/OTR]
TOURIST ATTRACTIONS, which includes parks, museums, attractions within locations, events of attractions (world cup)
[ATTR] Pabirik Festival [/ATTR]

Manual annotation was done using the following steps:
The corpus was inspected on a per sentence basis.
A named-entity is most of the time, capitalized, so if it was capitalized, it was identified as to what type of named-entity it is. The type of named-entity depends on prior knowledge of the author whether it is a person, an artistic or sports group, a credit card company, an organization, a currency unit, a language, a location, a restaurant, a hotel, a road or a street, transportation, food, brand, artistic creation or a tourist attraction.
If the author is not sure as to what type of named-entity it is, the way the named-entity was used in the sentence was checked or another resource was consulted like a search engine or another person. This is because some named-entities are ambiguous in the Philippine setting, like, a person’s name can be a name of a street or a road, a location can be a tourist attraction in itself or if it is a brand or an organization.
Some named-entities have annotations within annotations. The titles of persons were also identified and annotated, as well as places within restaurants, tourist attractions or hotels names.
To build the lexicon, the named-entities were extracted from the corpus manually. Named-entities were classified according to the specifications from NECTEC. The count of each entity (but not limited to) is shown in Table 5:

Table 5 Quantity of each named-entity from the Lexicon

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERSONS</td>
<td>311</td>
</tr>
<tr>
<td>ARTISTIC AND SPORTS GROUP</td>
<td>22</td>
</tr>
<tr>
<td>CREDIT CARD COMPANIES</td>
<td>99</td>
</tr>
<tr>
<td>ORGANIZATIONS</td>
<td>208</td>
</tr>
<tr>
<td>CURRENCY UNITS</td>
<td>3</td>
</tr>
<tr>
<td>LANGUAGES</td>
<td>35</td>
</tr>
<tr>
<td>GEO-POLITICAL LOCATIONS</td>
<td>584</td>
</tr>
<tr>
<td>RESTAURANTS</td>
<td>319</td>
</tr>
<tr>
<td>HOTELS</td>
<td>285</td>
</tr>
<tr>
<td>ROADS AND STREETS</td>
<td>146</td>
</tr>
<tr>
<td>TRANSPORTATION</td>
<td>243</td>
</tr>
<tr>
<td>FOODS</td>
<td>169</td>
</tr>
<tr>
<td>BRANDS</td>
<td>103</td>
</tr>
<tr>
<td>ARTISTIC CREATIONS</td>
<td>35</td>
</tr>
<tr>
<td>TOURIST ATTRACTIONS</td>
<td>2611</td>
</tr>
</tbody>
</table>

4. Evaluation
The evaluation of the corpus was done using the BLEU score generated by MOSES, a statistical based machine translation system.

The training process in MOSES takes in the parallel data and uses co-occurrences of words and segments (known as phrases) to infer translation correspondences between the two languages of interest (Koehn, 2014).

MOSES either uses phrase-based machine translation or syntax-based translation.

In phrase-based machine translation, the correspondences are simply between continuous sequences of words, whereas in hierarchical phrase-based machine translation or syntax-based translation, more structure is added to the correspondences (Koehn, 2014).

For example, a hierarchical MT system could learn that the word Filipino word kain X corresponds to the English word eat X, where the Xs are replaced by any Filipino-English word pair.

This project made use of the outputs generated by MOSES. After the training process, MOSES generates information about the corpus like the contents of the corpus and the frequency of the words in the corpus. This project made use of that information to count the number of function words, the named-entities and the numbers or the digits present in the corpus.

MOSES also generates the English-Filipino and the Filipino-English lexicon which this project made use to determine the correctness of the translation on a per lexicon basis.

The tourism corpus was evaluated using the BLEU score computed by MOSES. The corpus was tested according to category where the category of Tourist Attraction got a BLEU score of 76.74. Also, to check the translation quality it was also tested according to who did the manual translation and BLEU scores of 31.59, 31.87, 24.6 and 64.02 were computed based on the translations of Translator1, Translator2, Translator3 and Translator4 respectively.
The corpus was further tested according to translator per category and a BLEU score of 76.57 and 69.69 for categories Provincial Profile and General Information under translator Translator2 and 65.73 for translator Translator4 under category Tourist Attractions.

To analyse the generated BLEU score, the corpus was further analysed according to the following factors:

The amount of named-entities and number per category and per category per translator. It was shown that the number of named-entities and amount of numbers decreases the BLEU score because these things were not translated and causes a penalty from the BLEU score. And based on the percent of number and named-entities the category Provincial Profile got the lowest BLEU score among the categories.

The number of function words. The number of function words affect translation quality because it causes a difference in the number of words between the source and target language, causing problems with alignment.

The percentage of the generated lexicon in proportion to the corpus was also measured to see how much of the translation pairs were learned by the engine and it shows that only category Provincial Profile under Translator 3 got more than 100%.

The percentage of unique words were also measured to check how much of the lexicon multiple translations have or are ambiguous. It shows a less than 60% of unique words in proportion to the size of the lexicon which indicates that 40% of the words have multiple translations.

Based on the analysis, although the corpus generated a relatively high BLEU score when broken down into per category, per translation and per category per translation, the overall BLEU score of the corpus as a whole was low because of percentage of the named-entities, numbers, function words, average difference of words, the proportion of the lexicon versus the corpus and the percentage of uniqueness.

5. Conclusions and future work

The Philippines is a country thriving with natural resources that have been celebrated tourism spots for years. These tourism spots have proven their mettle against other celebrated tourist spots from neighboring countries, and are known throughout the world. With this, it is a known fact that people from all over the world flock to the Philippines to visit the aforementioned tourist spots. To support this fact, it is in the statistics that Philippine Tourism is a major supplier to its economy, contributing 5.9% of the total Gross Domestic Product of the country. This is exactly the reason why the Philippines has to take advantage of its wonderful natural places, resources and sceneries to further boost its tourism and consequently fuel up the economy.

It is important to take note that efficient communication is a major aspect of tourism. With the expected inflow of tourists into the country, a means of communicating with the foreign people needs to be established. With better communication between the natives of the Philippines and the foreigners, tourism in the Philippines will surely continue to flourish as better understanding and relationships with the foreigners will follow. Having this in mind, building an English-Tagalog travel corpus and lexicon for a statistical machine translation system such as the ASEAN-MT, will prove to be beneficial to both local and tourists. This will allow both parties to understand each other whenever communication is necessary. Such a system is expected to help boost Philippine Tourism as foreigners will get to better experience their vacation with the satisfaction of having good communication.

The machine translation system must, however, learn from bilingual data of high quality of translation, with most of the word correctly aligned. As for the Filipino language and the English language, a near perfect alignment is relatively hard to achieve. There are words in
Filipino which when translated to English refers to multiple words and vice versa. This makes translation quality poor. Some words in Filipino are ambiguous, which can also be a problem in translation. Words may also have multiple translations like the Filipino word, *lahat* which can be translated to *everyone, everything, everybody or all*, depending on the usage in the sentence.

After the analysis of the data and based on the discussion of the results and analysis of the corpus as to why the overall BLEU score of 34.42, the following factors affect the translation quality:

**Number of function words**  
Based on frequency per word in the lexicon, there was a maximum of 15 percent of Filipino function words per category. Although function words are important in the sentence construction, it does not contribute much to the translation quality because it causes problems on alignment.

For the example below, we can see that there are a relatively more number of words for the Filipino translation, because of the function words, than the English translation and there is really no one is to one correspondence which produces poor training result.

This cove of pristine white sand is indeed a place for tranquil contentment.  
*Tunay nga na isang lugar ang cove na ito ng malinis na puting buhangin para sa matiwasay na kasiyahan.*

**The percentage of numbers and the named-entities in the Lexicon**  
According to (Papineni, 2002) test sets should have a minimal of dates, numbers and names for these entities are not usually translated. BLEU score’s focus are on the translation of words. Based on the tests and since travel information usually contains names of places, persons, numbers and dates, there was as high as 39% of named-entities and 12% of numbers per category.

**Ambiguity or ambiguous words**  
The percentage of unique words show that at most 40% of the words in the lexicon were not unique and have multiple translations. For example, there were multiple translations for the word *kasiyahan*, it was aligned with words like fun, enjoyment, leisure, and pleasures.

Difference with the number of words upon translation. For example, the word *pinagkatuwaan* was aligned with made a and fun, where its translation must be the two words made fun as a whole.

And so we have the following recommendations:  
Guidelines must be followed for the manual translation like:  
The order of translation must follow the structure of the source language.  
Although structure is really an issue with regards to the Filipino language, it is recommended that the translation must more or less follow the structure of the source language. For example, instead of translating, *She is a beautiful girl* to *Magandang babae siya*. , it must be translated to *Siya ay isang magandang babae*.  
The translation of a word must be consistent.

Also, although lexical ambiguity is always an issue for certain languages, it is recommended that a translation of a word be consistent as much as possible. For example, the word *beach* must be translated to only one translation like *dalampasigan* in the entire document.

To translate or not to translate named-entities, it must be consistent.
As for the named-entities, there should be certain rules on whether to further translate them or not. For example, should Tinago Falls still be translated to Talon ng Tinago or retain it as is.

The human translations should be further evaluated manually.
In this project, the manual translations were assumed to be correct and were not strictly manually evaluated if the translations conformed to the guidelines specified by NECTEC. And so it is recommended that it be further evaluated manually by giving each of the translators the same set of documents to be translated. These documents may be first evaluated manually by an expert then it may be evaluated again using the machine translation engine to see which will get the higher BLEU score.

The named-entity lexicon may be generated automatically from the annotated corpus by searching for the open annotations first ([TAG]) and identifying the tags inside each square bracket whether it is a person, a location, a tourist attraction, a food, an artistic group, organization, currency, or a language. Annotations within annotations or recursive annotations may be identified by searching first for a close annotation ([\TAG]) of an opening annotation before identifying which type of named entity it is.

The content of the corpus and the lexicon, based on the list of named-entities provided by NECTEC, it is recommended to add more information about Filipino food because aside from tourist spots, tourists are also interested about food when going to other countries. We can get more data from local restaurants, resorts and hotels.

Lastly, aside from machine translation, the tourism corpus may be used in other natural language processing applications like

Compiling or revising dictionaries
The meaning or other meanings of a word can be derived from this corpus because some words were used in a different manner because of the tourism domain. Like the word maligo, may also mean to swim and not only to take a bath. Some colloquial words may be included also, the word, mag-swimming may now be part of the Filipino dictionary.

Named-entity recognition in Filipino
This corpus is annotated with 15 different kinds of named-entities. An NER system may be able to learn some rules on how named-entities are recognized in a Filipino documents because there were several instances of them in this corpus.

Information extraction
This tourism corpus contains information about different festivals and events, profiles of provinces and tourist attractions in the Philippines, one can use this data to build a dialogue system that will extract information from the corpus which may also help tourists in making decisions on where to go, places to stay, and where to eat. Although we already have several tourism websites, it is very tedious to read, so a dialogue system where a tourist may “ask” for information may prove to be very useful.

References

Japanese Idioms in Corpus Data and Dictionaries:
Variability and Patterning

Priscilla Ishida
University of Tsukuba, Japan
ishida.priscilla.fn@u.tsukuba.ac.jp

Abstract

Traditional views of Japanese idioms hold that one of their defining properties is lexico-structural invariability, and many dictionaries present idioms as stable, fully fixed expressions. However, evidence from written texts shows that some idioms have variant forms with shared lexis/structure and similar or related meanings. This paper investigates how Japanese idiom variants are represented in phraseological dictionaries, in light of evidence from corpus data. First, past views of the phenomenon of variability are reviewed and some problems of identification and description are discussed. Then partial results of a corpus analysis of a set of 56 Japanese idioms are reported and compared to descriptions of the target idioms in seven monolingual phraseological dictionaries. Data were collected from the Balanced Corpus of Contemporary Written Japanese (BCCWJ) and the Tsukuba Web Corpus (TWC). The analysis focuses on two types of variants: synonymous and antonymous. Results show that, although many dictionaries list some variants, either as independent headwords or as alternate forms within a headword entry, most are inconsistent in their treatment of this phenomenon. Antonymous variants tend to be listed more often than synonymous ones; however, many variants of both types that are attested in corpus data are overlooked. The conclusion is that there is a strong need for the integration of past research, corpus analysis, and lexicographical practice.

Keywords: idioms, corpus data, lexical profiling, synonymous variants, antonymous variants

1. Introduction

Idioms are multi-word expressions characterized by strong bonding between their individual constituents. According to traditional views, this means it is usually not possible to replace individual words with synonyms or antonyms (1), or to delete them (2).

(1) saji/*spūn o nageru (lit. throw the spoon [Jap. word]/spoon [Eng. loanword], ‘give up’), abura o uru/*kau (lit. sell/*buy oil, ‘loaf; waste time’)
(2) kumo no ko o chirasu yō/*kumo o chirasu yō (lit. as if one scatters spider babies/*spiders, ‘run off in all directions’)

However, newspapers, books, internet blogs, and other Japanese texts show that some idioms do allow the replacement (3) or deletion (4) of constituents.

(3) Shikashi, gakkō kankeisha wa ijime ga atta ka dō ka wa wakaranai to kuchi o nigosu. (‘However, the parties concerned at the school prevaricate (lit. muddy their mouths), saying they don’t know whether or not there was any bullying.’) [book] cf. kotoba o nigosu (lit. muddy one’s words, ‘speak evasively; prevaricate’)

5 Miyaji (1982a: 238) defines idioms as “tango no futatsu ijō no renketsutai de atte, sono musubitsuki ga hikaku-teki kataku, zentai de kimatta imi o motsu kotoba (strings of two or more words that have comparatively strong internal bonding and fixed, holistic meanings)”. This paper will adopt Miyaji’s (1982a) definition and deal with expressions that fall within its scope.
This paper takes the view that pairs of expressions like (3) _kotoba o nigosu/kuchi o nigosu_ and (4) _me kara uroko ga ochiru/me kara uroko_ are “idiom variants”, which are defined as two or more idiomatic expressions that share some lexis and meaning and are recognized as relatively stable expressions in the Japanese lexicon (Ishida 1998; cf. Itō 1990; Moon 1998). The aim of this paper is to show what corpus data reveal about idiom variants and to investigate how well the phenomenon of variability is represented in phraseological dictionaries.

2. Past views of fixedness and variability

“Fixedness” or “stability” is widely considered to be a fundamental characteristic of idioms. Past research defines Japanese idioms as strings of words that are habitually used together and resistant to the replacement, addition, and deletion of lexical constituents (Miyaji 1982a; Kunihiro 1985; Muraki 1991). However, there is evidence that some idioms have one or more variant forms that are strongly institutionalized.

Miyaji (1982b) and Morita (1994) point out that some transitive verb phrase idioms have intransitive counterparts, and vice versa (e.g. _yaridama ni ageru_ lit. raise s.o. on the tip of a spear, ‘make an example of s.o.’/ _yaridama ni agaru_ lit. s.o. rises on the tip of a spear, ‘be made an example of’). Miyaji (1985) shows that some verb and adjective phrase idioms have corresponding compounds (e.g. _me ga sameru_ lit. one’s eyes awake, ‘wake up to; realize’/ _mezameru_ ‘wake up to; realize’). Morita (1994) also notes that some idioms have synonymous or antonymous variants created by the replacement of a content word (e.g. _ude ga sagaru/ochiru_ lit. one’s arm goes down/falls, ‘one’s ability or performance declines’; _koshi ga tsuyoi/yowai_ lit. one’s hips are strong/weak, ‘firm; resolute/weak-kneed’).

Ishida (1998) reports the results of an extensive analysis of idiom variants collected from post-war Japanese novels and identifies two principal types of variation. The first is structural variation, which involves the addition or omission of constituents and a change in the internal structure and/or grammatical function of the idiom. The second is lexical variation, which arises from the alternation of a noun, verb, or adjective constituent.

Satō (2007) analyzes the idiom headwords in five Japanese dictionaries and points out inconsistencies in the representation of the lexical structure and/or phonological form of some expressions. He attributes some of these inconsistencies to the existence of idiom variants ( _ikei_ ), which are exemplified but not discussed in detail, and others to differences in individual lexicographers’ views of the degree of standardization of idiom forms.

In short, past research on Japanese idioms has begun to explore the phenomenon of idiom variability. However, until now, scholars have relied on intuition, manually collected data, and/or genre-specific electronic databases. This paper takes the view that corpus data can be used to gain a clearer picture of the phenomenon of variability, including types of variants, non-intuitive forms, and frequency. The following sections will report the results of a corpus analysis of idiom variants and investigate how well evidence from corpus data is reflected in monolingual Japanese phraseological dictionaries.

3. Extracting idiom variants

3.1 Target idioms, tools, and procedures
In order to identify variant forms, this researcher first selected a set of 100 commonly-used idioms dealt with in past studies and then collected potential candidates for variants from the Balanced Corpus of Contemporary Written Japanese (BCCWJ) and an opportunistic web corpus (TWC). In order to find all possible forms of the target idioms, queries were created from the smallest possible number of constituents (Moon 1998; Fellbaum et al. 2006; Philip 2008), using the following search tools.

1) **NINJAL-LWP for BCCWJ (NLB) ver. 1.20**: This is an online lexical profiling tool developed jointly by the National Institute for Japanese Language and Linguistics (NINJAL) and the Lago Institute of Language. It uses the 2011 DVD version of the BCCWJ, which includes 11 text types and approximately 105 million words. The purpose of using this tool was to identify potential variant forms by extracting all phrases containing a key word in the target idiom and filtering the results yielded for phrase structures and collocates.

2) **NINJAL-LWP for TWC (NLT) ver. 1.10**: NLT is an online lexical profiling tool for the Tsukuba Web Corpus (TWC), which is a corpus of 1.1 billion words compiled from Japanese websites in January of 2012 by researchers at the University of Tsukuba. It has the same interface as NLB (see above) and was used to confirm, in a larger corpus, the potential variant candidates previously identified in the BCCWJ.

Multiple queries were necessary for both tools. After obtaining the initial search results, irrelevant data (e.g. literal examples, duplicates) were eliminated manually. The criteria used to identify idiom variants were shared lexis, shared meaning, and a threshold of at least 10 occurrences in the BCCWJ. In borderline cases, native speaker judgments of familiarity/acceptability were also taken into account.  

Search results were compiled and the alternating forms of each idiom were then analyzed and classified. For the purpose of the dictionary analysis, the target idioms were limited to a set of 56 (with a total of 150 individual forms) exemplifying the five types of variants shown below. This paper will compare the results of the corpus and dictionary analyses of synonymous and antonymous variants.

1) synonymous variants 4) idioms/compounds  
2) antonymous variants 5) truncations/expansions  
3) transitive/intransitive variants

### 3.2 Problems of identification and description

3.2.1. The question of “canonical” forms

Idiom dictionaries tend to present idioms as “syntactically, morphologically, and lexically fixed canonical forms” (Fellbaum et al. 2006: 350), and language users tend to believe that an idiom form listed in the dictionary is its “standard” form. One possible view is that idioms have a canonical form and variants are deviations from this form. Judgments regarding which is the canonical form and which is (are) the variant(s) are usually made on the basis of frequency, derivational relationships, or provenance. However, in some cases multiple forms of an idiom are used frequently, as in (5), and language users may judge multiple forms to be “standard” in spite of differences in frequency, as in (5) and (6).

(5) hara ga tatsu (lit. one’s belly rises, ‘get angry’) 766 (BCCWJ)  
**hara o tateru** (lit. raise one’s belly, ‘get angry’) 570 (BCCWJ)  
(6) **shiishifti o utsu** (lit. strike (typewrite) a period, ‘put an end to s.t.’) 164 (BCCWJ)  
**periodo o utsu** (lit. strike (typewrite) a period, ‘put an end to s.t.’) 42 (BCCWJ)

---

6 The BCCWJ yields only 8 instances of kenka o kau (lit. buy a fight, ‘accept a challenge’; see Table 4). However, since native speakers judge this expression to be familiar and acceptable, it is considered here to be a stable variant.
Another view is that pairs such as those shown in (5) and (6) are one idiom with alternate forms (Moon 1998; Ishida 1998). In this view, hara ga tatsu is a variant of hara o tateru (5) and vice versa. This view places emphasis on the mutual relationships between alternate forms and also provides a framework flexible enough to handle clusters of forms with variable lexis and/or structure (Moon 1998; Philip 2008).

This paper takes the second view. Alternating idiom forms are regarded as mutual variants, and no attempt is made to distinguish between canonical and non-canonical forms.

3.2.2 Institutionalized variants vs. creative modifications
Evidence from Japanese corpus data shows that language users sometimes produce ad hoc variants for the purpose of emphasis or stylistic effect, as in (7b) and (8b) below:

(7a) me kara uroko ga ochiru (lit. the scales fall from one’s eyes, ‘be awakened to the truth’)
(7b) me kara uroko ga jū mai hodo ochita (lit. ten or so scales fell from my eyes, ‘be completely awakened to the truth’)
(8a) ne-mimi ni mizu (lit. cold water into one’s sleeping ear, ‘a bolt from the blue’)
(8b) ne-mimi ni ōame (lit. heavy rain into one’s sleeping ear, ‘a huge bolt from the blue’)

This evidence is consistent with findings in studies of English idioms, which present fixedness as a cline ranging from complete frozenness through restricted variation to creative modification for the purpose of wordplay or other stylistic effects (Moon 1998; Philip 2008). This paper distinguishes between creative modifications, which are low-frequency and context-bound ((7b), (8b)), and institutionalized variants, which are used relatively frequently and widely accepted by native speakers ((3)-(6)). The scope of this study is limited to institutionalized variants.

4. Idiom variants in corpus data

4.1 Synonymous variants
Some idioms allow the alternation of a content word, with little or no change to the internal structure and meaning of the idiom. Table 1 shows synonymous variants produced by the alternation of a noun.

<table>
<thead>
<tr>
<th>idioms/variants</th>
<th>meaning (literal/idiomatic)</th>
<th>BCCWJ</th>
<th>TWC</th>
</tr>
</thead>
<tbody>
<tr>
<td>mi ni shimiru</td>
<td>lit. permeate one’s flesh, ‘touch/sting s.o. deeply’</td>
<td>212(2.02)</td>
<td>2422(2.13)</td>
</tr>
<tr>
<td>honemi ni shimiru</td>
<td>lit. permeate one’s bones and flesh</td>
<td>23(0.22)</td>
<td>163(0.14)</td>
</tr>
<tr>
<td>shūshifu o utsu</td>
<td>lit. strike (typewrite) a period [Jap.], ‘put an end to s.t.’</td>
<td>164(1.56)</td>
<td>1397(1.23)</td>
</tr>
<tr>
<td>piriodo o utsu</td>
<td>lit. strike (typewrite) a period [Eng. loanword]</td>
<td>42(0.40)</td>
<td>340(0.30)</td>
</tr>
<tr>
<td>hadome o kakeru</td>
<td>lit. put on the brake/skid [Jap.], ‘slow down/stop s.t.’</td>
<td>96(0.92)</td>
<td>1229(1.08)</td>
</tr>
<tr>
<td>burēki o kakeru</td>
<td>lit. put on the brake(s) [Eng. loanword]</td>
<td>61(0.58)</td>
<td>614(0.54)</td>
</tr>
<tr>
<td>kotoba o nigosu</td>
<td>lit. muddy one’s words, ‘speak evasively/ambiguously’</td>
<td>80(0.76)</td>
<td>249(0.22)</td>
</tr>
<tr>
<td>kuchi o nigosu</td>
<td>lit. muddy one’s mouth</td>
<td>16(0.15)</td>
<td>59(0.05)</td>
</tr>
</tbody>
</table>

*Figures show the raw number of occurrences in each corpus and (in parentheses) frequency per million words.
In some cases, the alternating nouns include a noun of Japanese origin on one hand, and a loanword on the other. For example, shūshīfu ‘period’ and hadome ‘brake; skid’ are both nouns of Japanese origin, and piriodo ‘period’ and burēki ‘brake(s)’ are both borrowings from English. With respect to the pair shūshīfu o utsu and piriodo o utsu, the latter is held to be the original form, with shūshīfu o utsu appearing after shūshīfu was introduced as a Japanese equivalent for the loanword piriodo sometime after 1926 (Miyaji 1982a: 251). In present-day Japanese, shūshīfu o utsu is used more frequently (Table 1), but both expressions are common and familiar. Of the pair kotoba o nigosu and kuchi o nigosu, kotoba o nigosu is usually considered to be the “standard” form, but the BCCWJ also yields 16 occurrences of kuchi o nigosu.

<table>
<thead>
<tr>
<th>Idioms/variants</th>
<th>Meaning (literal/idiomatic)</th>
<th>BCCWJ</th>
<th>TWC</th>
</tr>
</thead>
<tbody>
<tr>
<td>kao-iro o ukagau</td>
<td>lit. (steal a) glance at s.o.’s face-colour, ‘judge s.o.’s feelings/thoughts from his/her expression’</td>
<td>98 (0.94)</td>
<td>778 (0.68)</td>
</tr>
<tr>
<td>kao-iro o miru</td>
<td>lit. look at s.o.’s face-colour</td>
<td>37 (0.35)</td>
<td>235 (0.21)</td>
</tr>
<tr>
<td>kao-iro o yomu</td>
<td>lit. read s.o.’s face-colour</td>
<td>10 (0.10)</td>
<td>17 (0.01)</td>
</tr>
<tr>
<td>kageguchi o tataku</td>
<td>lit. beat the shadow-mouth (=malicious gossip), ‘backbite’</td>
<td>59 (0.56)</td>
<td>257 (0.23)</td>
</tr>
<tr>
<td>kageguchi o iu</td>
<td>lit. speak/say the shadow-mouth</td>
<td>20 (0.19)</td>
<td>197 (0.17)</td>
</tr>
<tr>
<td>kageguchi o kiku</td>
<td>lit. make use of the shadow-mouth</td>
<td>14 (0.13)</td>
<td>16 (0.01)</td>
</tr>
</tbody>
</table>

There are also pairs or sets of variants with alternating verbs (Table 2). In some cases, the verbs themselves have broadly related meaning, as do (kao-iro o) ukagau ‘glance at’, miru ‘look’, and yomu ‘read’. Both the BCCWJ and the TWC add a few instances of other verbs to this paradigm, including saguru ‘search’ (BCCWJ 5) and tashikameru ‘check; see (if)’ (TWC 3). Because of their low frequency, the expressions with these two verbs do not have the status of variants as defined in this paper. However, they do provide further evidence of the productivity of the set kao-iro o ukagau/miru/yomu.

On the other hand, in the set kageguchi o tataku/iu/kiku ‘backbite’, the alternating idiom phrases are synonymous in spite of the fact that the meanings of the constituent verbs are unrelated (tataku ‘hit; beat’/iu ‘speak; say’/kiku ‘make use of’).

In addition to the synonymous pairs and sets shown in Tables 1 and 2, there are also cases of rather free alternation among a variety of noun or verb constituents (Table 3).

<table>
<thead>
<tr>
<th>Idioms/variants</th>
<th>Meaning (literal/idiomatic)</th>
<th>BCCWJ</th>
<th>TWC</th>
</tr>
</thead>
<tbody>
<tr>
<td>yumizu no yō ni (gotoku)*...\V...like hot and cold water, ‘(waste/be wasted)...like water’</td>
<td>43 (0.41)</td>
<td>371 (0.33)</td>
<td></td>
</tr>
<tr>
<td>yumizu no yō ni (gotoku)...tsukau</td>
<td>use...like hot and cold water, ‘waste/squander...like water’</td>
<td>[31 (0.30)]</td>
<td>[187 (0.16)]</td>
</tr>
<tr>
<td>chi no nijimu yō na N</td>
<td>N like blood oozing, ‘strenuous/backbreaking N’</td>
<td>28 (0.27)</td>
<td>255 (0.22)</td>
</tr>
<tr>
<td>chi no nijimu yō na do ‘ryoka</td>
<td>strenuous/backbreaking effort’</td>
<td>[19 (0.18)]</td>
<td>[157 (0.14)]</td>
</tr>
</tbody>
</table>

\* gotoku is the formal, literary equivalent of yō ni ‘like; as (if)’.
\[31\] indicates that out of a total of 43 instances of yumizu no yō ni...\V, 31 included the verb tsukau ‘use’.

265
For instance, although there is a strong tendency for yumizu no yō ni (lit. like hot and cold water) to be followed by the verb tsukau ‘use’, as in (9), this expression also combines with other verbs that mean ‘use’ or ‘flow out’ in a broad sense (10).

(9) Kain o fuyasu tame no kōkokuhi o yumizu no yō-ni tsukatta. (‘They squandered (lit. used like hot and cold water) the ad budget allocated to increasing membership.’)

In the same way, chi no nijimu yō na (lit. like blood oozing) combines most frequently with the noun do’ryoku ‘effort’ (11). However, it is also used to modify a variety of other nouns expressing an activity that demands effort or labour (12).

(11) Sono hi kara, Kazuo no chi no nijimu yō na do’ryoku ga hajimari... (‘From that day on, Kazuo’s backbreaking efforts (lit. efforts like blood oozing) began’.)
(12) renshū ‘practice’, kunren ‘training’, shugyō ‘training; apprenticeship’, kurō ‘hardship; suffering’, kutō ‘tough struggle; hard fight’...

What is evident in these cases is a tension between fixedness and productivity, as well as a kind of extended patterning: a particular verb or noun tends to be preferred, but other verbs or nouns with similar or related meaning can be substituted rather freely. The lexical structure of the expressions in (9-12) can be represented as yumizu no yō ni...V and chi no nijimu yō na...N, respectively. The V or N in each expression is an “empty slot” filled at the time of use by a verb or noun that belongs to the particular semantic class described above.7

4.2 Antonymous variants

Corpus data also yield cases of alternation of an adjective or verb that results in variants with shared lexis and structure and opposite meaning. Table 4 shows variant pairs in which an adjective or verb alternates with its usual antonym (e.g. tsuyoi/yowai, ‘strong/weak’, kasu/kariru, ‘lend/borrow’) and the idioms also have opposite or converse meaning (e.g. ki ga tsuyoi/yowai lit. one’s ki (spirit/mind) is strong/weak, ‘strong-willed; tough’/‘timid; fainthearted’, te o kasu/kariru lit. lend/borrow a hand, ‘give help’/‘get help’).

7 It might be argued that the V and N slots indicated above are not intra-idiomatic constituents but extra-idiomatic collocates. However, because the bonding with tsukau ‘use’ and do’ryoku ‘effort’, respectively, is relatively strong (Table 3), and because the range of alternating verbs and nouns is restricted, this paper takes the view that the V and N slots do in fact constitute parts of their respective idioms.

<table>
<thead>
<tr>
<th>idioms/variants</th>
<th>meaning (literal/idiomatic)</th>
<th>BCCWJ</th>
<th>TWC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ki ga (no) tsuyoi</td>
<td>lit. one’s ki (spirit/mind) is strong, ‘strong-willed’</td>
<td>161 (1.54)</td>
<td>1,326 (1.17)</td>
</tr>
<tr>
<td>ki ga (no) yowai</td>
<td>lit. one’s ki is weak, ‘timid; fainthearted’</td>
<td>154 (1.47)</td>
<td>1,060 (0.93)</td>
</tr>
<tr>
<td>ki ga (no) mijikai</td>
<td>lit. one’s ki (spirit/mind) is short, ‘short-tempered; impatient’</td>
<td>55 (0.52)</td>
<td>299 (0.26)</td>
</tr>
<tr>
<td>ki ga (no) nagai</td>
<td>lit. one’s ki is long, ‘patient’</td>
<td>27 (0.26)</td>
<td>260 (0.23)</td>
</tr>
<tr>
<td>hagire ga (no) yoi</td>
<td>lit. the teeth-sharpness is good, ‘clear and crisp; articulate’</td>
<td>98 (0.94)</td>
<td>584 (0.51)</td>
</tr>
<tr>
<td>hagire ga (no) warui</td>
<td>lit. the teeth-sharpness is bad, ‘inarticulate; evasive’</td>
<td>50 (0.48)</td>
<td>368 (0.32)</td>
</tr>
<tr>
<td>te o kasu</td>
<td>lit. lend a hand, ‘give help’</td>
<td>297 (2.83)</td>
<td>1,432 (1.26)</td>
</tr>
<tr>
<td>te o kariru</td>
<td>lit. borrow a hand, ‘get help’</td>
<td>99 (0.94)</td>
<td>905 (0.80)</td>
</tr>
<tr>
<td>kenka o suru</td>
<td>lit. sell a fight, ‘pick a fight’</td>
<td>74 (0.71)</td>
<td>683 (0.60)</td>
</tr>
</tbody>
</table>
However, the antonymous element is sometimes unpredictable (Table 5), as in the case of *kuchi ga karui* (lit. one’s mouth is light, ‘be unable to keep a secret’) and *kuchi ga katai* (lit. one’s mouth is hard, ‘be able to keep a secret’). In addition, the relationship between antonymous forms may be asymmetrical, as in *futokoro ga atatakai* (lit. one’s breast pocket is warm, ‘have a fat wallet’), on one hand, and *futokoro ga sabishii/samui* (lit. one’s breast pocket is lonely/cold, ‘be low on cash’), on the other.

<table>
<thead>
<tr>
<th>Idioms/variants</th>
<th>Meaning (Literal/Idiomatic)</th>
<th>BCCWJ</th>
<th>TWC</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>kuchi ga karui</em></td>
<td>lit. one’s mouth is light 1) ‘be talkative’ 2) ‘be a blabbermouth; can’t keep a secret’</td>
<td>24 (0.23)</td>
<td>115 (0.10)</td>
</tr>
<tr>
<td><em>kuchi ga omoi</em></td>
<td>1) lit. one’s mouth is heavy, ‘be reticent/slow to speak’</td>
<td>58 (0.55)</td>
<td>131 (0.12)</td>
</tr>
<tr>
<td><em>kuchi ga katai</em></td>
<td>2) lit. one’s mouth is hard, ‘be able to keep a secret’</td>
<td>46 (0.44)</td>
<td>193 (0.17)</td>
</tr>
<tr>
<td><em>futokoro ga atatakai</em></td>
<td>lit. one’s breast pocket is warm, ‘have a fat wallet’</td>
<td>13 (0.12)</td>
<td>34 (0.03)</td>
</tr>
<tr>
<td><em>futokoro ga sabishii</em></td>
<td>lit. one’s breast pocket is lonely, ‘be low on cash’</td>
<td>11 (0.10)</td>
<td>46 (0.04)</td>
</tr>
<tr>
<td><em>futokoro ga samui</em></td>
<td>lit. one’s breast pocket is cold, ‘be low on cash’</td>
<td>10 (0.10)</td>
<td>21 (0.02)</td>
</tr>
</tbody>
</table>

There are also complex networks of antonymous and synonymous variants that overlap partially in terms of lexis, meaning, and syntactic structure. Table 6 shows a cluster that exploits the metaphor of a “curtain” to indicate the beginning and ending of actions or events.

<table>
<thead>
<tr>
<th>Idioms/variants</th>
<th>Meaning (Literal/Idiomatic)</th>
<th>BCCWJ</th>
<th>TWC</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>maku o akeru</em></td>
<td>lit. open the curtain, ‘start s.t.’</td>
<td>75 (0.72)</td>
<td>706 (0.62)</td>
</tr>
<tr>
<td><em>maku o tojiru</em></td>
<td>lit. close the curtain, ‘put an end to s.t.’</td>
<td>101 (0.96)</td>
<td>1,403 (1.23)</td>
</tr>
<tr>
<td><em>maku o orosu</em></td>
<td>lit. lower the curtain, ‘put an end to s.t.’</td>
<td>39 (0.37)</td>
<td>411 (0.36)</td>
</tr>
<tr>
<td><em>maku o hiku</em></td>
<td>lit. draw the curtain, ‘put an end to s.t.’</td>
<td>14 (0.13)</td>
<td>146 (0.13)</td>
</tr>
<tr>
<td><em>maku ga aku</em></td>
<td>lit. the curtain opens, ‘s.t. starts to happen’</td>
<td>14 (0.13)</td>
<td>119 (0.10)</td>
</tr>
<tr>
<td><em>maku ga kitte otosareru</em></td>
<td>lit. the curtain is cut and dropped, ‘s.t. starts with a bang’</td>
<td>19 (0.18)</td>
<td>84 (0.07)</td>
</tr>
<tr>
<td><em>maku ga oriru</em></td>
<td>lit. the curtain comes down, ‘s.t. comes to an end’</td>
<td>10 (0.10)</td>
<td>69 (0.06)</td>
</tr>
</tbody>
</table>

First, there is a set of transitive synonyms/antonyms composed of *maku o akeru* (lit. open the curtain, ‘start s.t.’), on one hand, and *maku o tojiru/orosu/hiku* (lit. close/lower/draw the curtain, ‘put an end to s.t.’), on the other. Secondly, there is a set composed of the intransitive antonyms *maku ga aku/oriru* (lit. the curtain opens/comes down, ‘s.t. starts to happen’/‘s.t. comes to an end’), as well as the expression *maku ga kitte otosareru* (lit. the curtain is cut and dropped, ‘s.t. starts with a bang’), which is an emphatic synonym of *maku ga aku* that contains the passive form of the verb *otosu* ‘drop’.

As noted in Section 1, many idioms do not have synonymous or antonymous variants, even if they contain a word with a commonly-used synonym or antonym in the Japanese lexicon (see (1) *saji/spun o nageru, abura o uru/kau*). Insofar as lexico-structural frozenness is a defining property of idioms (Section 2), it is possible to say that idioms without synonymous
or antonymous variants are prototypical idioms with a relatively high degree of “idiomaticity”. Indeed, idioms with unrealized antonymous variants often have a meaning that is either specialized to one end of a dimension (e.g. kao ga hiroi/*semai lit. one’s face is wide/*narrow, ‘know a lot of people’) or lacks the relational orientation of a constituent verb (e.g. abura o uru/*kau lit. sell/*buy oil, ‘loaf; waste time’). On the other hand, idioms like kotoba/kuchi o nigosu (lit. muddy one’s words/mouth, ‘speak evasively’) and te o kasu/kariru (lit. lend/borrow a hand, ‘give/get help’), along with the others in Tables 1 to 6, exhibit a relatively lower degree of lexico-structural frozenness, as well as regular lexical relations akin to those found among words. These idioms are thus relatively less prototypical—and less “idiomatic”—than those without synonymous/antonymous variants.

5. Idiom variants in dictionaries

5.1 Materials and procedures

In order to investigate the extent to which Japanese phraseological dictionaries provide a reliable account of idiom variants, the results of the corpus analysis were compared to representations of the target idioms and their synonymous/antonymous variants in seven commonly-used dictionaries:

<table>
<thead>
<tr>
<th>Semasiological Dictionaries</th>
<th>Onomasiological Dictionaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>YWKJ (2007): 3,000 entries</td>
<td></td>
</tr>
</tbody>
</table>

The dictionary analysis targeted the 150 variant forms previously identified in the corpus analysis, investigating if and how each form was presented in each dictionary. Four types of listings were found and coded as follows:

- ☎ = listed as a headword/entry;
- ○ = not listed as a headword but shown somewhere in the entry of a related variant form (in a usage note and/or illustrative example, or under a marker such as rui ‘synonym’);
- ● = listed as a headword only (no entry) and cross-referenced to a variant form with an entry;
- × = not listed or shown in any way.

5.2 Dictionary analysis of synonymous variants

Results of the dictionary analysis of synonymous variants are shown in Table 7. Some dictionaries list both members of a pair of synonymous variants as individual headwords (12.5%). For example, both mi ni shimiru (lit. permeate one’s flesh) and honemi ni shimiru (lit. permeate one’s bones and flesh) are listed in three dictionaries (HKKJ/RKJ/NKJ), and shūshifu o utsu (lit. strike (typewrite) a period [Jap.]) and piriodo o utsu (lit. strike (typewrite) a period [Eng. loanword]) appear in two (YWKJ/NKJ).

In more cases (36%), one expression is listed as a headword and a synonymous variant is listed elsewhere in the entry, sometimes under the marker rui(ku) ‘similar (phrases)’. Kageguchi o tataku (lit. beat the shadow-mouth) is listed as a headword with kagekuchi o kiku (lit. make use of the shadow-mouth) specified as its synonymous variant, in four dictionaries (YWKJ/SKKKJ/RKJ/IHKJ). This may reflect lexicographers’ intuitions that the variant with

8 Of these dictionaries, only KIY and NKJ are theory- and text-based. KIY has a relatively small number of entries in comparison to the others; however, it also includes a “Table of Commonly Used Idioms” (Jōyō Kanyōku Ichiran) with a total of about 1280 items. Variants listed in the KIY table are shown in Table 7, for the sake of comparison; however, these are not included in the analysis reported in Sections 5.2-5.3.
tataka ‘hit; beat’ is used more frequently. On the other hand, in some dictionaries kao-iro o miru (lit. look at s.o.’s face-colour) is listed as a headword with kao-iro o ukagai (lit. (steal a) glance at s.o.’s face-colour) as its variant (SKK/J/RKJ), despite the fact that the latter has a considerably higher frequency in both the BCCWJ and TWC.

### Table 7: Synonymous variants in Japanese phraseological dictionaries

<table>
<thead>
<tr>
<th>synonymous variants</th>
<th>frequency</th>
<th>HK</th>
<th>KJ</th>
<th>YW</th>
<th>SK</th>
<th>KJ</th>
<th>IH</th>
<th>KJ</th>
<th>KIY (entry)</th>
<th>KIY (table)</th>
<th>NKJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>mi ni shimiru</td>
<td>212</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>honemi ni shimiru</td>
<td>23</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>shūshifu o utsu</td>
<td>164</td>
<td>×</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>×</td>
<td>○</td>
<td>×</td>
<td>×</td>
<td>○</td>
</tr>
<tr>
<td>piriodo o utsu</td>
<td>42</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>×</td>
<td>○</td>
<td>○</td>
<td>×</td>
<td>○</td>
</tr>
<tr>
<td>hadome o kakeru</td>
<td>96</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>○</td>
<td>○</td>
<td></td>
<td>○</td>
</tr>
<tr>
<td>burēki o kakeru</td>
<td>61</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>×</td>
<td>○</td>
<td>○</td>
<td></td>
<td>○</td>
</tr>
<tr>
<td>kotoba o nigosu</td>
<td>80</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>×</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>kuchi o nigosu</td>
<td>16</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>×</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>kao-iro o ukagai</td>
<td>98</td>
<td>○</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>○</td>
<td>○</td>
<td></td>
<td>×</td>
</tr>
<tr>
<td>kao-iro o miru</td>
<td>37</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>○</td>
<td>×</td>
<td>○</td>
</tr>
<tr>
<td>kao-iro o yomu</td>
<td>10</td>
<td>●</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>○</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>kageguchi o tatatu</td>
<td>59</td>
<td>×</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>×</td>
<td>○</td>
<td>×</td>
<td>○</td>
</tr>
<tr>
<td>kageguchi o ii</td>
<td>20</td>
<td>×</td>
<td>○</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>○</td>
</tr>
<tr>
<td>kageguchi o kiku</td>
<td>14</td>
<td>×</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>×</td>
<td>×</td>
<td>○</td>
<td>×</td>
<td>○</td>
</tr>
<tr>
<td>yumizu no yō ni</td>
<td>43</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>○</td>
<td>○</td>
<td></td>
<td>×</td>
</tr>
<tr>
<td>yumizu no yō ni tsukau</td>
<td>[31]</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
<td>○</td>
</tr>
<tr>
<td>chi no nijimu yō (na)</td>
<td>[187]</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td></td>
<td>○</td>
</tr>
<tr>
<td>chi no nijimu yō na do’ryoku</td>
<td>[19]</td>
<td>×</td>
<td>●</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>○</td>
<td>○</td>
<td></td>
<td>○</td>
</tr>
</tbody>
</table>

Finally, in many cases, only one of the variants is listed, with no indication of an alternate form attested in corpus data (30%). HKKJ lists piriodo o utsu (lit. strike (typewrite) a period [Eng. loanword]), but not shūshifu o utsu (lit. strike (typewrite) a period [Jap.]). The only dictionary that lists kageguchi o ii (lit. speak/say the shadow-mouth) as a synonym for kageguchi o tatatu (lit. beat the shadow-mouth) is YWKJ, even though this variant is used more often than kageguchi o kiku (lit. make use of the shadow-mouth). Hadome o kakeru (lit. put on the brake/skid) and kuchi o nigosu (lit. muddy one’s mouth) also tend to be overlooked. This may be for reasons of space (cf. Miyaji 1982b: 94), or it may be related to lexicographers’ intuitions of currency or stability. However, considering the frequency and familiarity of these variants, it is arguable that they should be included in dictionaries, in order to meet users’ lookup needs and to provide a clear picture of the variability of individual idioms.

The “empty slot” idioms identified in this paper are not treated consistently. Four dictionaries list yumizu no yō ni tsukau (lit. use like hot and cold water) as a headword, with no indication of potential alternation with other verbs. An exception is NKJ, which lists yumizu no yō ni (lit. like hot and cold water) as a headword and presents the combination with tsukau in usage note and examples. On the other hand, all of the dictionaries that list chi no nijimu yō (na) (lit. like blood oozing) as a headword use this form, and only a few (YWKJ/KIY/NKJ) show that this idiom tends to combine with the noun do’ryoku (‘effort’).

### 5.3 Dictionary analysis of antonymous variants

Some antonymous variants are treated systematically in Japanese idiom dictionaries. In 46% of all possible cases, both members of an antonym pair (Table 4) or at least two members of an antonym set (Tables 5 and 6) are listed as independent headwords. This includes the pairs ki ga mijikai/nagai (lit. one’s ki (spirit/mind) is short/long) and te o kasukariru (lit.
lend/borrow a hand), which each appear in four dictionaries (HKKJ/YWKJ/SKKKJ/RKJ and YWKJ/SKKKJ/RKJ/NKJ, respectively), as well as the set kuchi ga karui/omoi/katai (lit. one’s mouth is light/heavy/hard), which appears in five (HKKJ/YWKJ/SKKKJ/RKJ/NKJ). However, while kuchi ga karui (lit. one’s mouth is light) is marked as an antonym of both kuchi ga omoi (lit. one’s mouth is heavy) and kuchi ga katai (lit. one’s mouth is hard) in YWKJ/RKJ (cf. Table 5), most dictionaries present it as an antonym of kuchi ga katai only.

In fewer cases (22%), one variant is listed as a headword, and an antonymous form is listed somewhere in its entry. In five dictionaries (YWKJ/SKKKJ/RKJ/IHKJ/NKJ), hagire ga ii (lit. the teeth-sharpness is good) appears as a headword with hagire ga warui (lit. the teeth-sharpness is bad) listed as its variant, usually under the marker tai/tsuiku ‘antonym’ or han(tai) ‘opposite’. Six dictionaries list both futokoro ga atatakai (lit. one’s breast pocket is warm) and futokoro ga samui (lit. one’s breast pocket is cold) as headwords; however, futokoro ga sabishii (lit. one’s breast pocket is lonely) appears as a headword in only two of these (HKKJ/NKJ). The other four list this variant as a synonym under the entry for futokoro ga samui. This preference may reflect lexicographers’ judgments that samui ‘cold’ is a better antonym for atatakai ‘warm’, even though the variant with sabishii ‘lonely’ is no less stable than the other two (Table 5).

Some variants do not appear at all (6%). RKJ/NKJ have an entry for ki ga tsuyoi (lit. one’s ki (spirit/mind) is strong) but not ki ga yowa (lit. one’s ki is weak), and SKKKJ lists kenka o uru (lit. sell a fight) but not kenka o kau (lit. buy a fight). Treatment of the maku o akeru (lit. open the curtain) cluster shown in Table 6 also lacks consistency. NKJ lists all variants of this cluster, with most as independent entries. However, maku o orosu (lit. lower the curtain) is not shown in IHKJ, maku ga kitte otosareru (lit. the curtain is cut and dropped) is not in RKJ, and both of these dictionaries lack maku o hiku (lit. draw the curtain). HKKJ lists none of the transitive variants (maku o akeru/tojiru/orosu/hiku lit. open/close/lower/draw the curtain).

At the same time, it is true that about two thirds (68%) of the antonymous variants surveyed here do appear in phraseological dictionaries. Attention paid to this type of variant likely reflects lexicographers’ intuitions that antonymous variants have the status of independent lexical items.

6. Conclusions

This paper has shown that corpus tools are useful for the systematic extraction of idiom variants and the identification of patterns of correspondence that may be difficult to discover based on introspection alone. Future research on Japanese idioms should take advantage of corpus data to clarify the degree of variability of individual idioms and to investigate further the phenomenon of variability, including creative modification (Section 3.2.2). The analysis of large-scale corpora is essential because many idioms are low frequency (Moon 1998; Fellbaum et al. 2006).

The dictionary analysis reported here shows that, although some Japanese phraseological dictionaries are more consistent in their treatment of variants than others (NKJ/YWKJ), many variants attested in corpus data are overlooked. In addition, antonymous variants tend to be treated more systematically than synonymous ones. A separate analysis has indicated that transitive/intransitive variants are treated less systematically than antonymous variants, but more systematically than idioms/compounds (Ishida 2013). Future studies should extend these results and suggest a principled basis for the representation of different types of variants in dictionaries.

The results suggest a strong need to integrate the findings of past research, corpus analysis, and lexicographical practice. An important purpose of dictionaries is to provide information on usage, so phraseological dictionaries should move beyond the received view that idioms...
are fixed expressions and present reliable descriptions of their relative variability. Future research should also address the practical issues of representing variants, such as principles for the selection of headwords, cross-referencing, and the use of markers to indicate semantic relationships and frequency.

**References**


**Dictionaries**


Data and tools
A Corpus-based Difficulty Level Evaluation of Example Sentences with the CEFR Levels in CALD4

Yoshiyuki Notohara
Doshisha University, Japan
ynotohar@mail.doshisha.ac.jp

Abstract

This study evaluates the reading ease levels of example sentences with the CEFR levels in Cambridge Advanced Learner’s Dictionary Fourth Edition (CALD4) in terms of two perspectives: the BNC/COCA word family lists (Nation, 2012) and the readability formulae (e.g., the Flesch Reading Ease Score). Example sentences in corpus-based learner’s dictionaries are likely to be difficult for learners to read due to recent authenticity-first approach to dictionary compilation policies (Walter, 2010). For example, CALD4 is a corpus-based dictionary partially based on the Cambridge Learner Corpus (CLC). Due to the data, the CEFR levels can be attached to definitions of words in the dictionary. However, the relationships between the CEFR proficiency levels and objective reading ease levels of attached example sentences seem to be unclear. Then, first of all, this study extracts and makes six CEFR-based example sentence data files (A1, A2, B1, B2, C1, and C2) from the Example Corpus of CALD4 referring to attached CALD4’s CEFR levels. Second, certain example sentences are randomly selected from each CEFR level data file and the levels of words in examples are categorized in terms of the BNC/COCA word family lists through AntWordProfiler 1.4.1.m. (Anthony, 2014). Third, the reading ease levels of example sentences are evaluated by the readability formulae. Finally, this study evaluates the reading ease levels of example sentences with the CEFR levels and discusses whether the word family levels and the readability formulae could be appropriate post-hoc measures in dictionary writing.

Keywords: CEFR levels, difficulty level evaluation, example sentence, the readability formulae, word family

1. Introduction

Recently, corpus-based approaches to dictionary writing have been increasingly common among learner’s dictionaries (Hanks 2012). These dictionaries are ‘reliable’ based on lexicographic evidence for example, citations in ‘real’ written, spoken, and electronic texts. They can show us ESL/EFL learners typical (i.e., frequent and well-dispersed) linguistic features. Specifically, example sentences of word meanings and usages play crucial roles in showing ‘real’ linguistic samples such as contexts, syntactic patterns, collocations, and multiword expressions, etc. However, such authenticity-first approaches to dictionary writing might cause ESL/EFL learners to struggle to read and understand word meanings or usages through ‘difficult’ example sentences. Some lexicographers argue that pedagogically invented examples are suggested to be more suitable for ESL/EFL learners than real ones (Walter 2010: 438).

According to Atkins & Rundell (2008: 458-461), the following three basic criteria of ‘good’ example sentences in a dictionary are suggested: 1) naturalness and typicality, 2) informativeness, and 3) intelligibility. In this paper, the last ‘intelligibility’ issue is tackled by using two objective measures: The BNC/COCA word family lists (Nation 2012) and the readability formulae (e.g., the Flesch Reading Ease Score). Then, this study evaluates the difficulty levels of example sentences with the CEFR levels and discusses whether the word family levels and the readability formulae could be appropriate post-hoc measures in dictionary writing.

273
2. Literature review

2.1. CALD 4
Cambridge Advanced Learner’s Dictionary Fourth Edition (CALD4), the first learner’s dictionary with the CEFR level labels, is a unique corpus-based learner’s dictionary. This gives us example sentences based on not only a large native corpus, Cambridge International Corpus (CIC), but also a large learner corpus, Cambridge Learner Corpus (CLC). Due to English Vocabulary Profile project (http://www.englishprofile.org/index.php/wordlists), learner profiles at different CEFR levels (A1-C2) are becoming increasingly clear in terms of words, meanings, and phrases (CALD4 2013: ix). However, the relationships between the CEFR levels and attached example sentences are not clear. Some examples of a particular easy word meanings or usages with A1 level might be quite difficult; therefore, the difficulty levels of example sentences are needed to be confirmed by other objective perspectives.

2.2. The word family levels
Bauer and Nation (1993) selected and presented six level word family lists from 1,000,000 token Lancaster-Oslo-Bergen (LOB) corpus in terms of frequency, productivity, predictability, and five regularities (i.e., written forms, spoken forms, spellings of the affix, spoken forms of the affix, and functions). Recently, Nation (2012) compiles better-balanced and slightly spoken-based lists, the BNC/COCA word family lists composed of twenty nine word family lists. The first 1000 word family lists (6,857 word types) and the 2nd 1000 word family lists (6,374 word types) were selected from the English corpus (approximately 10 million tokens), which can be divided into two components composed of several US and UK/NZ subcorpora: spoken component (approximately 6 million tokens) and written one (approximately 4 million tokens).

2.3. The readability formulae
So far, several different readability formulae have been proposed and used as measures for evaluating text difficulty. The following six measures are relatively common: 1) the Flesch Reading Ease Score (FRES), 2) the Flesh-Kincaid Grade Level (FKG), 3) the Gunning Fog Index (GFI), 4) the Coleman-Liau Index (CLI), 5) the Simple Measure of Gobbledygook (SMOG), and 6) the Automated Readability Index (ARI). These formulae are actually different although they are commonly based on syllables numbers, word numbers, or sentence numbers. Therefore, it is crucial for us to compare and discuss their similarities and differences. Truly, these formulae are mainly based on linguistic information on vocabulary in given texts although Bailin & Grafstein (2001) argue that the validity of these measures are doubtful in checking actual reading ease from psycholinguistic perspectives.

1) The Flesch Reading Ease Score (FRES)
   \[ \text{FRES} = \frac{206.835 - 1.015 \times \text{total sentences}}{\text{total syllables/total words}} - 84.6 \]
2) The Flesh-Kincaid Grade Level (FKG)
   \[ \text{FKG} = \frac{0.39 \times \text{total words/total sentences}}{\text{total syllables/total words}} + 11.8 \]
3) The Gunning Fog Index (GFI)
   \[ \text{GFI} = \frac{0.4 \times \text{words/sentences} + 100 \times \text{complex words/words}}{\text{total sentences}} - 15.59 \]
4) The Coleman-Liau Index (CLI)
   \[ \text{CLI} = \frac{0.0588 \times \text{L (average number of letters per 100 words)}}{\text{100 words}} - 0.296 \]
   \[ \frac{0.296 \times \text{S (average number of sentences per 100 words)}}{\text{100 words}} - 15.8 \]
5) The Simple Measure of Gobbledygook (SMOG)
   \[ \text{SMOG} = \frac{1.0430 \times \sqrt{\text{number of polysyllables} \times 30 \times \text{number of sentences}}}{\text{number of sentences}} + 3.1291 \]
6) The Automated Readability Index (ARI)
   \[ \text{ARI} = \frac{4.71 \times \text{characters/words} + 0.5 \times \text{words/sentences}}{\text{number of sentences}} - 21.43 \]
2.4. Example corpus studies
Corpus-based studies on example sentences in lexicography area (e.g., Herbst 1996; Ishii 2011; Xu, 2005) have been conducted. Multiple comparisons between example corpora of major dictionaries (e.g., OALD, CIDE, COBUILD, LDOCE etc.) show us several lexicographic features in dictionaries (e.g., definitions, examples, valency information, collocation and phrases, etc.). Although there are a few studies on CALD4 (e.g., Dziemianko 2014), difficulty level evaluation of example sentences with the CEFR levels remains unexplored.

3. Method

3.1. Research Questions
Here, the following three research questions are posed to confirm difficulty levels of example sentences with the CEFR level labels in CALD4:

What frequency distributions of the BNC/COCA word family levels can be seen in example sentences with the CEFR level labels in CALD4?
How difficult example sentences with the CEFR level labels in CALD4 are in terms of readability formulae?
Are the BNC/COCA word family levels and the readability formulae could be appropriate post-hoc measures in the CEFR-based dictionary writing?

3.2. Corpus data
Token and type frequencies of example sentences in the example corpus of CALD4, which categorically extracted and complied from CALD4 referring to CALD4’s CEFR levels, are summarised with AntWordProfiler 1.4.1m (2014) in Table 1.

Table 1 Token and type frequencies in example sentences with the CEFR level labels in the Example Corpus of CALD4

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lines</td>
<td>3,190</td>
<td>4,090</td>
<td>6,732</td>
<td>7,998</td>
<td>3,475</td>
<td>4,783</td>
</tr>
<tr>
<td>Tokens</td>
<td>28,632</td>
<td>38,152</td>
<td>66,814</td>
<td>83,664</td>
<td>36,119</td>
<td>51,344</td>
</tr>
<tr>
<td>Types</td>
<td>3,452</td>
<td>4,465</td>
<td>6,670</td>
<td>8,127</td>
<td>5,913</td>
<td>7,229</td>
</tr>
</tbody>
</table>

Actually, it is quite difficult for us to directly compare their difficulty levels and appropriateness with raw data. One problem is that each CEFR level files include not only sentences but also phrases. Another one is that the total number of lines in CEFR level files are quite different. Therefore, the randomly selected version is needed to be prepared for further delicate analysis. With Microsoft Excel for Mac 2011’s random function, 1,000 example sentences (i.e. sentence level only) are randomly selected and equally extracted from each CEFR level-based files of the Example Corpus of CALD4. Tokens and types of the randomly selected version are calculated with AntWordProfiler 1.4.1m (2014) as shown in Table 2 and Figure 1.

The 1,000 randomly selected version of the Example Corpus of CALD4 shows us that token frequencies in the CEFR level files are gradually increasing as the CEFR levels increase from A1 to C2. Similarly, type frequencies are mildly increasing as well. Such a corpus comparison tells us that vocabulary in each CALD4 example sentence is slightly richer across the CEFR levels; however, the differences in both numbers and kinds do not seem to be so big between levels. With cluster analyses...
(Ward method), token and type frequencies in the 1,000 randomly selected version can be divided into three clusters: Cluster 1 (A1 and A2), Cluster 2 (B1), and Cluster 3 (B2, C1, C2).

Table 2 Token and type frequencies in the 1,000 randomly selected version

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lines</td>
<td>1,000</td>
<td>1,000</td>
<td>1,000</td>
<td>1,000</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Tokens</td>
<td>10,014</td>
<td>10,057</td>
<td>10,704</td>
<td>11,293</td>
<td>11,351</td>
<td>11,552</td>
</tr>
<tr>
<td>Types</td>
<td>1,905</td>
<td>2,222</td>
<td>2,668</td>
<td>2,996</td>
<td>3,040</td>
<td>3,102</td>
</tr>
</tbody>
</table>

Figure 1 Token and type frequencies in the 1,000 randomly selected version

3.3. Data Analysis

In order to confirm the relationships between the CEFR proficiency levels and objective difficulty levels of attached example sentences, the following three steps are conducted as follows: (1) frequency distributions of words in example sentences are described in terms of the BNC/COCA word family lists through AntWordProfiler 1.4.1.m. (Anthony, 2014), (2) the reading ease levels of example sentences are evaluated by the readability formulae, and finally, (3) which word family levels and which readability formulae could be appropriate post-hoc measures in writing the CEFR-based dictionaries are discussed.

Statistically, (1) cluster analysis (Ward method) are used to categorise token and type frequency distributions in terms of two measures: the BNC/COCA word family levels and the readability formulae, (2) correspondence analysis is used to confirm the relationships between six CEFR proficiency levels and two objective measures.

4. Results & Discussions

4.1. The BNC/COCA word family levels

Cluster analyses (Ward method) show us token frequencies of word family levels can be divided into four clusters: Cluster 1 (the 1st 1000 word family), Cluster 2 (the 2nd 1000 word family, Others B2-C2), and Cluster 3 (AWL 570 A1-B1), and Cluster 4 (AWL 570 B2-C2, Others A1-B1). Similarly, type frequencies of word family levels can be divided into three clusters: Cluster 1 (the 1st 1000 word family), Cluster 2 (the 2nd 1000 word family A1, AWL 570, Others A1), and Cluster 3 (the 2nd 1000 A2-C2, Others A2-C2).

As shown in Figure 2. and Figure 3., both token and type frequencies of words in the 1st 1000 word family level are quite high in example sentences of all CEFR levels; in contrast, both token and type frequencies of words in the 2nd 1000 family level, AWL 570 level, and others are relatively low. It could be said that example sentences with the CEFR level labels in
CALD4 are written in relatively easy words regardless of the CEFR levels. Additionally, the numbers of words in AWL 570 level and Others are gradually increasing from A1 to C2.

![Figure 2 Token frequencies of word family levels in the 1,000 randomly selected version](image1)

![Figure 3 Type frequencies of word family levels in the 1,000 randomly selected version](image2)

![Figure 4 Token frequencies of word family levels](image3)

![Figure 5 Type frequencies of word family levels and the CEFR levels](image4)
Correspondence analysis was conducted using IBM SPSS Statistics Ver. 23. Figure 4 shows us the relationships between token frequencies of word family levels and the CEFR levels. Example sentences in B1 level are written in both the 1st 1000 and the 2nd 1000 word family levels. At B2 and C1 levels, words in AWL 570 level can be seen; furthermore, at C2 level, the numbers of words in Others are increasing. In contrast, in Figure 5., type frequencies show us that different words in AWL 570 level are used in C1; furthermore, different words in Others can be seen at C2 level.

4.2. The readability formulae
Cluster analyses (Ward method) show us the readability formulae can be divided into four clusters: Cluster 1 (FRES), Cluster 2 (FKG B2-C2, GFI A1-B1, CLI A1-A2, SMOG B1-C2), Cluster 3 (CLI B2-C2, GFI B2-C2), and Cluster 4 (ARI A1-A2), Cluster 5 (FKG A1-B1, SMOG A1-A2, ARI B1-C2). Except for FRES, it could be said that scores of other formulae increase as the CEFR levels are higher.

4.3. The BNC/COCA word family levels and the readability formulae
To confirm the relationships between four BNC/COCA word family levels and six readability formulae, Pearson's correlations were confirmed. As a result, token and type frequencies of three BNC/COCA word family levels are closely related to six readability formulae; however,
in terms of word family levels, the 2nd 1000 word family levels only are not related. One main reason for that seems to be that the features of vocabulary in the 2nd 1000 word family tend to be relatively less frequent, long and context-dependent word for instance, advertise, blame and compete. As for AWL 570 and Others, they are actually rare in frequencies. However, it is assumed that lexicographers in CALD4 intentionally select academic, technical, and proper words at the higher levels.

Table 3 Correlation between token frequencies of The BNC/COCA word family levels and the readability formulae

<table>
<thead>
<tr>
<th></th>
<th>The 1st 1000</th>
<th>The 2nd 1000</th>
<th>AWL 570</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRES</td>
<td>-.925**</td>
<td>-.365</td>
<td>-.984**</td>
<td>-.915**</td>
</tr>
<tr>
<td>FKG</td>
<td>.934**</td>
<td>.384</td>
<td>.981**</td>
<td>.918**</td>
</tr>
<tr>
<td>GFI</td>
<td>.939**</td>
<td>.348</td>
<td>.984**</td>
<td>.931**</td>
</tr>
<tr>
<td>CLI</td>
<td>.931**</td>
<td>.408</td>
<td>.975**</td>
<td>.922**</td>
</tr>
<tr>
<td>SMOG</td>
<td>.945**</td>
<td>.370</td>
<td>.981**</td>
<td>.933**</td>
</tr>
<tr>
<td>ARI</td>
<td>.942**</td>
<td>.423</td>
<td>.970**</td>
<td>.930**</td>
</tr>
</tbody>
</table>

*p<.01**

Table 4 Correlation between type frequencies of The BNC/COCA word family levels and the readability formulae

<table>
<thead>
<tr>
<th></th>
<th>The 1st 1000</th>
<th>The 2nd 1000</th>
<th>AWL 570</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRES</td>
<td>-.923**</td>
<td>-.664</td>
<td>-.991**</td>
<td>-.919**</td>
</tr>
<tr>
<td>FKG</td>
<td>.926**</td>
<td>.678</td>
<td>.990**</td>
<td>.923**</td>
</tr>
<tr>
<td>GFI</td>
<td>.900*</td>
<td>.649</td>
<td>.995**</td>
<td>.933**</td>
</tr>
<tr>
<td>CLI</td>
<td>.938**</td>
<td>.702</td>
<td>.985**</td>
<td>.928**</td>
</tr>
<tr>
<td>SMOG</td>
<td>.907*</td>
<td>.668</td>
<td>.993**</td>
<td>.936**</td>
</tr>
<tr>
<td>ARI</td>
<td>.936**</td>
<td>.714</td>
<td>.983**</td>
<td>.935**</td>
</tr>
</tbody>
</table>

*p<.01**, *p<.05*

5. Conclusions

In terms of the BNC/COCA word family levels, it could be said that example sentences with the CEFR level labels in CALD4 are written in relatively easy words (the 1st 1000 word family level mainly) regardless of the CEFR levels. Furthermore, as the CEFR level are higher, vocabulary in AWL 570 word family and others are increasing. Considering these results, CALD4’s example sentences are user-friendly and suitable for ESL/EFL dictionary readers different levels.

All six readability formulae here show us that difficulty levels of example sentences with the CEFR level labels in CALD4 are gradually higher and higher as the CEFR levels increase from A1-C2. This could imply that example sentences with the CEFR level labels in CALD4 are well-selected and proficiency-based one. However, in terms of differences of the readability formulae, it could be pointed out here that FRES tend to be related to A1, A2, B1; on the other hand, other formulae tend to be relatively higher CEFR levels (above B2). Among them, ARI could be less effective.

Finally, the correlations between word family levels and the readability formulae show us that the relationships between the BNC/COCA word family levels and the readability formulae are closely related; however, words in the 2nd 1000 word family level are exceptional and less related. Thus, when evaluating difficulty levels of example sentences by two measures (the
BNC/COCA word family levels and the readability formulae), it could be said that vocabulary in the 2nd 1000 word family level and ARI are inappropriate measures.

References
A Corpus-based Study of Noun-adjective Continuum in Japanese and English

Masayuki Hirata, Alex Chengyu Fang
City University of Hong Kong
masayuki.hirata@my.cityu.edu.hk

Abstract
Not all languages share the same number or variety of parts of speech. For instance, the word class keiyō-dooshi, meaning adjectival verb or adjectival noun, is unique to Japanese and it does not have an equivalent word class in English. Such cross-lingual incongruities in word classes should impact on the fundamental design of bilingual dictionaries. This paper is an analysis of the noun-adjective continuum in Japanese. In particular, based on a parallel corpus, it investigates how such a feature influences the equivalent translations in English, focusing on distributional differences of nouns and adjectives in each language. To compare and study the features of nouns and adjectives in the two languages, a corpus-based approach and theories in cognitive linguistics were adopted in this research. The analysis was based on 150,000 pairs of morphologically and syntactically annotated sentences in each language. The results show empirically that the distribution of nouns and adjectives between Japanese and English is highly skewed: Japanese uses approximately 20% more nouns than English, and English has around 10% more adjectives than Japanese. Nouns and adjectives form a continuum in terms of modification functions in Japanese, whereas the same functions are predominantly achieved by adjectives in English. The corpus evidence to be presented in this paper highlights a strong preference for noun modifications in Japanese. In addition, the noun modification schema involving the use of a compound or ‘noun phrase+conjunctive particle no’ can be also observed repeatedly in the use of adjective stems. These observations justify the existence of word classes such as ‘adjectival noun’ in Japanese and illustrate a continuum between noun usage and adjective usage. As will be duly addressed in the paper, while these findings can be logically explained by the notions of conceptual space and semantic map developed in the field of cognitive linguistics, the frequent cross-mapping between noun modifiers in Japanese and corresponding translations via adjectives in English needs to be considered in the design of bilingual dictionaries.

Keywords: corpus-based bilingual lexicography, parallel corpus, cognitive linguistics, conceptual space, semantic map

1. Introduction

1.1 Incongruities in terms of parts of speech in different languages
Mismatches in terms of parts of speech between different languages have been widely reported in several areas of linguistics. For example, Kuno (1973) introduced the term transitive adjective to account for the fact that ‘to like sb/sth’9 (sb/sth-ga sukida) or ‘to want sth’ (sth-ga hoshii) are often translated by adjectives in Japanese together with an object-like noun marked by the nominative case marker ga.

In reality, not all languages share the same number or variety of parts of speech. According to Croft (2001: 63), “many linguists claim that there are languages that lack Adjectives or even lack Noun-Verb distinction”. Such incongruities pose a problem in the fundamental design of bilingual dictionaries because entries in bilingual dictionaries are usually arranged according to the same word classes at the same lexical level.

This may not cause immediate problems for decoding bilingual dictionaries in which the primary function is to consult meanings of unknown words. However, in the case of encoding bilingual dictionaries, where users consult dictionaries to encode a meaning in one language into another form in a different language, it could happen that cross mappings beyond word classes or even beyond a lexical level may produce a better translation.

9 sb: somebody; sth: something
The distinction between semasiological and onomasiological approaches has been suggested in the literature. For example, Siepmann (2005: 8) writes that “[s]emasiological dictionaries tend to consist of an alphabetical word list leading to the user from the word to its meaning, while onomasiological dictionaries allow the user to proceed from a particular concept and find the most appropriate word for it”. Such an entry model which starts from a concept is ideal because the mapping between different languages do not have to be restricted by word classes, hence it allows cross-mappings beyond the parts of speech boundaries. However, the actual compilation of such a bilingual onomasiological dictionary is not easy because it is difficult to comprehensively list all concepts and meanings from scratch that human beings are likely to want to encode. Furthermore, we need to know how such concepts and meanings can be actually mapped between different languages. To implement an onomasiological model, the first step must be to study actual translation processes from one language to another in detail, based on empirical data.

1.2 Goals and organisation of this paper
In the present article, we explore a part of such mappings between Japanese and English, i.e., focusing on noun-adjective continuum in Japanese, and investigate how this feature is reflected in the equivalent English translations based on parallel corpus data. Note that this research does not intend to investigate individual translation patterns between Japanese and English but it focuses more on macro-view of distributional differences of word classes for Japanese and English, attempting to provide quantitative analysis for the cross mappings between nouns and adjectives for the two languages.

Section 2 introduces a word class in Japanese, namely, keiyoo-dooshi, meaning adjectival verb or adjectival noun. This unique word class in Japanese naturally raises a question how such a word class should be translated in English in which the equivalent category does not exist.

Section 3 explicates the theoretical background of this study, by briefly giving an overview of conceptual map and semantic map theories developed in the field of cognitive linguistics. The continuity of nouns and adjectives can be logically explained by these theories, and the arguments justify the existence of the word class keiyoo-dooshi (adjectival noun) in Japanese.

Section 4 gives an account of the research design and the tools used. The specifications of the corpus used in this research are given, and the process of data annotation is explained.

Section 5 begins by overviewing the POS tag information for Japanese and English with their frequency distribution. The results of corpus-based analyses provide the highly skewed distribution of content words between Japanese and English, especially for adjectives. The data also exhibit the influence of Japanese POS distribution in the English side of the parallel corpus. Based on the data, the following subsections present bilingual semantic map for part-of-speech constructions in Japanese and English, for the semantic class of property and for the semantic class of object respectively.

Section 6 confirms that the above analyses justify the continuum of nouns, adjectival nouns, and adjective proper in Japanese, and the arguments lead to a conclusion that object modification and property modification need to be considered together in explaining the huge discrepancy of adjective in terms of frequency between Japanese and English.

Based on the findings and arguments explained in the previous sections, the final conclusion is drawn in Section 7. The implication of the above analyses for bilingual dictionary design is discussed.

2. Keiyoo-dooshi (adjectival verb or adjectival noun)
It is well-known that Japanese has a unique category called *keiyoo-dooshi*. There are two basic categories for adjectives in Japanese, although the labels for each category differ according to disciplines and linguists. One category has the names such as *i-keiyooishi* (i-ending adjective), or *keiyooishi* (adjective). For instance, *yasu-i* (cheap) belongs to this category. In the present article, we use the term *adjective proper* for this category.

The other category is called *na-keiyooishi* (na-ending adjective), *keiyoo-dooshi* (adjectival verb), or *nominal adjective* or *adjectival noun*. For example, *kirei-na* (beautiful/pretty) falls into this category. In this article, we use the term *adjectival noun* for this category.

The existence of such variety of names derives from their complex behaviour and numerous combinations of adjectival endings or particles in modification, and copular patterns in predication. See the following examples in Table 1, which is expanded from Croft (2001: 95). Note that OBJECTS at the top of the left column indicates the features of nouns, and PROPERTIES at the bottom shows the domain of adjectives in Japanese (more detailed explanation for this distinction is given in Section 3.1).

**Table 1 Examples for Japanese adjective constructions (adapted from Croft 2001: 95)**

<table>
<thead>
<tr>
<th>(English)</th>
<th>MODIFICATION</th>
<th>PREDICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OBJECTS:</strong></td>
<td>book</td>
<td>hon-no hyoshi book-no-CONJ&lt;sup&gt;10&lt;/sup&gt; cover</td>
</tr>
<tr>
<td></td>
<td>peace(ful)</td>
<td>heiwa-no saiten peace-no-CONJ festival heiwa-na saiten peaceful festivities</td>
</tr>
<tr>
<td></td>
<td>health(y)</td>
<td>kenkou-no himitsu health-no-CONJ secret kenkou-na hito healthy person</td>
</tr>
<tr>
<td></td>
<td>pretty</td>
<td>kirei-na hito pretty person</td>
</tr>
<tr>
<td></td>
<td>warm</td>
<td>atataka-na hi warm day atataka-i hi warm day</td>
</tr>
<tr>
<td></td>
<td>small</td>
<td>chiisa-na yorokobi small happiness chiisa-i kuruma small car</td>
</tr>
<tr>
<td><strong>PROPERTIES:</strong></td>
<td>cheap</td>
<td>yasu-i hon cheap book</td>
</tr>
</tbody>
</table>

*Hon* (book) is a noun which can be used as a modifier in combination with conjunctive particle *no* as in *hon-no hyoshi* (book cover). It takes copula *da* in predication. The second group of words, *heiwa(na)* (peaceful) and *kenko(na)* (health(y)) are classified as adjectival nouns; however, their status is in between nouns and adjectival nouns in terms of taking both conjunctive particle *no* and adjectival noun ending *na* in modification. For example, we could expect combinations such as ‘heiwa-no saiten’ (festival of peace) or ‘heiwa-na saiten’ (peaceful festivities), and ‘kenkou-no himitsu’ (secret of health) and ‘kenkou-na hito’ (healthy person). Taking the copula *da* in predication is the same.

<sup>10</sup> CONJ: CONJUNCTIVE PARTICLE
<sup>11</sup> TOP: TOPIC MARKER
<sup>12</sup> COP: COPULA
The next word in the middle of the list is the adjectival noun ‘kirei-na’ (pretty), which no longer takes no for modification but only na, and copula da in predication. The next two words exemplify the mixture of characteristics between adjectival noun and adjective proper (i-keiyooshi) in Japanese. Atataka-i (warm) can take both -na and -i ending for modification as in atataka-na hi and atataka-i hi (warm day). In the case of predication, the word can be used in both patterns, taking the copula as in ‘kyo-wa atataka da.’, and ending with basic form of i-keiyooshi (i-ending adjective) as in ‘kyo-wa atataka-i.’ (It is warm today.). Adjective proper (i-keiyooshi) can predicate without any copula and it has its own inflectional forms for tense and negation.

Chiisa-i (small) also has mixed characteristics, however, there is a slight difference with the previous word. For modification, both -na and -i endings are possible in the same way as illustrated for atataka-i (warm), but for predication, chiisa-i (small) cannot take the copula da, unlike atataka-i (warm).

The final word in the list demonstrates a clear departure from the previous patterns. The word yasu-i (cheap) is a pure adjective proper (i-keiyooshi), that uses -i ending both for modification and predication as in ‘yasu-i hon’ (cheap book), and ‘kono hon-wa yasu-i.’ (This book is cheap).

Croft (2001) explains the above phenomena, i.e. the continuum of noun, adjectival noun, and adjective proper in Japanese, through the notions of conceptual space and semantic map. Section 3 explicates the concepts and illustrates how these notions can be utilised for the present research.

3. Theories from cognitive linguistics

The theoretical framework developed by Croft (2001), Radical Construction Grammar, introduced the notion conceptual space and argued that “conceptual space represents a universal structure of conceptual knowledge for communication in human beings” and “language-specific and construction-specific grammatical categories should map onto connected regions of conceptual space” (Croft 2001: 105).

The significance of applying the notion of conceptual space in the present research lies in the fact that the conceptual space for parts of speech is believed to represent a universal structure for human communication. Although there may be different constructions in different languages to express the same meanings, those different constructions should map onto a connected region in conceptual space. Croft’s definitions of Semantic Map Connectivity Hypothesis suggests that “any relevant language-specific and construction-specific category should map onto a connected region in conceptual space” (Croft 2001: 96).

As shown in Figure 1, ‘conceptual space for parts of speech’ can be represented by placing three semantic classes in the left column (OBJECT, PROPERTY, ACTION), and placing three propositional act functions in the top row (REFERENCE, MODIFICATION, PREDICATION). The following subsections account for how this grid should represent the universal structure for parts of speech in human communication.
3.1 Three semantic classes
Table 2 below reproduced from Croft (2001: 87) summarises the differences and the similarities for the three semantic classes with regards to those four semantic properties.

Table 2 Semantic properties of three semantic classes (Croft 2001: 87)

<table>
<thead>
<tr>
<th></th>
<th>RELATIONALITY</th>
<th>STATIVITY</th>
<th>TRANSITORINESS</th>
<th>GRADABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBJECT</td>
<td>non-relational</td>
<td>state</td>
<td>permanent</td>
<td>non-gradable</td>
</tr>
<tr>
<td>PROPERTY</td>
<td>relational</td>
<td>state</td>
<td>permanent</td>
<td>gradable</td>
</tr>
<tr>
<td>ACTION</td>
<td>relational</td>
<td>process</td>
<td>transitory</td>
<td>non-gradable</td>
</tr>
</tbody>
</table>

The semantic classes of OBJECT and ACTION are placed at opposite poles on the left column because the two are maximally different, with the only semantic property shared by the two classes being non-gradability. In terms of the other semantic properties, OBJECT is non-relational, state, and permanent, whereas ACTION is relational, process, and transitory. OBJECT and PROPERTY share two of the semantic properties, namely being stative and permanent. PROPERTY and ACTION share the characteristic of being relational.

The ordering of the three semantic classes in Figure 1 reflects the characteristics of each semantic property summarised in Table 2. That means higher up in the conceptual map (Figure 1) represents the semantic properties such as being less relational, more stative and permanent, and non-gradable. Lower down in the grid portrays the characteristics of being more relational, more dynamic and less permanent, and also being non-gradable. Being in the middle row indicates the mixture of such properties, and the clearest difference from each pole is that it is gradable.

For PROPERTY and ACTION, being relational means that these semantic classes are conceptually dependent on other concepts, that is, in order to define these semantic classes, it implicitly requires other concepts. Croft (2001) gives a couple of examples illustrating that “one cannot conceive of an action such as running without the involvement of a runner, or of a property such as height without something that is tall” (Croft 2001: 87).

3.2 Three propositional act functions
The three propositional act functions illustrated in the top row of Figure 1 are REFERENCE, MODIFICATION, and PREDICATION. The term ‘propositional act function’ originates in Searle’s (1969) distinction of ‘propositional acts’ from other speech acts such as ‘utterance acts’ and ‘illocutionary acts’. Searle lists two examples of propositional acts, viz referring (reference) and predication. According to Croft (1990b) taking in Searle’s account, the propositional act function of reference is “that of identifying some entity, that is, the entity that the speaker intends to talk about”, and the propositional act function of predication is defined as “the act of ascribing a property to a referred-to entity” (Croft 1990b: 250). Although Searle (1969) did not mention the act of modification, Croft supplements the acts of modification as a third propositional act, “which appears to serve the purpose of ‘enriching’ or adding to the description of an entity being referred to”. (Croft 1990b: 250)
Though simple, by having the three semantic classes in the columns, and having these three propositional act functions in the rows, the conceptual space serves to be an underlying universal structure to represent a prototypical part of speech found in many different languages of the world. The greyed-out area in Figure 1, i.e. object reference, property modification, action predication, are considered to represent prototypical noun, prototypical adjective, and prototypical verb respectively.

The conceptual space also connects language specific and construction specific grammatical categories in overlapping regions. This fact is particularly relevant for the present article in explaining the continuum of noun, adjectival noun, and adjective proper in Japanese. Furthermore, it is also highly applicable in investigating how such a feature is reflected in English translations.

Note that conceptual space becomes a semantic map once the space is filled with actual constructions from a particular language. In Section 5.4 and 5.6, this study suggests a bilingual semantic map of Japanese and English parts of speech constructions.

Table 1 in Section 2 demonstrated the continuity of noun, adjectival noun, and adjective proper in Japanese. This analysis reflects the structure of conceptual space in Japanese speakers’ minds. Being located in the centre of conceptual space, semantic class of PROPERTY exhibits continuum with semantic class of OBJECT and semantic class of ACTION. This justifies the existence of word class such as adjectival noun in Japanese, which has the features of not only adjectives but also those of nouns and verbs. The variety of names given for the word class is understandable given the nature of the word class. See the below quotation from Miyagawa (1987).

As the name indicates, and Adjectival Noun (AN) shares properties of adjectives and nouns. […] ANs share some traits with verbs as well (this is reflected in the traditional terminology, *keiyou doosi* ‘adjectival verb’). Among the traditional Japanese grammarians, there is a lively controversy on whether to establish a separate category for the set of words labeled AN. (Miyagawa 1987: 42)

Based on the observations and analyses given in previous subsections, this article seeks to answer the following questions.

Research questions

• Are there any proportional difference between Japanese adjectives and English adjectives? What about Japanese nouns and English nouns?
• Do adjectives in Japanese straightforwardly correspond to adjectives in English?
• How are Japanese and English adjectives used according to the three different functions distinguished by conceptual space? What are the main functions of Japanese and English adjectives?
• How is the continuum observed in Japanese, i.e. the continuum among noun, adjectival noun, and adjective proper, reflected in the parallel corpus data?
• How does the non-existent word category, i.e. *keiyou-dooshi* (adjectival noun) can be translated in English?

4. Methodology

The corpus that played an important role in the present research was the parallel corpus known as Japanese-English News Articles Aligned Data (JENAAD). Actual analyses of the difference between Japanese and English were carried out based on this set of aligned sentences. The other corpus, namely, the Rank frequency list of grammatical word classes (Leech et al. 2001) was used to compare some of the results observed in the JENAAD corpus.
The rationale for using the parallel corpus in the present study is that it allows us to concentrate our attention on the analysis of forms and constructions. In other words, the first task of matching the semantics between Japanese and English is already done computationally. By using the parallel corpus data, we were able to observe cross-mappings beyond word classes for the same meanings between Japanese and English.

4.1 The JENAAAD Corpus

The Japanese-English News Articles Aligned Data (JENAAD) was used to compare Japanese and English. The original data of JENAAD consists of 2,000,000 articles from the Japanese newspaper *Yomiuri Shinbun* and 110,000 articles from its English counterpart *The Daily Yomiuri*. The period covered by these newspaper articles ranges from September 1989 to December 2001.

The articles in *The Daily Yomiuri* are either translations of the original Japanese newspaper, or were written based on Japanese articles. The data has 47,000 article alignments, 150,000 one-to-one sentence alignments, and 38,000 one-to-many alignments between Japanese and English. This study used the 150,000 one-to-one sentence aligned data. According to Utiyama and Isahara (2003), the alignment was done computationally based on the cross-language information retrieval method and dynamic programming. See Utiyama and Isahara (2003) for the details of the alignment.

Syntactic information was vital for the analysis of functions and constructions for the present research. The data in the JENAAD were not annotated, therefore, the first step was to annotate the data for both Japanese and English. The tag sets used for part-of-speech tagging were JUMAN for Japanese and the Penn Treebank for English. Both sets of data were then parsed with syntactic information, using the KNP for Japanese, and the Stanford parser for English. Note that the translation process is unidirectional and limited to Japanese to English in this study.

4.2 Rank frequency list of grammatical word class: written English

The data from the Rank frequency list of grammatical word class: written English (Leech et al. 2001) were used to compare the proportion of content words between the BNC and the English side of JENAAD. Leech et al. (2001) lists the frequencies of the grammatical word classes based on the two-million-word Sampler Corpus of the BNC. For the purpose of the present study, the data from ‘Rank frequency of grammatical word classes: written (compared with spoken) English’ (Leech et al. 2001: 299) was selected to compare the distributional difference of content words between the BNC and the English side of JENAAD. See Section 5.3 for more details.

5 Results

5.1 Overview of distribution of parts of speech for Japanese and English

The data in focus, the Japanese-English News Articles Aligned Data (JENAAD), are fixed in terms of equality in semantics between Japanese and English. Given the fact that meanings are equal between the sentences, the comparison and analysis of parts of speech and their frequency distribution should give us a general idea how words are used similarly or differently to express the same meanings between Japanese and English.

---

13 Ideally, parallel corpus data that have more genres and domains would be desirable; however, this is the largest parallel corpus of Japanese and English available. Others are too small for the purpose of this study.

14 English tagging was done in the process of syntactic parsing by the Stanford parser. The English side of JENAAD was also tagged by CLAWS for a comparison purpose. See Section 5.3.
The frequency distribution of parts of speech for Japanese and English in JENAAD is summarised in Table 3. In total, Japanese has 3,882,390 tokens, and English has 3,608,481 tokens. There are 273,909 discrepancies in the number of tokens between the two languages.

Table 3 Frequency distribution of POS for Japanese and English

<table>
<thead>
<tr>
<th>JAPANESE</th>
<th>FREQ</th>
<th>%</th>
<th>ENGLISH</th>
<th>FREQ</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Meishi</strong> (NOUN)</td>
<td>1687174</td>
<td>43.46%</td>
<td><strong>NOUN</strong></td>
<td>1169939</td>
<td>32.42%</td>
</tr>
<tr>
<td><strong>Doushi</strong> (VERB)</td>
<td>484583</td>
<td>12.48%</td>
<td><strong>VERB</strong></td>
<td>623548</td>
<td>17.28%</td>
</tr>
<tr>
<td><strong>Setsubijii</strong> (SUFFIX)</td>
<td>350326</td>
<td>9.02%</td>
<td><strong>PREPOSITION or CONJUNCTION, subordinating</strong></td>
<td>477555</td>
<td>13.23%</td>
</tr>
<tr>
<td><strong>Jyoshi</strong> (POSTPOSITION)</td>
<td>338196</td>
<td>8.71%</td>
<td><strong>DETERMINER</strong></td>
<td>419284</td>
<td>11.62%</td>
</tr>
<tr>
<td><strong>Jyoshi</strong> (CONJUNCTIVE)</td>
<td>293117</td>
<td>7.55%</td>
<td><strong>ADJECTIVE or NUMERAL, ordinal</strong></td>
<td>320261</td>
<td>8.88%</td>
</tr>
<tr>
<td><strong>Jyoshi</strong> (CASE MARKER)</td>
<td>259905</td>
<td>6.69%</td>
<td><strong>ADVERB</strong></td>
<td>132932</td>
<td>3.68%</td>
</tr>
<tr>
<td><strong>Jyoshi</strong> (TOPIC MARKER)</td>
<td>134196</td>
<td>3.46%</td>
<td><strong>“to” as PREPOSITION or INFINITIVE MARKER</strong></td>
<td>120114</td>
<td>3.33%</td>
</tr>
<tr>
<td><strong>Keiyoushi</strong> (ADJECTIVE)</td>
<td>114946</td>
<td>2.96%</td>
<td><strong>CONJUNCTION, coordinating</strong></td>
<td>102814</td>
<td>2.85%</td>
</tr>
<tr>
<td><strong>Jyoshi</strong> (ADVERBIAL PARTICLE)</td>
<td>74613</td>
<td>1.92%</td>
<td><strong>PRONOUN, personal</strong></td>
<td>92426</td>
<td>2.56%</td>
</tr>
<tr>
<td><strong>Fukushi</strong> (ADVERB)</td>
<td>50909</td>
<td>1.31%</td>
<td><strong>NUMERAL, cardinal</strong></td>
<td>69949</td>
<td>1.94%</td>
</tr>
<tr>
<td><strong>Settouji</strong> (PREFIX)</td>
<td>42965</td>
<td>1.11%</td>
<td><strong>MODAL VERB</strong></td>
<td>(54103)</td>
<td>(1.49)%</td>
</tr>
<tr>
<td><strong>Shijishi</strong> (DEMONSTRATIVE)</td>
<td>35870</td>
<td>0.92%</td>
<td><strong>WH</strong></td>
<td>42483</td>
<td>1.18%</td>
</tr>
<tr>
<td><strong>Hanteiishi</strong> (COPULAR)</td>
<td>(31461)</td>
<td>(0.81)%</td>
<td><strong>GENITIVE MARKER</strong></td>
<td>28794</td>
<td>0.80%</td>
</tr>
<tr>
<td><strong>Jyodoushi</strong> (AUXILIARY VERB)</td>
<td>(23950)</td>
<td>(0.62)%</td>
<td><strong>EXISTENTIAL THERE</strong></td>
<td>4706</td>
<td>0.13%</td>
</tr>
<tr>
<td><strong>Setsuzokushi</strong> (CONJUNCTION)</td>
<td>9822</td>
<td>0.25%</td>
<td><strong>PRE-DETERMINER</strong></td>
<td>2761</td>
<td>0.08%</td>
</tr>
<tr>
<td><strong>Rentaishi</strong> (ADNOMINAL)</td>
<td>4647</td>
<td>0.12%</td>
<td><strong>FOREIGN WORD</strong></td>
<td>522</td>
<td>0.01%</td>
</tr>
<tr>
<td><strong>Jyoshi</strong> (MODAL MARKER)</td>
<td>923</td>
<td>0.02%</td>
<td><strong>LIST ITEM MARKER</strong></td>
<td>251</td>
<td>0.01%</td>
</tr>
<tr>
<td><strong>Kandoushi</strong> (INTERJECTION)</td>
<td>187</td>
<td>0.0048%</td>
<td><strong>INTERJECTION</strong></td>
<td>89</td>
<td>0.0025%</td>
</tr>
<tr>
<td><strong>Miteigigo</strong> (UNDEFINED)</td>
<td>11</td>
<td>0.0003%</td>
<td><strong>SYMBOL</strong></td>
<td>53</td>
<td>0.0015%</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>3882390</td>
<td>100%</td>
<td><strong>TOTAL</strong></td>
<td>3608481</td>
<td>100%</td>
</tr>
</tbody>
</table>

The frequency order of content words for Japanese and English is the same (highlighted with tinted background in Table 3) in JENAAD. Noun is the most frequent part of speech followed by verb, adjective, and adverb. Although the order is the same, the proportions of parts of speech for these content words appear to be different between Japanese and English.

Note that hanteiishi (copular marker) and jyodoushi (auxiliary verb) are listed separately from verbs, but their numbers are included in the category of verb. Similarly, English has the part...
of speech tag MD for auxiliary verb. It is also listed separately but its frequency is included in the verb category.

What is obvious from the information in Table 3 is that many other grammatical classes rank highly between verb and adjective in Japanese. *Settouji* (prefix) ranks as the third most frequent grammatical class for Japanese, and a variety of *jyoshi* (particle) occupies the following five positions. Although the characteristics of *jyoshi* is not homogeneous, the total percentage for *jyoshi* amounts to 28.36% altogether.

With the involvement of these other grammatical classes besides content words, it is unclear whether the proportional differences for content words between Japanese and English is still significant or not. Hence, the next subsection considers the proportional differences of content words between Japanese and English without considering other word classes.

5.2 Proportional differences of content words between Japanese and English without other word classes

Table 4 lists the percentages of content words for Japanese and English in the JENAAD corpus. Those percentages were calculated using only the content words for both languages, excluding any other word classes.

<table>
<thead>
<tr>
<th>JUMAN</th>
<th>PENN TREEBANK</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>FREQ (raw)</td>
</tr>
<tr>
<td>NOUN</td>
<td>1687174</td>
</tr>
<tr>
<td>VERB</td>
<td>484583</td>
</tr>
<tr>
<td>ADJECTIVE</td>
<td>114946</td>
</tr>
<tr>
<td>ADVERB</td>
<td>50909</td>
</tr>
<tr>
<td></td>
<td>2337612</td>
</tr>
</tbody>
</table>

The total frequency of content words in Japanese is 2,337,612, whereas that of English is 2,246,680. There are 90,932 discrepancies in the number of tokens between the two languages.

The most frequent word class, noun, exceeds 72% for Japanese, and 52% for English. Japanese uses approximately 20% more nouns than English. For verbs, the percentage reaches over 27% for English, and slightly over 20% for Japanese. English has about 7% more verbs than Japanese. The proportion of adjectives in Japanese is 4.92%, whereas in English it is 14.25%. The gap between the two languages is around 10%. The proportion of adverbs for both Japanese and English is smaller than other content words, but still the proportion of English adverbs is 2.7 times larger than that of Japanese adverbs.

If we take a simplistic view, assuming that Japanese parts of speech are merely translated into the same parts of speech in English, the proportion of them, especially content words, should not be drastically different between the two languages. Figure 2 represents the proportion of content words between Japanese and English based on the normalised frequency summarised in Table 4. For example, the ratio of NOUN is observed to be roughly 6:4 between Japanese (black bar) and English (grey bar).
For verbs, the ratio reverses. English uses more verbs than Japanese. The proportion is 42.76% for Japanese and 57.24% for English. There are more verbs in English than there are in Japanese; however, the overall category of ‘verb’ needs a careful inspection. For instance, English verbs include auxiliary verbs used in passive constructions, negations, questions, perfectives and so on. Japanese morphemes used for passive, causative, progressive, and others are categorized under the verbal suffix (doushisei-setsubiji). Categorising verbs in terms of their grammatical functions, constructions, and also as a lexical verb is necessary for a more sensible analysis and comparison.

The most striking difference between the two languages emerges from the use of adjectives. The proportion of adjectives is 25.65% for Japanese and 74.35% for English. This finding stresses just how unlikely it is that the translation process is governed by parts of speech. This huge discrepancy already suggests that functions achieved, or meanings expressed by adjectives in English are not necessarily reflected by those of the same parts of speech in Japanese.

There is similarly a sharp contrast between Japanese and English in the use of adverbs. The proportion of adverbs is 26.90% for Japanese and 73.10% for English.

Figure 2 clearly reveals the different use of content words between Japanese and English. If the translation process is governed predominantly by part-of-speech information, the proportion of content words should centre around 50% for each category. These findings at least confirm the necessity of rearranging or reorganising the concept of ‘word’ or ‘parts of speech’ when we compare different languages. The next subsection considers the distribution of English content words in JENAAD compared with the British National Corpus.

5.3 Comparison of content words between BNC (written) and JENAAD (English)
Section 5.2 confirmed the uneven distribution of content words between Japanese and English in JENAAD. There is also a possibility that the distribution of content words in the English side of JENAAD corpus is somehow affected through the process of translation.

To investigate the point, this section utilised the data from Leech et al. (2001). Leech et al. (2001) lists the frequencies of the grammatical word classes based on the two-million-word Sampler Corpus of the BNC. According to them, the rationale for using the Sampler Corpus rather than the whole BNC in counting word classes is that word classes are sufficiently frequent and do not require a large corpus, and the Sampler Corpus was manually checked for the accuracy of tagging.

For the purpose of the present study, the data from ‘Rank frequency of grammatical word classes: written (compared with spoken) English’ (Leech et al. 2001: 299) was selected because we can efficiently exclude the spoken portion of the Sampler Corpus from the BNC.

Note that JENAAD was also tagged with the CLAWS tag sets, and the information was used for this subsection. See Table 5 below.

| Table 5 Comparison of content words between the BNC (written) and JENAAD |
|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
|                | BNC (written English) |                | JENAAD (English) |                |
| POS | FREQ (raw) | FREQ (per million) | % | POS | FREQ (raw) | FREQ (per million) | % |
| NOUN | 283679 | 283677.01 | 48.22% | NOUN | 1135668 | 314717.54 | 51.94% |
| VERB | 172872 | 172870.79 | 29.39% | VERB | 604803 | 167603.66 | 27.66% |
| ADJECTIVE | 78836 | 78835.45 | 13.40% | ADJECTIVE | 319386 | 88508.59 | 14.61% |
| ADVERB | 52876 | 52875.63 | 8.99% | ADVERB | 126477 | 35049.44 | 5.78% |
| | 588263 | 100.00% | | | 2186334 | 100.00% | |

In accordance with the last subsection, only the frequencies for the content words for written English was extracted from ‘Rank frequency of grammatical word classes: written (compared with spoken) English’. The same procedure was repeated for the JENAAD corpus (English) and only frequencies for the content words were extracted.

Table 5 reveals that the proportion of nouns in JENAAD is 3.72% higher than the BNC. For verbs, the proportion is 1.72% higher in the BNC. For adjectives, the proportion is 1.21% higher in JENAAD. For adverbs, the proportion is 3.20% higher in the BNC.

Pearson’s chi-square test was conducted to see whether the difference between the BNC and JENAAD was statistically significant. The statistical software R was used for the test. Note that Yate’s correction for continuity was not used. See the following result.

X-squared = 9576.05, df = 3, p-value < 2.2e-16

As indicated in the above result, the p-value for the test proved to be far lower than a significance level of 0.05, indicating the proportional difference of content words between the BNC and JENAAD is statistically significant. In other words, given that the BNC is supposed to represent standard English, the distribution of content words in JENAAD deviates from the standard.

It can be hypothesised here that the high percentage of nouns in the English side of JENAAD compared to that of the BNC could be explained by the influence of the high percentage of nouns in the Japanese side of the JENAAD corpus. Likewise, the higher proportion of nouns in JENAAD corpus is somehow affected through the process of translation.

15 The data can be accessed from http://ucrel.lancs.ac.uk/bncfreq/flists.html.
adjectives in JENAAD could also be attributed to the high frequency of nouns in the Japanese side of JENAAD. In later sections, it will be observed that many nouns in Japanese are actually used for modification functions, and many of those noun modifiers are translated by adjectives in English.

The proportion of verbs is naturally higher for English compared to Japanese, therefore, it may not be surprising that the BNC has a higher percentage of verbs than JENAAD. For adverbs, the proportion is 3.20% lower in JENAAD compared with the BNC. The smaller proportion of adverbs in the Japanese side of JENAAD could also contribute to the lower percentage for the English side of JENAAD; however, the characteristics of so-called adverbs in English cut across several grammatical classes in Japanese. These facts could influence the generally lower proportion of adverbs in JENAAD than in the BNC.

### 5.4 Bilingual semantic map for Japanese and English (Semantic class of PROPERTY)

The result shown in Section 5.2 portrayed the skewed distribution of content words between Japanese and English. Figure 2 particularly highlighted the huge discrepancy of adjectives in terms of proportion between the two languages. To understand how Japanese adjectives and English adjectives are distributed differently, the parallel corpus data were further analysed according to the three propositional act functions distinguished by the conceptual map.

Table 6 exemplifies the bilingual semantic map of Japanese and English part-of-speech constructions for the semantic class of PROPERTY. ADJECTIVEs with the tinted background in the centre represent the prototypical parts of speech for this semantic class, and the frequencies quoted there are from the POS-tagged information shown in Table 3. Each raw frequency is followed by the normalised frequency (occurrences per million words). The overall purpose of this bilingual semantic map is to indicate the actual use of those adjectives according to their functions. The box at the bottom encircled with the thick black line reveals the use of adjectives in each function, i.e. REF (REFERENCE), MOD (MODIFICATION) and PRED (PREDICATION).

Brief explanations of constructions under each propositional function are given below. Space only permits a brief overview for each constructions.
Table 6 Bilingual semantic map for Japanese and English: Semantic class of PROPERT

<table>
<thead>
<tr>
<th>REFERENCE</th>
<th>MODIFICATION</th>
<th>PREDICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAPANESE</td>
<td>ENGLISH</td>
<td>JAPANESE</td>
</tr>
<tr>
<td>adj. stem: independent noun</td>
<td>adj. stem: independent noun</td>
<td>predicative</td>
</tr>
<tr>
<td>9349 2408.05</td>
<td>(-ness: suffix)</td>
<td>17693 4557.24</td>
</tr>
<tr>
<td>1441 (399.34)</td>
<td>898752.30</td>
<td>(27533) 7416.42</td>
</tr>
<tr>
<td>ADJECTIVE</td>
<td>ADJECTIVE</td>
<td>predicative</td>
</tr>
<tr>
<td>11496 29607.02</td>
<td>attributive:</td>
<td>248265: (278086)</td>
</tr>
<tr>
<td>34657 8926.72</td>
<td>(rentai shushoku)</td>
<td>68800.42</td>
</tr>
<tr>
<td>attributive:</td>
<td>attributive:</td>
<td>predicative: (other POS+ adjectival suffix)</td>
</tr>
<tr>
<td>(26502) 7344.36</td>
<td>(# of adj. as modifiers)</td>
<td>(predicative)</td>
</tr>
<tr>
<td>adj. stem: independent noun</td>
<td>attributive:</td>
<td>object complement</td>
</tr>
<tr>
<td>7706 1984.86</td>
<td>(ity: suffix)</td>
<td>attributive:</td>
</tr>
<tr>
<td>(15127) (4192.07)</td>
<td>attributive:</td>
<td>object complement: passive</td>
</tr>
<tr>
<td>adj. stem: (modifier)</td>
<td>attributive:</td>
<td>verbless clause</td>
</tr>
<tr>
<td>19292</td>
<td>postpositive:</td>
<td>contingent clause</td>
</tr>
<tr>
<td>4969.10</td>
<td>1874</td>
<td>exclamatory clause</td>
</tr>
<tr>
<td>9172</td>
<td>519.33</td>
<td>preposed clause</td>
</tr>
<tr>
<td>(other deadjectival nouns)</td>
<td>(other deadjectival nouns)</td>
<td>(other deadjectival nouns)</td>
</tr>
<tr>
<td>(26502) 7344.36</td>
<td>(modifier)</td>
<td>(other POS+ adjectival suffix)</td>
</tr>
<tr>
<td>1284</td>
<td>282.82</td>
<td>(attributive)</td>
</tr>
<tr>
<td>the+adj.</td>
<td>1098</td>
<td>QP: Quantifier Phrase</td>
</tr>
<tr>
<td>1304 361.37</td>
<td>282.82</td>
<td>UCP: Uncoordinated Phrase</td>
</tr>
<tr>
<td>1304</td>
<td>282.82</td>
<td>verbless clause</td>
</tr>
<tr>
<td>361.37</td>
<td>(other POS+ adjectival suffix)</td>
<td>contingent clause</td>
</tr>
<tr>
<td>13535</td>
<td>(attributive)</td>
<td>exclamatory clause</td>
</tr>
<tr>
<td>(3486.25)</td>
<td>(attributive)</td>
<td>preposed clause</td>
</tr>
<tr>
<td>REF</td>
<td>14.48%</td>
<td>PRED</td>
</tr>
<tr>
<td>17055 4392.91</td>
<td>0.41%</td>
<td>PRED</td>
</tr>
<tr>
<td>1304 361.37</td>
<td>47.89%</td>
<td>17693 4557.24</td>
</tr>
<tr>
<td>MOD</td>
<td>88.79%</td>
<td>(other POS+ adjectival suffix)</td>
</tr>
<tr>
<td>55047 14178.64</td>
<td>14.48%</td>
<td>284373</td>
</tr>
<tr>
<td>7806.84</td>
<td>0.41%</td>
<td>MOD</td>
</tr>
<tr>
<td>88.79%</td>
<td>14.48%</td>
<td>PRED</td>
</tr>
</tbody>
</table>
In the region of PROPERTY REFERENCE, the key concept is ‘nominalisation’, and the focus is on deadjectival nouns. Some of the adjectives in Japanese are used in the form of an adjective stem. They can stand alone the same as a noun, or be incorporated as a part of a compound. The KNP distinguishes such adjective stems with the feature <名詞的形容詞語幹>, which can be translated as ‘noun-like adjective stem’. This ‘noun-like adjective stem’ is a feature assigned only to na-keiyooshi (adjectival nouns). The examples include jiyu (freedom) from jiyu-na (free), kenkou (health) from kenkou-na (healthy), and so forth. When the noun-like adjective stems stand alone, they are in fact the same as deadjectival nouns. Their frequency is 9,349 (2,408.05 pm) in JENAAD.

Adjective stems (in this case both for adjectival noun and adjective proper) can be also combined with noun suffixes to form nouns out of adjectives. Not many suffixes are available to convert adjective proper into nouns in Japanese. Kaiser et al. (2013) lists only a couple of them, i.e. -sa, and -mi. The examples are taka-sa (height) from taka-i (high), omo-mi (heaviness) from omo-i (heavy), and so on. JUMAN also has a POS tag for noun-creating suffixes (meishi-sei setsubiji), which combine with stems of adjectival nouns. For instance, jyuyou-sei (importance) from jyuyou-na (important), jiyu-ka (liberalisation) from jiyu-na (liberal) and so on. These patterns amount to be 7,706 (1984.86 pm) in JENAAD.

Care must be taken in counting deadjectival nouns in English. The situation is obviously different from counting Japanese deadjectival nouns. In Japanese, the parts of speech counted as deadjectival nouns derived from suffixation or adjective stems is based on the number of adjective tags. Whereas, the English deadjectival nouns extracted in Table 6 are based on the number of noun tags. Thus, the number of adjectives is not influenced by the number of deadjectival nouns in English. Therefore, the frequencies for deadjectival nouns in English are indicated in parentheses.

However, there is a category of adjectives in which the adjectives serve as the head of a noun phrase in English. The construction ‘the+adjective’ is such a construction. The examples are phrases such as ‘the rich and the poor’, ‘the unemployed’, ‘the elderly’, and so on. The extraction of such constructions from the JENAAD corpus turned out to be low in frequency. There were only 1,304 (361.37 pm) occurrences even though the pattern included ordinal numbers and other patterns such as ‘the same’ and ‘the other’.

PROPERTY MODIFICATION represents constructions that are considered as prototypical adjectives. The frequencies of adjectives for both Japanese and English quoted from Table 3 are shown as 114,946 (29,607.02 pm) and 320,261 (88,752.30 pm) respectively with the tinted background in Table 6. Those numbers indicate the total number of adjectives tagged as such in JENAAD.

In Japanese, attributive use of adjectives that can be extracted through the KNP’s feature rentai shushoku is 34,657 (8926.72 pm). The attributive use includes the straightforward use of adjectival noun and adjective proper such as tanjyun-na shitsumon (simple question) and tsuyoi kanshin (strong interest).

The adjective stem mentioned in the previous paragraphs can be also used as modifier in compounds or as part of modifier in ‘adjective stem+conjunctive particle no’. Those patterns are counted as attributive use and their frequencies are 19,292 (4,969.10 pm). Similarly, ‘adjective stem+noun suffix’ can be used for the same patterns, i.e. in compounds and in patterns with conjunctive particle no, and their frequencies are 1,098 (282.82 pm). See Section 6 for the examples for these patterns.

Note that under ACTION REFERENCE, the relevant category would be deverbal nouns (e.g. government, running, confirmation and so forth). See Croft (2001) for detailed discussion.

meishi-sei setsubiji: 名詞性接尾辞.
rentai shushoku: <連体修飾>
There are derived adjectives that are created out of other parts of speech with adjectival suffixes. For example, the suffix -tekina combines with nouns to derive adjectives, similar to the English suffix ‘-ic’, and rekishi (history)+tekina creates rekishi-tekina (historic). The frequency of these patterns is put in the parentheses because these numbers do not affect the original adjective tags in JENAAD.

The total number of adjective used attributively in English is 278,086 (77,064.56 pm). Postpositive adjectives (e.g. rules clear to the public), adjectives in uncoordinated phrase (e.g. government and private sector), and adjectives in quantifier phrases (e.g. at least) are considered to be attributive use and their frequencies are 1,874 (519.33 pm), 1,558 (431.76 pm), 2,855 (791.19 pm) respectively.

The patterns such as other than, due to, such as are considered as complex prepositions and they were excluded from the adjective use. Thus their frequency is in the parenthesis in Table 6.

In PROPERTY PREDICATION, the predicative use of adjective in Japanese is 17,693 (4557.24 pm). ‘Other POS+adjectival suffix’ that was introduced in the previous paragraphs can be also used predicatively, however, they do not influence the adjective tag in JENAAD, therefore, these patterns are excluded from the count in Table 6.

Some varieties are observed for the English side of the predicative use of adjectives. The total number of adjective used predicatively is 27,533 (7630.08 pm). Those are straightforward use of adjectives in predicate, for example, “Administrative reform is necessary.”

Quirk et al. (1985) distinguishes several adjective constructions, which are referred to in Table 6. Constructions such as ‘object complement’ (e.g. made this vision possible), ‘verbless clause’ (e.g. Needless to say), ‘contingent clause’ (e.g. When necessary, …), ‘exclamatory clause’ (e.g. How nice!) and ‘preposed adjectives’ (e.g. Even more effective would be …) are considered as a predicative use of adjectives in English. These constructions are relatively infrequent compared to the ‘copula+ADJP’ construction with the exception of the ‘object complement construction’ (2,397 (664.27 pm)). The ‘object complement’ construction is considered to be a predicative use because there is a copula relationship between the object and the complement. ‘Object complement’ constructions can be also expressed in passive forms (e.g. … were considered as superior) and the frequency is 367 (101.70 pm).

5.5 Functional distribution of adjectives in Japanese and English
The usage of adjectives in Japanese and English is drastically different. In Japanese, the distinction between adjectives and nouns is not clear-cut. Furthermore, the division between adjectives and adverbs is equally fuzzy. This continuity between parts of speech makes it hard to grasp the whole range of functions associated with adjectives in Japanese. On the other hand, the difference between English adjectives and nouns or adverbs are generally well-established. It is relatively easy to understand the proportion of functions performed by adjectives in English. See Figure 3.

Figure 3 summarises the functional distribution of Japanese and English adjectives for the three propositional act functions, i.e. REFERENCE, MODIFICATION, and PREDICATION. A category for adverb (the bar with horizontal lines) is added for Japanese because around 20% of inflected adjectives (e.g. tanoshii (happy) -tanoshiku (happily)) are actually adverbs in Japanese. English has the category for ‘others’ which includes adjectives used as a part of complex prepositions (e.g. such as, other than, due to).
Japanese uses around 15% of adjectives for PROPERTY REFERENCE. They consist of deadjectival nouns by means of suffixation and adjective stems acting as independent nouns. The percentage for the same category in English is comparatively small (0.41%). This does not mean that English hardly has deadjectival nouns. There are deadjectival nouns, however, they are tagged as ‘nouns’ in the JENAAD corpus, whereas the tag for Japanese deadjectival nouns and adjective stems are keiyooshi (adjective) and nominal suffixes. The distinction between nouns and adjectives is clearly established in English.

The white bar in the middle in Figure 3 visibly highlights that the main function of adjectives in English is undoubtedly modification (close to 90%). For Japanese, though it is also the main proportion of all adjectives, the number of adjectives used for modification is less than 50%. Recall and refer to Figure 2 that indicated the huge discrepancy between the proportion of Japanese adjectives and English adjectives. The comparatively limited use of Japanese adjectives is distributed across four functions, whereas the greater use of English adjectives is predominantly for the function of modification.

5.6 Semantic class of OBJECT
To answer the question as to why the Japanese and English adjectives exhibit huge discrepancy in terms of frequency, we need to consider OBJECT MODIFICATION together with PROPERTY MODIFICATION. Table 7 exemplifies the bilingual semantic map of Japanese and English part-of-speech constructions for the semantic class of OBJECT. NOUNs with the tinted background represent the prototypical parts of speech for this semantic class, and the frequencies quoted there are from the POS-tagged information shown in Table 3.
Table 7 Bilingual semantic map for Japanese and English: Semantic class of OBJECT

<table>
<thead>
<tr>
<th>OBJECT</th>
<th>REFERENCE</th>
<th>MODIFICATION</th>
<th>PREDICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>JAPANESE</td>
<td>ENGLISH</td>
<td>JAPANESE</td>
</tr>
<tr>
<td>NOUN</td>
<td>1687174</td>
<td>434570.97</td>
<td>multiple NP compound (#of noun modifiers)</td>
</tr>
<tr>
<td>NOUN</td>
<td>1169939</td>
<td>324219.25</td>
<td>multiple NP compound (#of noun modifiers)</td>
</tr>
<tr>
<td></td>
<td>13467</td>
<td>3468.74</td>
<td>NP+kakuyoshi+no+NP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NP+genitive+NP</td>
</tr>
<tr>
<td>REF</td>
<td>772302</td>
<td>198924.37</td>
<td>45.77%</td>
</tr>
<tr>
<td>MOD</td>
<td>640661</td>
<td>177543.13</td>
<td>54.76%</td>
</tr>
<tr>
<td>MOD</td>
<td>864912</td>
<td>222778.24</td>
<td>51.26%</td>
</tr>
<tr>
<td>PRED</td>
<td>509366</td>
<td>141158.01</td>
<td>43.54%</td>
</tr>
<tr>
<td>PRED</td>
<td>49960</td>
<td>12868.36</td>
<td>2.96%</td>
</tr>
<tr>
<td></td>
<td>19912</td>
<td>5518.11</td>
<td>1.70%</td>
</tr>
</tbody>
</table>
The number of nouns computed based on the POS tag was 1,686,174 (434,570.97 pm) for Japanese and 1,169,939 (324,219.25 pm) for English. These figures represent the total number of nouns in the JENAAD corpus. The number of nouns used for the propositional act function of REFERENCE is 772,302 (198,924.37 pm) for Japanese and 640,661 (177,543.13 pm) for English. These figures were calculated by subtracting the use of nouns as MODIFICATION (e.g. computer game) and the use of nouns as PREDICATION (e.g. This is a computer) from the total number of nouns for each language.

In the region of OBJECT MODIFICATION, the frequency of multiple NP compound constructions is significantly high in Japanese, and the actual number of nouns used in these constructions exceeds 630,000 (162,736.61 pm). The ‘NP+no (conjunctive/case particle)+NP’ construction is also highly frequent in Japanese (56,572.88 pm). This construction can be particularly equated with the English constructions such as the use of the genitive case, the possessive form of pronouns, and the prepositional phrase using ‘of’. ‘NP+no’ can be often replaced by adjectives.

‘NP+kakujyoshi (case particle)+no (conjunctive particle)+NP’ includes examples such as nihon shijyou de no kyouou (competition in the Japanese market). The addition of a case particle such as de represents meanings often denoted by a preposition in English.

The most frequent construction under OBJECT MODIFICATION in English is the use of a prepositional phrase. The frequency totals 212,507 (58,890.98 pm) if the frequency for the use of the preposition ‘of’ and other prepositions are combined together.

Multiple NP compounds are also frequent; there are 179,908 (49,856.99 pm) constructions of this kind, and the actual number of nouns used as a modifier amounts to 232,956 (64,557.91 pm). The frequency of the genitive case with ‘-’s is 28,532 (7,906.93 pm), and the frequency of the possessive form of a pronoun is 35,371 (9,802.19 pm).

The target construction under OBJECT PREDICATION is ‘NP+hanteishi’ for Japanese, and ‘copula+(article)+NP’ in English. As shown in Table 7, the frequency of ‘NP+hanteishi’ is 49,960 (12,868.36 pm), and the frequency of ‘copula+NP’ constructions in English is 19,912 (5,518.11 pm). The proportion of nouns used for OBJECT PREDICATION is 2.96% for Japanese and 1.70% for English. The percentages are much lower than those for OBJECT REFERENCE and OBJECT MODIFICATION.

Some distributional differences are observed between Japanese and English. See the bottom lines encircled with the thick black line. In Japanese, around 46% of the total number of nouns are used for the reference function, whereas the percentage for English is around 55% of total nouns, although the concrete number of nouns used for referencing in Japanese exceeds that of English by 92,610 because of their high frequency. The purely referencing function of nouns appears to be higher in proportion for English than Japanese.

Consequently, the proportion of nouns used for the function of MODIFICATION in Japanese is higher: over 51% of the total number of nouns. That is, 864,912 (222,778.24 pm) nouns are used for the propositional act function of modification. English nouns used for modification constitute less than 44%, or 509,366 (141,158.01 pm) occurrences. Apparently, 355,546 more nouns are used for modification in Japanese compared to English.

From the figures summarised in the previous paragraphs, it appears that the major function of nouns for Japanese is MODIFICATION (51.26%), followed by REFERENCE (45.77%). On the other hand, the primary function of English nouns is REFERENCE (54.76%), followed by MODIFICATION (43.54%).
Compared to REFERENCE and MODIFICATION, nouns used for the function of PREDICATION is infrequent for both languages. Less than 3% of noun predicates are observed for both languages, viz. 2.96% and 1.70% for Japanese and English respectively.

6. Discussion

See Figure 4 below that summarises figures and facts about OBJECT MODIFICATION and PROPERTY MODIFICATION in Japanese and in English. Figure 4 enables us to grasp empirically that the function of modification in Japanese is primarily achieved by the use of nouns (OBJECT MODIFICATION) rather than by adjectives (PROPERTY MODIFICATION). On the other hand, the main function of adjectives in English is clearly modification (see the bar graph in the middle).

![Figure 4 Proportions of modification between Japanese and English](image)

Although the proportions of OBJECT MODIFICATION and PROPERTY MODIFICATION are massively different between Japanese and English, the figures balance out when the two types of the modification are considered together. See the bottom bar graph between Japanese modification and English modification in TOTAL. They divide much closer at 50%.

Under the nominal use of adjectives in Japanese, there is a noun modification schema repeatedly used. See the following examples and Figure 5.

- *anzen* (adj. stem) *hoshou rijikai ketsugi*¹ (Security Council resolution)
- *jiyu* (adj. stem) *no hitsuyou* (adj. stem) *sei*² (freedom no-CNJ necessity sei-NOMINAL SUFFIX)
- *kokkou seijyou* (adj. stem)-*ka koushou*³ (diplomatic relation normalising-ka-NOMINAL SUFFIX negotiations)
- *shijyou jiyu* (adj. stem)-*ka no nagare*⁴ (market liberalisation-ka-NOMINAL SUFFIX no-CNJ stream)
- *yowa* (adj. stem)-*sa no ishiki*⁵ (weakness-sa-NOMINAL SUFFIX no-CNJ awareness)

---

¹ 安全保障理事会決議
² 自由の必要性
³ 国交正常化交渉
⁴ 市場自由化の流れ
This noun modification schema, which consists of ‘compound’ and ‘noun/adj. stem+no (conjunctive particle)’, appears recurrently in the JENAAD data. This observation was confirmed in Section 5.6 that this noun modifier constructions are actually massively frequent, thus it could be argued that the pattern is extended to the use of adjective stem. Table 7 lists three constructions in the region of OBJECT MODIFICATION in the bilingual semantic map, i.e. ‘multiple noun compound’, ‘NP+no+NP’, and ‘NP+postposition+no+NP’. All these constructions can be used by ‘adjective stems’.

The continuity of nouns and adjectives supports the ‘Semantic Map Connectivity Hypothesis’ that was introduced in Section 3. The repeated use of noun modifying schemas is possible because OBJECT MODIFICATION and PROPERTY MODIFICATION are located in connected regions of the bilingual semantic map in focus. Furthermore, the huge discrepancy between Japanese adjectives and English adjectives in terms of their frequency can be explained by the strength of OBJECT MODIFICATION in Japanese and the strength of PROPERTY MODIFICATION in English.

The above analyses in the previous sections also provided ample evidence for noun-adjective continuum in Japanese and illustrated how the unique word class keiyoo-dooshi (adjectival noun) is interwoven in the continuum of nouns and adjectives in Japanese. This continuity and the underlying structure of conceptual space and semantic map, and also the strength in each language’s modification function make it possible to cross-map beyond the word classes between Japanese and English for the same meanings and functions.

7. Implications for bilingual dictionary designs and conclusion

The constructions exemplified in the bilingual semantic maps of parts of speech constructions for Japanese and English need to be considered as a model for encoding bilingual dictionaries. The semantic maps presented in this research successfully illustrated the cross mappings of constructions between Japanese and English beyond word classes and also beyond lexical boundaries. The data and analyses in this research quantitatively revealed the fact that is difficult to grasp without empirical data and theories from cognitive linguistics. Further work is obviously necessary to apply these maps and data, for instance, for an onomasiological model of bilingual dictionaries. However, the empirical data provided in this study represent a part of actual translation processes based on the actual usage. Therefore, for the compilation of practically usable bilingual dictionaries, especially for an innovative encoding bilingual dictionary model, this study should provide insightful perspectives.

Acknowledgments
Research described in this paper was supported in part by a grant received from the General Research Fund of the University Grant Council of the Hong Kong Special Administrative Region, China (Project no. CityU 142711) and City University of Hong Kong (Project nos. 6354005, 7002793, 7004091, 700433, 9041694, 9610283, 9610226, and 9610188).

*1. Shaska no seishiki*
References


Lavender Filipino: Computational Models of Twitter Swardspeak

Nathaniel Oco, Raymart Fajutagana, Christine Mae Lim, Judi Diane Miñon,
Julie-Ann Morano, Ryan Christian Tinoco
National University, the Philippines
{nathanoco,raymartfajutagana1,cm_lim23,diane.fetalino}@yahoo.com,
{moranojulieann,ryeryeryeryeye7}@gmail.com

Abstract
In this paper, we model Swardspeak – as depicted in Twitter – through computational means. Swardspeak is an argot used by gays in the Philippines (Ponce, 2008) and this paper contributes to the growing literature of studies regarding gay language. Language modeling, the process of creating smaller representations of language, was applied and word and character sequences were generated. Language modeling provides systematic as it gives an overview of frequent occurring sequences and makes the data to be analyzed more manageable (Pablo et al., 2014). Analyses of word sequences reveal that Swardspeak use in Twitter is mostly limited to nonce-borrowing or single-word code-switching while analyses of character sequences using an approach that measures language similarity (Oco et al., 2014b) show similarities with other Philippine languages. Character differences were also generated to show replacement conventions (example: \textit{kili-kili} to \textit{jili-jili}). The results are discussed and also detailed are future directions. As application, two computer programs were implemented using a rule-based engine (Oco et al., 2014a): (1) collected tweets were used as training data to implement a language identification system; (2) and a dictionary integrated into a word editor was also built using existing resources (Chiong et al., 2009; Komisyon sa Wikang Filipino, 2004) on Swardspeak as guide.

Keywords: Filipino Lavender, swardspeak, language modeling, character difference, trigrams, language tool

1. Introduction to the history of the Filipino languages

The Philippines is a Southeast Asian country which consists of 7,107 islands with 186 number of individual language listed.\footnote{http://www.ethnologue.com/country/PH} “Alarmingy, according to current estimates, only one-tenth of today’s languages will remain by the coming of the 22nd century. Languages are disappearing at an astonishing rate and the pace shows no sign of abating. Many Filipino languages are at high risk of extinction.”\footnote{http://www.ethnicgroupsphilippines.com/people/languages-in-the-philippines/} The Filipino languages have been influenced by many other language groups throughout the history, as being a part in the Pacific Ring of Fire where many other countries surrounding the Philippines had an access for many opportunities for trade and correspondence with other nations and languages. And throughout these languages, one of the sub-languages that came was the Gay-Lingo language or swardspeak. Let’s take a look on the evolution of languages in the Philippines.

The first Indonesians are thought to have come to the Philippines in groups, beginning some 5,000 to 6,000 years ago and again about 1500 B.C. (Bautista). Linguistic evidence connects Tagalog with Bahasa Indonesia as having common roots, so the main root of the modern Filipino languages probably came with these people (although other groups of people are thought to have come to the Philippines much earlier). In the 16th Century, Spain claimed the Philippines for its own. The new Republic of the Philippines did not last long until American occupation began. The Americans began English as the official language of the Philippines.
There were many reasons given for this change. English began to be taught in the schools to all. During World War II, Japan occupied the Philippines for three years (Bautista). At this time English was still the official language of the Philippines; however, Japanese certainly influenced the various dialects during this time as well. (Stevens, 1999)

Additionally, the languages of the Philippines was not really from the country itself due to the reason of the occupation of different countries throughout the centuries. Over the course of its development, Tagalog (and other languages of the Philippines) have been influenced by Chinese, Japanese, Spanish, English, and many other languages, in trade and in occupations by various countries. Moreover, Philippines have taken and adapted different words from all of these languages to make part of their own languages. They have, however, still maintained their own languages and were able to have their new sub-languages. Example of this was the G-words (adding letter “G” to every word), the Jejemon (uses numbers and capitalization of letters) and the most common used Gay-Lingo (language from gays).

Gay-Lingo was from gays which was said to be their own version of language such that not all people can understand it unlike the English Language which was very common. Girls and most commonly teenagers were now using these words but do not expect males to do so. The use of these words depends on the people who tend to use it. It may be for fun, for teasing, for an occasion, or just for everyday lives seems to be like natural in using the gay-lingo words.

In this paper, we model Swardspeak (Gay-Lingo) language and how it was understood by the people using and just even hearing it. By this, we might lessen the difficulty of understanding the gay-lingo words and the meaning behind those words. Such that, misunderstanding and communication might be lessen and people can understand what each others was saying especially when a person is talking to a gay person.

2. Different sub-languages and related works

2.1 Sub-languages
Human beings can communicate with each other. We are able to exchange and share each others knowledge, beliefs, opinions, wishes, threats, commands, thanks, promise and etc,. We can laugh to express amusement, happiness, we can smile to express amusement and approval, we can shriek to express anger, excitement, or fear, and lastly we can clench our fists to express determination, anger or a threat, and so on, but the root of having this communication between human beings is because of language.

Language is a system for communicating. Written languages use symbols to build words. The entire set of words is the language's vocabulary. The ways in which the words can be meaningfully combined is defined by the language's syntax and grammar. The actual meaning of words and combinations of words is defined by the language's semantics. 8

Table 1 Tagalog words and its value in different languages.

<table>
<thead>
<tr>
<th>Tagalog</th>
<th>English</th>
<th>G-words</th>
<th>Jejemon</th>
<th>Gay-Lingo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wala</td>
<td>None</td>
<td>Wagalaga</td>
<td>W4l4</td>
<td>Waley/Wit/Wasiwas</td>
</tr>
<tr>
<td>Babae</td>
<td>Girl</td>
<td>Bagabae</td>
<td>B4b43</td>
<td>GirLalu</td>
</tr>
<tr>
<td>Nanay</td>
<td>Mother</td>
<td>Naganagay</td>
<td>NhaN4y</td>
<td>Mudra/Mudrakels</td>
</tr>
<tr>
<td>Tatay</td>
<td>Father</td>
<td>Tagatagay</td>
<td>T4t4y</td>
<td>Pudra/Pudrakels</td>
</tr>
</tbody>
</table>

8 http://www.webopedia.com/TERM/L/language.html
Table 1 shows the Tagalog words and its value in English language, G-words and Jejemon which was commonly known in the Philippines and lastly, the Gay-Lingo language which was the main topic of this paper.

Each of these sub-languages has its own way in changing the Tagalog words into these sub-languages. Mainly, this paper tackles the Gay-Lingo words and how it was changed and what was the value of the word being used.

**2.2 Related works**

According to Lunzaga, Ariel B. “Morphological Analysis of Gay’s Spoken Discourse” (April 18, 2011), “The language of gays known as gay speak has now earned respect from the community and observably been infused in the mainstream language of the society. Language is evolving time by time due to many aspects of the environment. This study focused on the morphological analysis of the language of gays who were studying at SLSU-Tomas Oppus in SY 2009-2010. Specifically, this identifies the common vocabulary of local gays and the corresponding meanings through a self-administered questionnaire and an informal conversation. Based from a conducted survey, it was learned that gay speak is simply an adaptation of the mainstream language like English, Filipino, and Cebuano. Gays construct their own language through simple reversal, syllabic reversal, simple reversal with affixation, clipping with affixation, straight words with affixation, and connotation through images. It was concluded that the spoken discourse of college gays in the campus violates the rules of English grammar yet used by gays for purposes of expressing themselves in a way unrecognizable by the discriminating straight men in highly patriarchal society.
According to Garcia, Nestlie U “Study on the usage of gay lingo among heterosexual students in University of the Philippines College of Arts and Letters”, “Language is not stable and static; it moves, evolves and progresses as our society take a step forward for improvement. Because of this, many sub-languages are being formed and created for different purposes. One of these sub-languages is 'Gay lingo', the language being spoken by gays. However, some heterosexuals now use it for their everyday conversation. In view of this, the researchers were led to do a study on the problem, Why is there a usage of gay lingo among heterosexual students in University of the Philippines College of Arts and Letters (UPCAL) The thesis aimed to identify the reasons behind the usage of Gay lingo, the source/s through which the participants were being influenced, cite situation/s wherein Gay lingo was commonly used, and enumerate Gay terms commonly used by heterosexual students in UPCAL. Participants that take place was said to be in the number of 30 heterosexual students in UPCAL, male and female alike. The study found out significant results, which met the objectives. First, that Gay lingo was used for communication purposes, for entertainment or fun, and also because it has become a fad. Second, that the participants were greatly influenced by the community or environment and media to speak the language. Third, that the research participants use Gay lingo only in informal occasions or situations. Lastly, the study was also able to enumerate Gay terms commonly used by the participants.”

3. Methodology

In this chapter, researchers will discuss the processes that will serve as the medium in the study. This chapter will also show the researchers method on gathering the needed corpora in this study. The procedure for gathering the different corpora where divided into two parts: Parallel and Online Corpus. Since the research was all about the Gay-Lingo language, researchers should be able to collect many resources for the testing of the research.

3.1 Collection of corpora

A. Parallel Corpus
This method was applied by collecting the text from a book and online sources via copy-paste with the word (gay-lingo), the definition of it and what was the meaning of the word in the Tagalog Language. Figure 1 shows the representation how the corpus was gathered from different sources through parallel corpus method. Around 1500 training data for parallel corpus were collected.

![Figure 1 Parallel Corpus Method](image)

B. Online Corpus
This method was applied by collecting the text and converting the words from different videos by listening to each one of them. Figure 2 show the representation how the corpus was gathered from a main source of videos and then wrote them manually by using online corpus method. Around 1000 training data for online corpus was collected. Figure 2 shows the representation how the online corpus were gathered.
3.2 Computational Comparison
For the researchers to track and distinguish the differences between the speak words to sward speak. We used an application that identifies the similarities and differences from the two sets of model thus, these sets of models were implemented.

Almost 5,000 tweets were collected via the Tweet Collector by random with Gay-Lingo words in it were used as training data, 5000 tweets were generated into one-gram, bi-gram, tri-gram and four-gram (Oco et al., 2014) giving the results of the most used word(s) in each section to the least one via Apache Nutch application.\(^1\) We used the one-gram and filtered the top 100 Gay-Lingo words that were most used in the tweets collected via VLookUp on Excel. Note that, word gram classify punctuation marks as part of the numbering of the words. Now, considering from the parallel corpus, what is found in the corpus and found in the one-gram via the tweets searched were considered to be Gay-Lingo words. While there are 20 keywords been used also as our training data, we searched for it manually through twitter and analyze whether these keywords were true positive or false positive.

Next, on the parallel corpus (dictionary), researchers filtered the different categories and look where the word diff\(^2\) can be applied we rank the 8 categories naming: Neolohismo o Paglika, Paghalu-halo, Metatesis, Malayang Pagdaragdag, Pag-aangkop, Pagpapalit, Reduksyon at Pagkakaltas and from most that can be used to for the word diff to the least. Rank 1 was Pagpapalit, next is Reduksyon and last is Pagkakaltas. We only used three categories then apply it to word diff where the character differences can be seen whether addition and replacement of letters happened. Follows was how to distinguish the character differences in each word.

“\([+\{\text{character}\}]\)" means an insertion of character;
“\([-\{\text{character}\}]\)" means a deletion of character; and
“\([-\{\text{character}\}, +\{\text{character}\}]\)" means a change of character from set 1 to set 2.

4. Results and discussion
After the processing of different corpora and adding it to different tools, researchers now came up with the following results.

<table>
<thead>
<tr>
<th>Word</th>
<th>GL Word</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>tulog</td>
<td>shulog</td>
<td>Pagpapalit</td>
</tr>
<tr>
<td>mongoloid</td>
<td>mongi</td>
<td>Reduksyon</td>
</tr>
<tr>
<td>inom</td>
<td>jinum</td>
<td>Pagkakaltas</td>
</tr>
</tbody>
</table>

---

\(^1\) http://nutch.apache.org/
\(^2\) http://www.gnu.org/software/wdiff/manual/wdiff.html
Table 2 shows the example for the both sets of word which undergone infiltration to categorize it accordingly, whereas 3 categories were being represented: Reduksyon, Pagpapalit, Pagkakaltas.

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>itey</td>
<td>979</td>
</tr>
<tr>
<td>ditey</td>
<td>856</td>
</tr>
<tr>
<td>keribels</td>
<td>813</td>
</tr>
<tr>
<td>sinetch</td>
<td>760</td>
</tr>
<tr>
<td>imbiyerna</td>
<td>396</td>
</tr>
</tbody>
</table>

Table 3 shows the example of how the word one-gram was applied on the tweets.

The following tables show the set of models that would undergo in the program and the output differences of it from each other.
Table 4 Word Diff

<table>
<thead>
<tr>
<th>SPEAK WORD (set 1)</th>
<th>SWARD SPEAK (set 2)</th>
<th>DIFFERENCE(S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ito</td>
<td>itechiwarA</td>
<td>i t [-o-] {+echiwarA}</td>
</tr>
<tr>
<td>Akin</td>
<td>aketch</td>
<td>a k [-i n-] {+etch}</td>
</tr>
<tr>
<td>punta</td>
<td>jonta</td>
<td>[-p u-][+j o+] n t a</td>
</tr>
<tr>
<td>hintay</td>
<td>jontay</td>
<td>[-h i-][+j o+] n t a y</td>
</tr>
<tr>
<td>buntis</td>
<td>jontis</td>
<td>[-b u-][+j o+] n t i s</td>
</tr>
</tbody>
</table>

Table 4 shows how the Tagalog words were changed into Gay-Lingo word by seeing the Character Difference through word diff.

Table 5 Searched Keywords

<table>
<thead>
<tr>
<th>True Positive</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jinet</td>
<td>Chaka</td>
</tr>
<tr>
<td>Ditey</td>
<td>Echos</td>
</tr>
<tr>
<td>Imbyerna</td>
<td>Chos</td>
</tr>
<tr>
<td>Wetpaks</td>
<td>Gora</td>
</tr>
<tr>
<td>Kayiz</td>
<td>Chaka</td>
</tr>
<tr>
<td>Imbey</td>
<td>Echos</td>
</tr>
<tr>
<td>Kemerut</td>
<td>Chos</td>
</tr>
<tr>
<td>Keribels</td>
<td></td>
</tr>
<tr>
<td>Itey</td>
<td></td>
</tr>
<tr>
<td>Sinetch</td>
<td></td>
</tr>
<tr>
<td>Chenelin</td>
<td></td>
</tr>
<tr>
<td>Churvalu</td>
<td></td>
</tr>
<tr>
<td>Eklavu</td>
<td></td>
</tr>
<tr>
<td>Mudrakels</td>
<td></td>
</tr>
<tr>
<td>Pudra</td>
<td></td>
</tr>
<tr>
<td>Junakis</td>
<td></td>
</tr>
</tbody>
</table>

Table 5 shows the True positive and False positive within the top 20 keywords used as training data.

From about 269 set of paired words being gathered, filtered and implemented in the word diff, based from the results of the program most of the gay-lingo word were added by some letters such as ‘ch’ which is commonly based from the Japanese words, others were being replaced by the letter ‘j’ which showed from table 4 (see Table 4.2), it showed that the most of the words we’re being substituted frequently with ‘ch’ and ‘j’ as shown in table 4.2.

However, 3 categories were used to label each words (see Table 2) whether what category does it belong, thus this adds a little help to analyze on how the speak word was being
transformed into a gay-lingo word depending on the category it was been ranked. For every rank or category speaks how a certain word being changed does. For the first rank which is ‘Pagpapalit’ this category means that a certain word’s character was been changed and added another character that leads it longer. Another rank was ‘Reduksyon’ in English term reduction, which means to reduce, whereas a certain word’s character is being reduce and nor not been replaced. And lastly is the ‘Pagkakaltas’ in which word is being cut/replaced and added before or after of a certain character and leads on the transformation of speak word into a gay lingo word.

Another approach is keyword searching, by the use of twitter we’ve manually searched for each training data to determine whether if this keyword has different meaning for the user or else they mean it as usual meaning as a gay lingo word. These 20 keywords that had been used results 16 true positive and 4 false positive as determined in searching via twitter.

After doing some processes with the Grammar Checker by inputting the Gay-Lingo words within the three (3) categories only, researchers tried to input a gay-lingo word and see how the grammar checker would react in the Gay-Lingo word. For example, the word “ades” in the Gay-Lingo word, after inputting it in the Grammar Checker, results would give, “Correction: Pera” which was the meaning of the “ades” word in Tagalog Language. Results have shown that accuracy of the right meaning of the Gay-Lingo word would give the right meaning to the word. Also, those Gay-Lingo words which were found in the dictionary were the only one to take place in the tool. The Grammar Checker was not case sensitive which means not having a capital letter at the first word would still give the meaning if it is in the dictionary but would give a Message(see fig. 4.1). Thus, saying that Gay-lingo words not inside the Grammar Checker would not be included in the checking of the results because definitely no results will be shown.

![Fig. 3 Grammar Checker Result](image)

Figure 3 shows that the Grammar Checker gives the meaning of the Gay-Lingo word ades which is pera. It was checked in the Text Language of Tagalog.

5. Conclusion and recommendation

Researchers now conclude that with the use of the dictionary and different corpora, fed it up to different tools, Gay-Lingo words could be easily understood by people which may lessen the misunderstanding and miscommunication among people who are familiar with the language and to those who are not. The categorization of in-groups versus out-groups can be generated when customers recognize certain cues, such as language, that delineate cultural differences (Brickson, 2000; Nkomo and Cox, 1996). Not being able to identify with other cultural groups may lead to negative attitudes towards such groups (Bartel, 2001).
Consequently, the understanding of the link between language and social identity patterns is of great importance to the community.

Since researchers focus on the new sub-languages that were formed in the Philippines, we would like to recommend future researchers to seek and find other sub-languages that were being used by the Filipinos. Examples of these were the G-words and the Jejemon which was common to few Filipinos and was given a bit of background from the statements above.

References


The Philippines: Then and Now. Excerpted from The Filipino Americans (From 1763 to the Present): Their History, Culture, and Traditions, by Veltisezar Bautista.


The Old and New Tagalog Orthography: A computational comparison

Manolito Octaviano Jr., Nathaniel Oco, Rachel Edita Roxas
National University, the Philippines
manolito.octaviano,nathanoco, rachel_roxas2001@yahoo.com

Abstract
This paper reports the orthographic changes between present-day Tagalog and its counterpart from the Spanish colonial era (c. 1521 – 1898). We empirically showed how the language has evolved. As training data, the following were used: (1) a 16,000-word collection of Tagalog literary works (Oco et al., 2015) to represent the early orthographic system; and (2) a 16,000-word subset of the text version of the Tagalog Wikipedia (Oco and Roxas, 2012) to represent the present-day use of the language. The differences and similarities of words from the two sets of data were automatically derived using edit distance (Levenshtein, 1966), string length, and frequency count (i.e., words with less than three characters or those with low frequency counts were removed). Character changes such as ‘c’ vs. ‘k’ (e.g., cayo vs. kayo) and ‘u’ vs. ‘w’ (e.g., auit vs. awit) were noted. Character trigrams (Oco et al., 2014) were generated and a similarity metric was applied (Oco et al., 2013) and results showed 69% similarity. On the other hand, the remaining 31% showed Spanish influence and evolved through the years (e.g., cusina to kusina). These results validate that “language is a system of arbitrary vocal symbols by means of which society cooperates” (Bloach and Trager, 1942). Results of this research contribute to the growing list of studies in language change in the country (Ilao et al., 2011).

Keywords: orthography, Tagalog, edit distance, character trigrams, similarity metrics

1. Introduction

Kamusta, Huwag, at Kami are some of the familiar words of the modern orthography of Tagalog. These are few of the words that average Filipinos can understand in a snap being Tagalog as the most frequently spoken language. (Roxas, 2009). Will Filipinos still be able to recognize the words written in the old orthography of Tagalog like the words Ciya, Come esta, and auit? Probably the current generation does not, since words written in the old orthography are used negligibly nowadays. The words Ciya, Come esta, and auit came from the Spanish system that greatly influenced the Filipino language today.

Looking back in the history of the Philippines before Spanish came; we can see the influence of different foreign language group in the Filipino language because of the different trade and correspondence that happened with the other nations (Stevens, 1999). Not only did the foreign language groups influenced the language but also the country’s barangay groups that traded freely amongst each other during those time.

In the 16th century, colonization of Spanish in the Philippines (c. 1521 – 1898) had a great impact in the culture of the Filipinos. One notable influence of the Spanish is the alphabet they implemented in Filipino schools - the Abecedario alphabet. The alphabet system introduced by Spanish variously had 28, 29, 31, or 32 characters. It replaced the Baybayin script, the ancient Tagalog alphabet, which was believed to have existed long before Spanish colonization. It also became the Tagalog orthography of the generation in the Spanish era. Filipinos followed the 32-abecedario system in the beginning of the 17th century from the time when the Spanish took over the country. The first ever book printed (Doctrina Cristiana) in the Philippines (c. 1590) was believed to be an evidence of the Abecedario alphabet (Tan, 2014).

3 http://filipinokastila.tripod.com/FilSpa.html
Years after Spanish colonization, different sets of alphabet were introduced to Tagalog: from the *Abecedario* alphabet to the present-day *Modern Filipino alphabet*. The new set of alphabet used in the present-day orthography consists of 28 characters, omitting some characters from the previous sets of alphabet (e.g. ll, rr). However, orthography in the *Abecedario* alphabet is still in use even after the transition of different sets of alphabet.

In this paper, we reported the orthographic changes between present-day Tagalog (i.e. used of *Modern Filipino alphabet*) and its counterpart (i.e. used of *Abecedario alphabet*) from Spanish colonial era. Knowing these changes, we can determine how much the language has evolved over the past years from the influence of Spanish.

2. Language and related works

2.1. Filipino language

“Language is a symbol system based on pure or arbitrary conventions… infinitely extendable and modifiable according to the changing needs and conditions of the speakers,” Robins said when he defined language. From this definition, we can conclude that the language is a collection of symbols organized to form meaning that human can understand. For example, the symbol (or character, considering the spelling symbol) ‘e’, ‘a’, and ‘t’ when structured will produce the word –‘cat’; combinations of words will then produce a sentence (used for communication). These words and sentences form a language used as means for communication (The Nature of Language Linguistics, 2010).

Language reflects the culture and history of one’s civilization from one generation to the other. When the language is handed down to other generations, words are constantly changing considering its meaning and spelling; new words are supplemented to adapt the needs of community. The Philippine language, specifically the Tagalog, is a good example since it changed for a number of times already.

Tagalog is the widely used language in the Philippines. It took a long haul before it was declared the national language with a formal name of Filipino. Looking back on the history, the influence of different countries changed the life and the culture of Filipino people. Spanish, the longest colonist of the Philippines for approximately 400 years, influenced the country in a lot of things. An example is their influence in the Filipino language that had a great effect on the current language Filipinos use today (Bautista, 1998). About 20% of the Tagalog words used today are from them (Rebecca, 2009). For instance, the word “Kamusta” (means how are you) originated from the Spanish “Come esta”.

Before Spanish came to the Philippines, Filipinos had their own alphabet and their own system of writing, Baybayin – consists of 3 vowels and 14 consonants (Gugol, 2011). When the Spanish ruled the country, they romanized the Baybayin and introduced their own alphabet: *Abecedario* and their own writing system. From these, Tagalog language underwent significant development. Spelled words from previous alphabet were replaced and new words came out (e.g. cusina, auit, and cayo). Table 1 shows the 31 *Abecedario* letters with its corresponding pronunciation.

<table>
<thead>
<tr>
<th>Letter</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>CH</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Spanish) Pronunciation</td>
<td>Ah</td>
<td>Be</td>
<td>Se</td>
<td>Che/Se-hache</td>
<td>De</td>
<td>Eh</td>
<td>Efe</td>
<td>Ge/He</td>
</tr>
<tr>
<td>Letter</td>
<td>H</td>
<td>I</td>
<td>J</td>
<td>K</td>
<td>L</td>
<td>LL</td>
<td>M</td>
<td>N</td>
</tr>
</tbody>
</table>
Abecedario began to decline in the 17th century and became obsolete in the 18th century (Morrow, 2007). Different sets of alphabet were introduced to Filipinos before attaining the Modern Filipino alphabet being used today.

Table 3 shows the summary of the evolution of alphabets after the Abecedario alphabet of Spanish. The Modern Filipino alphabet is also the same as the English alphabet but with additional characters from Spanish (ñ and ng). It replaced the old orthography Spanish introduced. In addition, it is the alphabet that is currently being taught in Philippines.

| Abakada alphabet | a, b, k, d, e, f, g, h, i, l, m, n, ng, o, p, q, r, s, t, u, w, y |
| Pilipino alphabet | a, b, c, ch, d, e, f, g, h, i, j, k, l, ll, m, n, ñ, ng, o, p, q, r, rr, s, t, u, v, w, x, y, z |
| Modern Filipino alphabet | a, b, c, d, e, f, g, h, i, j, k, l, m, ñ, ng, o, p, q, r, s, t, u, v, w, x, y, z |

2.2. Computational comparison

One way in determining the distance of words of the training data is by using Levenshtein distance also known as edit distance (Levenshtein, 1966). Levenshtein distance computes for the difference of two strings. It identifies the cost of making insertions, deletions, and substitutions of transforming one string to the other. For example, the distance between “kitten” and “sitting” is 3:

- kitten → sitten (substitution of "s" for "k")
- sitten → sittin (substitution of "i" for "e")
- sittin → sitting (insertion of "g" at the end)

Another technique used in this study is the Trigram modeling. Trigram modeling is commonly used in implementing a language identification system (LID system). It slices the word in 3-character group. For instance, the word “test” will produce the following trigram: “_te”, “es” “est”, “st_”. Different studies use the modeling approach to further understand the Philippine languages such as measuring the similarities of the languages (Oco, 2013) and implementing an LID system for Philippine languages (Oco, 2015).

3. Methodology

3.1. Collection and modeling

As training data, we used the collections of Filipino literary works to represent early orthography (Spanish system) and Tagalog Wikipedia to represent the present-day orthography (Oco, 2015). The training data were placed in Notepad++. Regular expressions were utilized on the training data to remove unnecessary keywords (e.g. numbers and symbols). Afterwards, the first 16,000-word of both training data were extracted and fed in an automated tool4 to produce unigram models - collection of unique words that are used in the documents. Generated models were then filtered (i.e. leaving all words that has character

---

4 http://www.speech.sri.com/projects/srilm/
length greater than 2) and saved in text file. From the 16,000-word document, 4,265 unique words remained for early orthography while 4,808 unique words for present-day orthography.

3.2. Computational comparison

In order to track the evolution of words, an application was implemented to identify the differences and similarities from the two sets of models. Integrating the edit distance algorithm, sets of models were fed in the application. Table 3 shows the example output of the program that the word undergone 1-character change from column 1 (set 1) to column 2 (set 2). The difference of words are shown in column 3, explanations are as follows:

- “[+ {character}]” means an insertion of character;
- “[– {character}]” means a deletion of character; and
- “[– {character}, + {character}]” means a change of character from set 1 to set 2.

<table>
<thead>
<tr>
<th>Table 4 Edit distance between words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set1</td>
</tr>
<tr>
<td>Din</td>
</tr>
<tr>
<td>Anot</td>
</tr>
<tr>
<td>Cay</td>
</tr>
</tbody>
</table>

Another approach used to determine how much the language has evolved was to generate character trigrams from the training data. Using Apache Nutch, trigrams were generated and similarity metric was applied.

Table 5 shows the generated trigram models from the training data.

<table>
<thead>
<tr>
<th>Table 5 Trigrams of training data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set1 (julieta)</td>
</tr>
<tr>
<td><em>ca</em></td>
</tr>
<tr>
<td>Ing</td>
</tr>
<tr>
<td><em>ma</em></td>
</tr>
</tbody>
</table>

4. Results and discussion

Table 6 Top 10 result of edit distance (frequent character changed)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Character difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[–C, +K]</td>
</tr>
<tr>
<td>2</td>
<td>[–g]</td>
</tr>
<tr>
<td>3</td>
<td>[+g]</td>
</tr>
<tr>
<td>4</td>
<td>[+n]</td>
</tr>
<tr>
<td>5</td>
<td>[–a, +i]</td>
</tr>
<tr>
<td>6</td>
<td>[–i]</td>
</tr>
<tr>
<td>7</td>
<td>[–I, +A]</td>
</tr>
<tr>
<td>8</td>
<td>[–A, +O]</td>
</tr>
<tr>
<td>9</td>
<td>[–u, +w]</td>
</tr>
<tr>
<td>10</td>
<td>[–n]</td>
</tr>
</tbody>
</table>

About 50% (2135 word-pairs) of the old orthography was noted that changed when the edit distance was used. Table 6 shows the top 10 character changed. The substitution of the letter

---

5 http://nutch.apache.org/
‘c’ to ‘k’ is the most frequent change that happened in the Tagalog orthography (e.g. cung to kung, Aco to ako, and daco to dako). The reason behind the changes in orthography was due to the evolution of the alphabet system in the Philippines. Different sets of alphabet were introduced: Abakada alphabet, removal of Spanish characters (e.g., rr, ll); Pilipino alphabet, removed Spanish characters were again used to accommodate English and Spanish words; Modern Filipino, diagraph letters in the alphabet were dropped. The orthography was revised accordingly to adapt in the transition of alphabet. Based on the similarity metric applied in the generated trigram models, it showed 69% similarity while remaining 31% showed the influence of the Spanish (e.g. _co, arr, laq ). Table 7 shows the top 10 generated trigram models for each set (top as the most frequent).

Table 7  Top 10 generated trigram models of training data

<table>
<thead>
<tr>
<th>Rank</th>
<th>Set1 (julieta)</th>
<th>Set2 (wiki)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ng_</td>
<td>ng_</td>
</tr>
<tr>
<td>2</td>
<td>Ang</td>
<td>Ang</td>
</tr>
<tr>
<td>3</td>
<td>_ma</td>
<td>_pa</td>
</tr>
<tr>
<td>4</td>
<td>_pa</td>
<td>an_</td>
</tr>
<tr>
<td>5</td>
<td>_na</td>
<td>_ma</td>
</tr>
<tr>
<td>6</td>
<td>an_</td>
<td>_na</td>
</tr>
<tr>
<td>7</td>
<td>_ca</td>
<td>_ka</td>
</tr>
<tr>
<td>8</td>
<td>Ing</td>
<td>Ala</td>
</tr>
<tr>
<td>9</td>
<td>Ala</td>
<td>Pag</td>
</tr>
<tr>
<td>10</td>
<td>Ina</td>
<td>Ina</td>
</tr>
</tbody>
</table>

5. Conclusion

Results validated that “language is a system of arbitrary vocal symbols by means of which society cooperates” (Bloach and Trager, 1942). The influence of other nations to the country, like how the Spanish influenced the Filipinos, has been a big factor why evolution of language happened. The Philippine adjusted its own orthography on the alphabet implementation of the colonist in order to adapt. Furthermore, result of the research have contributed to the growing list of studies in language change in the country (Ilao, 2011) since the Philippines lacks of study in its own languages.

References


FilCon: Filipino sentiment lexicon generation using word level-annotated dictionary-based and corpus-based cross lingual approach

Darlan Keen Domingo¹, Nyker Matthew King², Jerome Lorenzo Lopez³, Alexandra Mondares⁴, Charmaine Ponay⁵

Department of Computer Science
Institute of Information and Computing Sciences
University of Santo Tomas
España, Manila, The Philippines
¹2011031273, ²2011032329, ³2011032399, ⁴2011031293, ⁵csponay}@ust-ics.mygbiz.com

Abstract

The FilCon is a generated subjective lexicon for the Filipino language. It contains 22,380 words primarily based on a Filipino-English bilingual dictionary and aligned with SentiWordNet 3.0’s polarity values. The existing identical entries from the first iteration of lexicon generation are filtered according to their translation accuracy produced using Moses training sets. The final iteration involves sorting of words and the generation of final output in the form of a .txt file based on SentiWordNet’s format having: POS, ID, POS_SCORE, NEG_SCORE, ENG_WORD and with the additional attribute, FIL_WORD. The produced results were analyzed and interpreted by applying FilCon into various test cases of subjective classification of opinions to calculate the overall accuracy of the lexicon. The overall accuracy of FilCon based on a sentence sentiment analysis is 38.18%.

Keywords: Filipino, polarity lexicon, sentiment lexicon, subjective lexicon, sentiWordNet

1. The Problem and Its Background

1.1. Introduction

Several researchers have conducted many new approaches and algorithms to animate a computer with human-conversational capabilities. These approaches have led to the development of the branch of science known as Natural Language Processing (NLP). NLP, according to Liddy (2001), “is a range of computational techniques for analyzing and representing naturally occurring texts for the purpose of achieving human-like language processing for a range of tasks or applications”.

A language could convey messages delivered through verbal or non-verbal communication. A non-verbal communication contains textual information which can be categorized into two, facts or objective information and opinions or subjective information (Regaldo & Cheng, 2012). The subjectivity of a word can be determined by comparing to an existing subjectivity scores of the lexicon. Subjectivity scores contain positivity, negativity, and neutrality percentage values expressed in decimal form.

An example of an English lexicon is the WordNet, created by Princeton University in 2006 for the purpose of having a lexical-conceptual database. This database contains lexical units and their interrelationship (Bondoc, Garcia, Lacaden, Ping, & Borra, 2010).
One kind of WordNet is SentiWordNet. SentiWordNet is an enhanced lexical resource for sentiment analysis and opinion mining. This lexicon contains subjectivity scores, positivity, negativity, and neutrality polarity values.

This study was conducted in order to aid the studies in sentiment analysis. Before this study, the Filipino language polarity lexicon did not yet exist. FilCon is another tool like SentiWordNet; FilCon is in the Filipino language; whereas, SentiWordNet is in the English language. FilCon was constructed using a bilingual dictionary aligned with the polarity scores provided by SentiWordNet. Since neither the lexicon nor the bilingual dictionary provides information concerning the semantic meaning of the identical entries, FilCon has to rely on the most probable sense in the target language using Moses.

1.2. Background of the Study

Finding out people’s thoughts is an essential part of information gathering (Pang & Lee, 2008). Opinion mining and sentiment analyses are tools used to find out what people think. Sentiment analyses have been used by diverse disciplines and various industries in the community. Commercial, academe, and information technology industries are using sentiment analysis in analyzing customer and survey feedbacks.

Part of the sentiment analysis is subjectivity classification which describes a sentence as to whether it is objective or subjective. Subjectivity involves a bootstrapping approach. It involves the use a high precision classifier to automatically identify subjective and objective sentences. “It is based on manually collected lexical items and single words which are good subjective clues.” (Moghaddam & Ester, 2012)

With the current advancements in natural language processing, most sentiment analysis applications utilize a lexical resource that provides the scores needed to determine the positivity or negativity of a sentiment.

Since most sentiment analysis applications did not have a Filipino lexicon as a base language this study focused on generating a lexical resource in the Filipino language.

1.3. Theoretical Framework

Opinion lexicons, also called as subjective lexicons, are resources consisting of words annotated with sentiment polarity scores or subjectivity.

Two of the most frequently used opinion lexicons are OpinionFinder and SentiWordNet. OpinionFinder is a subjective lexicon. On the other hand, SentiWordNet is a polarity lexicon. OpinionFinder was compiled from manually developed resources augmented with entries learned from corpora. Its entries have been labeled for part of speech including subjectivity in to which, words that appeared in most subjective context are labeled strong while those that appear often are labeled weak.

The other lexicon, SentiWordNet, is created by Esuli and Sebastiani (2006) as a subset of WordNet. SentiWordNet assigns to each synset of WordNet three sentiment classifiers with scores of polarity, positive, negative, and objective.

Opinion lexicons were built using several approaches. The simplest approach that has been attempted for building these opinion lexicons in a new language is the dictionary-based approach, wherein the existing lexical resource is translated using a bilingual dictionary.
In a study conducted by Mihalcea, Banea and Wiebe (2008), a dictionary-based approach was used to generate a subjectivity lexicon for the Romanian language. Starting with the English opinion lexicon from the OpinionFinder, words were lemmatized for translation using an English-Romanian dictionary. However, in some cases, translations loses their subjectivity due to word ambiguity and lemmatization. Considering these problems, the researchers sought to find the most probable translation for each English entries with direct translation.

To evaluate the lexicon, two native speakers of Romanian annotated the subjectivity of 150 randomly selected entries. The subjectivity of every entry was judged in the context where it frequently appears on websites, accounting for its common meanings. Results show that 123 entries were correctly translated.

The Figure 1 is a block diagram of Mihalcea, Banea and Wiebe’s (2011) study in building Romanian Subjectivity Lexicon.

Furthermore, another study of Hamouda and Rohaim (2011) states that, the SentiWordNet can also be used as an important resource for sentiment classification tasks. These researchers conducted a study to classify sentiments and determine the subjectivity of English reviews. As a preparation for the sentiment classification, the following linguistic analysis processes were made for pre-defined input reviews:

- Tokenization process: splits the text into very simple tokens such as numbers, punctuation and words of different types.

- Sentence Splitting process: segments the text into sentences. This module is required for the Speech Tagging Process. The splitter uses a list of abbreviations to help distinguish sentence-marking full stops from other kinds.
• Speech Tagging Process: produces a part-of-speech tag as an annotation on each word or symbol.

• SentiWordNet Interpretation: matching the words with the SentiWordNet semantic entries.

• SentiWordNet Orientation: estimates the polarity score of a word in regards to a sentiment class (negativity, positivity and neutrality) and its relative frequency of appearance in that class is carried out using this method.

![Diagram of linguistic processes]

**Figure 2 Hamouda and Rohaim’s (2011) Review Sentiment Classification using SentiWordNet**

After the linguistic processes, the techniques to calculate the positive and negative scores for each review were proposed. These are the ‘Sum on Review’ and ‘Average on Sentence and Average on Review’. The results of the study extracted from these techniques show significant improvement in the overall accuracy to be 67% and 68.63%.

### 1.4. Conceptual Framework

Data mining is defined as the computer-assisted process of searching and modifying large sets of data followed by the extraction of its meaning (Alexander, n.d.). The data mining in FilCon aim to retrieve Filipino words as FilCon's primary input used for further processing.
In the first text file generation containing the retrieved words, each Filipino word was translated to English using a Filipino-English bilingual dictionary. The translated words were cross-referenced with the same English entries but with annotated polarity scores from SentiWordNet 3.0. These polarity scores were also retrieved and aligned together with the corresponding Filipino entries.

The SentiWordNet 3.0, a basis of polarity scores describing the positivity and/or negativity of a word, also provided the additional needed fields for FilCon such as the POS, ID, POS_SCORE, NEG_SCORE, ENG_WORD and FIL_WORD.

These processes could be compared with the Figure I-1 where the dictionary-based approach was used. However, in the Romanian lexicon, the subjectivity of the words were annotated according to the English entries retrieved from OpinionFinder while this study used Filipino words which are annotated with polarity scores according to the English entries from SentiWordNet.

Furthermore, the retrieval of the Filipino sentences served as the training corpora for the language tool, Moses, to find the most probable translation for each entry thus, eliminating identical entries. This task was the counterpart of the direct translation involving the accuracy of the Romanian lexicon.

The FilCon generation applied processes of text analysis which are similar to the approaches used in Figure I-2. For instance, tokenization, SentiWordNet interpretation, and orientation were also be adopted by FilCon. The translated Filipino words were tokenized and cross-referenced with the SentiWordNet English entries, aligning the proper polarity scores from
SentiWordNet entries to FilCon entries. The reconstructed lexicon was, then, the output of the system which can be used in opinion classification.

1.5. Statement of the problem

In the field of Sentiment Analysis, the need for a lexical resource in different languages as a basis for determining subjectivity remains an open domain. In this regard, this study aimed to generate a Filipino Subjective Lexicon, through application of the word level-annotated dictionary-based and corpus-based cross lingual approach with the use of a Filipino-English bilingual dictionary, Moses, and SentiWordNet. Specifically, it also intended to answer the following questions:

1. Were the Filipino-English bilingual dictionary, Moses and SentiWordNet a possible combination in generating a Filipino Subjective Lexicon?

2. What was the accuracy level of the translated words against the entries of SentiWordNet?

3. What was the accuracy level of the generated Filipino Subjective Lexicon using the Filipino-English bilingual dictionary, Moses and SentiWordNet?

4. Was there a significant difference between the sentiment scores of the generated Filipino Subjective Lexicon through the use of the Filipino-English bilingual dictionary, Moses and SentiWordNet and expert’s sentiment scores?

1.6. Objectives of the Study

The main objective of this study was to develop a Filipino Subjective lexical resource with acceptable polarity scores by application of the word level-annotated dictionary-based and corpus-based cross lingual approach through the use a Filipino-English bilingual dictionary, Moses and SentiWordNet which could contribute to the development of the Natural Language Processing, specifically, Sentiment Analysis, which is an area in NLP.

The FilCon generation undergo the following different phases and modules:

a. **Data Mining.** In this phase, the researchers seek to retrieve Filipino words from a Filipino-English bilingual dictionary to use as FilCon’s primary input and to retrieve Filipino-English sentences to serve as the training corpora for further processing of lexicon entries.

b. **Word Translation.** The second phase indicated the translation of Filipino words to English for the cross-referencing of entries with the SentiWordNet 3.0.

c. **Cross-Lingual Projection.** The translation of the Filipino words were matched with the same English word found in the SentiWordNet entries. The aligned polarity score of a SentiWordNet entry were retrieved and aligned with the same semantic meaning in Filipino.

d. **Corpus Training.** The retrieved Filipino-English sentences were trained in Moses to provide each Filipino entry with their corresponding translation accuracy. The translation accuracy indicated the most probable translation for each word.
e. **Word Filtering.** This phase conveyed the acceptable polarity scores aligned with the Filipino words retrieved in terms of identical entries. It also involved the application of the translation accuracy to filter multiple entries with different translations and part-of-speech.

1.7. **Scope and Limitations**

The study aimed to provide a lexical resource annotated with polarity scores of the Filipino language using the word level-annotated dictionary-based and corpus-based cross lingual approach through the use of a Filipino-English bilingual dictionary, Moses, and SentiWordNet 3.0.

The study focused on the retrieved Filipino words as the primary input of FilCon. The output of this study, FilCon, would only provide the following fields: (a) Part of Speech (POS), (b) ID, (c) Positivity Score (POS_SCORE), (d) negativity score (NEG_SCORE), (e) corresponding English translation of the Filipino word (ENG_WORD), and (f) the Filipino word (FIL_WORD).

This study was only limited in providing its users the lexical resource and the polarity scores determined through pairing of SentiWordNet English word’s polarity score with its corresponding translated Filipino word. Sentiment analysis of scenarios and instances would not be covered by the study, however, it may be used for the testing of accuracy.

The researchers emphasized on the fact that this study would only help writers and linguists construct their literary texts with consideration of having a good diction by referring to the sentiment values. The researchers also made the study to aid programmers develop a system geared towards an acceptable basis for Filipino sentiment analysis of their programs. The sentiment values of the Filipino words generated by this study were based on SentiWordNet 3.0 and, hence, based on their criteria.

1.8. **Significance of the Study**

According to a study conducted by Martha Perry, Face to Face (FtF) communication is preferred by couples rather than Computer Mediated Communications (CMC); because CMC supposedly contributes to misunderstandings and frustration, which can lead to escalated conflict (Perry, 2010). If it were said that technology bring people together (Chiala, 2013), then, this study could help clear out the ambiguities between in CMCs and further enable the understanding a person’s message.

The study aimed to provide a Filipino Subjective Lexicon (FilCon) to be used in future Natural Language Processing applications. This study was able to give a significant benefit to entities especially to:

1. **Sentiment Analysis Systems Programmers**
   FilCon can provide programmers with the utmost accurate automated Filipino lexicon in order to efficiently integrate the analysis of subjectivity of Filipino context into their sentiment analysis systems.

2. **Artificial Intelligence Systems Programmers**
   Artificial Intelligent (AI) systems mostly rely on subjectivity lexicon (pre-learned) and machine learned experience. With subjectivity lexicon mostly made from the English language, it would be of a great milestone to be able to develop an AI system in the Filipino language.
2. The FilCon

Scaling would be assigning a metric system as a standard unit of measurement for a certain object. To scale a subjective statement, sentiment analysis is required. Such analysis are conducted in order to draw acceptable conclusion regarding a certain matter based on the standard scale. This standard scale is the FilCon.

FilCon is generally an opinion polarity lexicon which determined the polarity (positivity and negativity degree) of a word which then contributes in the analysis of the subjectivity and the sentiment of a given document. With the use of FilCon, annotating Filipino words for subjectivity analysis would be automatically be available to its users without human intervention.

FilCon was basically constructed through the use of dictionary-based approach. This approach gave a developing language an advantage for a seamless generation of words in the lexicon through direct translation at a word/phrase-level. This was how FilCon came into existence. FilCon made use of an English-Filipino bilingual dictionary in translating synsets in the SentiWordNet lexicon.

Sampling of words from the FilCon was made through the use of MOSES training tool. MOSES did the training through aligning of English sentences with a human-provided counter-translation Filipino sentences. These sentences were fed into MOSES for aligning. This process, then, yielded translation probability accuracy for a group of words in FilCon. The testing of the accuracy of subjectivity was measured through the use of corpus-based approach.
3. FilCon Creation

3.1. System Architecture

The architecture represents the system flow of the generation of Filipino words aligned with polarity scores in FilCon.

3.2. Description of Modules and Interfaces

a. **Data mining:** The Filipino words were retrieved to serve as the initial entries for FilCon. There was no separation of subjective words from the objective words.

b. **Translation and POS Tagging:** The Filipino words, as input strings, undergone translation and POS Tagging processes. With the use of a Filipino-English bilingual dictionary as a basis, the English translation and the Part of Speech of a Filipino word was provided, hence, translation was done.
c. **Cross-Lingual Projection:** Once the Filipino words were given with English translations, the polarity scores were annotated. The English entries and their corresponding polarity scores from SentiWordNet 3.0 were matched with the same English translation of a specific Filipino word. Word-matching and polarity scores annotation was another term for this module.

d. **Moses Training:** This module trained sets of corpora. With the Filipino sentences having their English translations, the translation accuracy for each word can be obtained by determining the frequency of appearance of the Filipino word in the sentence and then, retrieving the English translation counterpart in the English sentence. In this way, the translation accuracy becomes a basis for the most probable translation of a word thus creating a new database for each Filipino word with the translation accuracy.

e. **Elimination of Identical Entries:** This module seek to filter the words by finding the most probable translation for a specific word, reducing the various translations made. Using the translation accuracy, the following cases are addressed:

   a. **Case 1:** If there were different translated entries but with the same translation accuracy for a specific Filipino word, the word filtering was done by manual selection of the most probable meaning of each word.

   b. **Case 2:** If there were different translation and different translation accuracy for a specific Filipino word, the word filtering was done by automatic selection of the word with the highest translation accuracy.

   c. **Case 3:** If there were no identical translated entries for a specific Filipino word, it was automatically included in the final output to be sorted.

   d. **Sorting:** Through PHP scripting, the filtered lexicon database obtained with translation accuracy were joined to form a sorted lexicon, FilCon, with POS, ID, POS_SCORE, NEG_SCORE, ENG_WORD and FIL_WORD.

3.3. Results of Conducted Research

The Filipino-English bilingual dictionary, Moses and SentiWordNet 3.0 were combined to generate a Filipino Subjective Lexicon. The Filipino words were translated to the most probable translation using the translation accuracy generated by Moses. FilCon was generated, containing 22,380 words annotated with polarity scores. However, there were around 30-40% part of the FilCon that were with lemma and the remaining were in the base form (root word). These results were used for analysis and interpretation.

4. Evaluation

4.1. Analysis and Interpretation of Results

In order to evaluate the accuracy of each word in the generated Filipino Subjective Lexicon using a Filipino-English bilingual dictionary, Moses, and SentiWordNet 3.0, the translation accuracy was used from the training sets in Moses. If there were more than one translations for a specific word, the translation with a high translation accuracy was interpreted to be of the nearest possible translation for that corresponding Filipino word. However, if there was at most one translation for that specific word, then it was considered as the probable translation available.
Moreover, the translation accuracy for these Filipino words were averaged and multiplied to 100% to evaluate the accuracy level of translation of entries. By applying the formula of averaging in Figure 5, the overall translation accuracy yielded an accuracy of 63.622%. Such result was obtained since some of the alignment during the MOSES training is not properly matched with the right translation; hence, it may result into a mistranslation in the FilCon entry.

For the overall evaluation of the accuracy level of FilCon, the researchers tested the output of the system by using sentences that were handpicked and randomly selected from various sources. The results of the system (see Appendix D, Table D-1) were then compared to the evaluation of an expert using a two-tailed t-test with \( \alpha = 0.05 \) having the hypotheses \( H_0: \mu_1 = \mu_2 \) and \( H_1: \mu_1 \neq \mu_2 \).

Table 1 Sentiments confusion matrix

<table>
<thead>
<tr>
<th>EXPERTS</th>
<th>Positive</th>
<th>Negative</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>FILCON</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>11</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Negative</td>
<td>8</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Neutral</td>
<td>12</td>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1 shows a confusion matrix of the results of the sentences that were evaluated by the system versus the evaluation of the experts having the expert’s evaluation the basis. The right diagonal of the matrix shows how many sentences were evaluated the same by the system and the experts. The other cells show the number of sentences that were evaluated differently by the system and the experts. There were 21 correctly evaluated sentences and 34 incorrectly evaluated sentences for a total of 55 sentences evaluated.

\[
Accuracy = \frac{\text{Number of Correctly Evaluated Sentences}}{\text{Total Number of Sentences}} \times 100\%
\]

Figure 6 Formula for calculating FilCon’s accuracy

FilCon’s accuracy were then calculated using the formula given on Figure IV-2. Based on the calculation, FilCon’s accuracy is 38.18%.

\[
t = \frac{\sum d}{\sqrt{\frac{n(\sum d^2) - (\sum d)^2}{n-1}}}
\]

where \( d \): difference per paired value

\( n \): number of samples

Figure 7 T-test formula for paired two sample for means
Table 2  Summary of the T-Test computation

<table>
<thead>
<tr>
<th></th>
<th>System</th>
<th>Expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>-0.058906422</td>
<td>0.140606061</td>
</tr>
<tr>
<td>Observations(n)</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>t computed</td>
<td>-2.63523816</td>
<td>±2.004879288</td>
</tr>
</tbody>
</table>

Table 2 shows the means of the means of the results of the system and t computed using the formula in Figure IV-4 and the t critical from the table of t-values. The t computed, -2.63523816, is outside the range of the t critical values ±2.004879288, hence we reject $H_0$. It denotes that there was a significant difference between the system’s sentiment scores and the expert’s sentiment scores.

4.2. Summary of Findings

The generation of the FilCon yielded the following results:
1. The Filipino-English bilingual dictionary, Moses, and SentiWordNet was a possible combination in generating a Filipino Subjective Lexicon.
2. The translation (SentiWordNet from English to Filipino) accuracy of the FilCon entries is 63.622%.
3. The overall accuracy of FilCon was 38.18%.
4. The generated FilCon were significantly different from expert’s scores.

5. Conclusions

The researchers were able to generate a Filipino Polarity Lexicon, FilCon adapting the combination of word level-annotation approach, dictionary-based approach and corpus-based cross lingual approach through the use of language tools involving the Filipino-English bilingual dictionary, Moses and SentiWordNet.
The translation of FilCon entries and its cross-lingual projection with the SentiWordNet entries were made possible. Out of 6,124 sample data from the training sets of Moses used for computing the translation accuracy, the result yields 63.622%. It only showed that the translation of the Filipino words is relatively accurate.
The overall accuracy of FilCon depended on the sentence sentiment analysis. Based on the results, 38.18% was the accuracy.
In the t-test computation, there was a significant difference between the sentiment scores of FilCon and sentiment scores of experts. It denoted that scores from the generated lexicon were significantly different from expert’s scores.

Acknowledgements

First of all, we would like to thank the Almighty God for giving us the strength, wisdom, knowledge and perseverance to make this great endeavor possible. We lift up all praises and glory to Him.
We would like to express our deepest gratitude to our dear adviser, Ms. Charmaine Ponay, for guiding us throughout this endeavor. Our coordinator, Assoc. Prof. Perla Cosme for contributing her knowledge and ideas for the improvement of this study.
Lastly, we would like to share the success and thank our families and friends for all the support and understanding they have extended to us all.
References


An Empirical Study of English Corpus as a Reference Tool for PhD Students

LI Lan
Hong Kong Polytechnic University
lan.li@polyu.edu.hk

Abstract
The important role of dictionaries in language studies is widely accepted. ESL learners rely on dictionaries to find word meaning, pronunciation, grammar and usage. However, with the advancement of modern technology, especially with open access to free online language corpora, the consultation of corpora offers new possibilities and perspectives to advanced learners. As a result, particularly in relation to the productive use of language, the reliance on dictionaries may have been changed. Although corpus as an aid to language teaching and learning has been embedded in many linguistic subjects in tertiary institutions, this valuable resource has not been widely used by research students in other disciplines. Also, whilst there has been an abundance of literature on how corpus is applied in education by teachers and textbook compilers, little research has been conducted into student use of corpora and their opinions, strategies and problems in corpus use.

This project aims to investigate whether reference skills of postgraduate students can change, or have changed, from using dictionaries to using corpora in high-level academic writing. After a corpus workshop, PhD students from different linguistic backgrounds all agreed that corpus would be useful in their academic writing. The questionnaire survey revealed their enthusiasm in using something new and relevant. The post-training interviews revealed that although online dictionaries are mostly-used reference tool, the use of corpus is gaining a solid increase among PhD students in productive use of the English language, especially in thesis and journal paper writing. Students also regarded training in corpus use as highly necessary. The results can help promote corpus application in high-level academic writing curricular, and can invite critical reflections on methodologies of using data with more effective reference skills.

Keywords: corpus, academic writing, reference skills, efficiency

Background
Reference skills are important for learners of any language. Among all types of reference tools, such as dictionaries, thesaurus, maps, almanacs, dictionaries are the mostly used reference books at all levels because of their convenience in use and immediacy of result. However, with emerging language corpora directly used in teaching and learning, the role of dictionary to learners, especially to advanced learners may start changing. In the last two decades a large number of scholarly books and research articles have been published on the application of corpus techniques in semantic studies, in lexicography, in grammar, in discourse analysis, and in textbook compilation. However, little is known about how far this new reference tool has reached students as end users, how learners think of corpus use and whether corpus really makes difference in academic writing.

Dictionary as a reference tool
Many reports in the recent literature have indicated that dictionary and corpus are significant reference tools in linguistic studies. Dictionaries are considered as a conventional reference book (Nesi, 2014), while corpora are emerging resource-pool that language users could discover appropriate language patterns in context (Granger 2003). There have been many user studies portraying student lexicographic practice (Nesi 2014; Tono 2012; East 2008; Li 1999) of how learners consult dictionaries to meet their productive or receptive needs in using a language.
Lexicography has a long history of user study. To meet different needs of users and to improve the reference tool, user studies have been tremendously conducted by linguists and
lexicographers in tandem with the development of dictionaries (Wiegand 1998). According to Hartmann, dictionary use and user studies have dealt with four thematic categories: dictionary typology, user typology, needs analysis, and skills analysis (Hartmann 1989: 105). The studies have the following functions and try to address related questions:

1. Research into the information categories presented in dictionaries: "What is in a dictionary?"
2. Research into specific dictionary user groups: "Who needs dictionaries?"
3. Research into the contexts of dictionary use: "What is a dictionary for?"
4. Research into dictionary look-up strategies: "How do you find what you need?"

(Hartmann 1987: 123)

In a recent article, Nesi reported the research into dictionary influence on language learning, and language learners' consultation behavior (Nesi 2014). Unquestionably, lexicographical studies have switched focus from dictionaries per se to the educational use of dictionaries, and this has been verified by some lexicographers (e.g. Laufer & Levitzky Aviad 2006; Ronald 2002). Ronald (2002), by testing 300 words from a reading test, proved the assumption that a monolingual dictionary is beneficial to a learner’s vocabulary growth. He found that students who more frequently consulted dictionaries remembered more words than students who had limited consultations. In another experiment by Laufer and Levitzky-Aviad (2006), participants were required to conduct Hebrew-English translation with target words retrieved from four dictionaries and complete a questionnaire. The results suggested that L1-L2-L2 dictionary was suitable to help writing tasks, because the most prevalent consulting pattern was “translation+ definition+ example”. Students were found to prefer computerized bilingual dictionaries to paper-based equivalents. In the perspective of how dictionary are consulted, Lew (2010), Nesi and Tan (2011) noticed the shortcuts in the interface of on-line dictionaries and figured out that the signposts were related to more accurate consultations. Nesi and Tan added that log-files could reveal the consultation process but not effectively construe the reasons for consultation decisions. Likewise, Tono (2012) employed eye-tracking technology to collect information about the electronic dictionary consultation process. Entry-initial items proved to be more useful than entry-internal signposts for learners with low English proficiency. Monolingual and bilingual entries were equally comprehensible. Generally speaking, although studies into the dictionary use only have several decades of research history, they have been conducted from different aspects, and seem to have reached, as claimed by Nesi (2014), a mature stage with a structured framework. Given that both dictionary and corpus can be used for reference, many research techniques in dictionary user studies can be applied to corpus user studies.

Many scholars and educators in the last twenty years have advocated direct corpus use in teaching and learning. Corpus, as claimed by Biber (1996), entails abundant information that could give a worthy complementary perspective to the existing teaching material and pedagogy. Dictionaries and course books derived from corpora offer learners with significant information in terms of lexical use and collocation. Corpus activities in classroom can contribute more in lexical studies, changing the design of teaching from deductive to inductive manners. John affirms that the “language learner is also, essentially, a research worker who’s learning needs to be driven by access to linguistic data” (1991: 2). Students in such a classroom should work with concordances and consult corpora in an exploratory way (John 2002). Despite all these comments and practices, it is surprising that very limited scholarly attention has been given to students’ opinion, strategies and problems in using corpus as a reference tool.
The role of corpus in education

The development of English corpora, together with the fast advance of computer technology, has inspired corpora to offer new perspectives for teachers ‘to see phenomena that previously remained obscure because of the limitation of our vintage points’ (Kenny 2001: xiii). John Sinclair (2003) believes that natural language use constitutes the best source of linguistic evidence, and such use can only be found in authentic communicative texts. Since its inception, corpus linguistics has transformed from a highly technical apparatus for lexicography and syntax to a handy, native-informant-like reference tool for both teaching and learning purposes. In the last few decades corpus has made remarkable contributions to higher education: The first is generating Academic Word Lists (Coxhead 1997; Gardner and Davies 2014), to genre-specific phrases and lexical bundles (Shin and Nation 2008; Vlach and Ellis 2010; Martinez and Schmitt, 2012). Corpus has also been used as a knowledge pool that provides primary resource to inform students of stylistic features frequently occurring in certain generic discourse. Corpus linguists have devised a tagging system to code and generate move structure. Thompson (2000) programmed the coding that would distinguish integral from non-integral citations in PhD dissertation. Flowerdew and Dudley-Evans (2002) identified moves in editorial letters to international journals and extracted interpersonal strategies in different rhetorical segments. More importantly, Upton and Connor (2009) presented a seven-step corpus-based approach for discourse analysis (2009), which started from annotation and analysis of individual texts, to figuring out typical organizational patterns, and then generalized them to all texts across the corpus.

Besides extracting linguistic information from corpora for teaching materials, educators have also probed into corpus-assisted language pedagogy, which Johns (1991) called Data Driven Learning (DDL), both in classroom teaching and for student self-learning. DDL is a corpus approach to language learning based on the assumption that the use of authentic language together with a concordancer will enable learners to gain insights into the language used in real-life situations. What is more, the process of DDL relies on the learner’s discovery of rules and patterns of language use, not on the teacher’s explanation and provision of examples of the language invented for pedagogical purposes. In the same vein, Römer (2006) made some suggestions to the DDL, including more relevant corpora, more user-friendly learning software and more connections between teaching practitioners and researchers. Römer (2009) claimed that corpus was a reliable reference tool to help solve teachers’ language problems, but before employment, teachers should be trained in corpus use by corpus linguistics. Pérez-Paredes (2010) and Boulton (2010) have contributed to the methodological development of DDL, such as, building a tagging system which is easily customized and manipulated by teachers, so that a corpus would become teaching materials (Pérez-Paredes 2010). Furthermore, integrating DDL elements into coursebook and other online materials could make the teaching approach more accessible (Boulton 2010). In addition to recommending some corpus-based activities to in-service teachers, Bennet (2010) also provided a framework for activity design. There are seven steps in the framework: 1) asking a research question; 2) determining the register on which your students are focused; 3) selecting a corpus suitable to the register; 4) using a concordancing program for quantitative analysis; 5) engaging in qualitative analysis; 6) creating exercises for students, and 7) engaging students in a whole-language activity.

Many scholars and educators in the last decades have advocated direct corpus use in teaching and learning. Dictionaries and course books derived from corpora offer learners with significant information in terms of lexical use and collocation. Corpus activities in classroom can contribute more in lexical studies, changing the design of teaching from deductive to inductive manners. As Park observed “a great number of studies have noted the potential benefits of using corpora to facilitate language learning” (Park 2012: 363), in which “language learner is also, essentially, a research worker who’s learning needs to be driven by access to linguistic data” (John 1991: 2). Students in such a classroom should work with concordances and consult corpora in an exploratory way (John 2002). Despite all these
comments and practices, most of studies report the corpus use from teacher’s perspective, that is, how teachers could produce instructional materials or design activities. Real learners are rarely involved.

Although very limited scholarly attention has been given to students’ opinion, strategies and problems in corpus use, a couple of field studies were found in literature exploring student corpus application. The researchers attempted to find out whether corpus consultation could enhance students’ writing performance, what student awareness of lexicogrammatical challenges and problem-solving abilities are. They generally covered four recurring themes similar to dictionary user research (Nasi 2014):

A English language learners’ preferences and attitudes regarding corpus use;
B The influence of corpus on English language learners’ text production;
C The role of corpora as an aid to English language learning, and
D English language learners’ corpus consultation behaviour.

The following table illustrates five corpus user studies found in literature:

<table>
<thead>
<tr>
<th>Investigators (s) and year</th>
<th>Participant</th>
<th>Method and findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun, Y. C. (2000)</td>
<td>Taiwanese university students</td>
<td>Questionnaire survey. The study revealed that most students were positive towards corpus approach. Web-based concordancing was especially useful to acquire knowledge about authentic use of vocabulary and phrases, but was least effective in understanding grammatical rules and improving writing proficiency. *The analysis of large number of concordance outputs was problematic to students.</td>
</tr>
<tr>
<td>Gaskell, D., &amp; Cobb, T. (2004)</td>
<td>Chinese university students</td>
<td>*An experiment with concordance lines and follow-up interview. *Reading concordances could help good learners to reduce and correct errors in writing. For slow learners, concordance identification was more straightforward in lexical development than in grammatical development. *Learners were willing to improve their grammar with corpus.</td>
</tr>
<tr>
<td>Yoon, H., &amp; Hirvela, A. (2004)</td>
<td>East Asian undergraduates</td>
<td>*A writing task plus questionnaire and interview. *The project examined the EFL students’ corpus use behavior in the completion of a writing assignment, and their perception about the corpora as a second language writing tool. *Corpus activities were beneficial to improve L2 writing, in terms of learning the common use of words, and building up confidence. *Students needed proper guide to use the corpus tools.</td>
</tr>
<tr>
<td>Park, K.</td>
<td>EFL undergraduates</td>
<td>*Screen-recording L2 learners’ actual interaction with corpus in</td>
</tr>
</tbody>
</table>
The studies can be regarded as pioneers in corpus user research, but the methodology and findings are rather rudimentary, leaving a research gap for a more comprehensive study. The current study therefore has the following research questions (RQs):

RQ1: What are the current reference tools used by research students?
RQ2: What are socio-psychological and psycholinguistic factors concerning corpus use by research students?
RQ3: What are the benefits and limitations of corpus perceived by research students?
RQ4: Is there a need to teach corpus in different disciplines at the higher education level?
RQ5: What are student reference behavior when consulting dictionaries and corpora?

The project seeks learners’ opinions and observes performance in corpus use not from a scholarly point of view, but from a student-user perspectives. I attempt to fill the gap between theoretical and practical corpus linguistics, particularly how student users communicate with corpus. However, only the first stage will be reported: student opinions. The study is also pedagogically significant in that it may help us to understand the legitimate role of corpus in language study. By understanding students’ needs, motivation and behavior in corpus use, we will be able to adapt corpus curricula to train students to become competent corpus users, which will not only benefit their ongoing study but also benefit their life-long learning.

**Research methodology**

A corpus workshop was run for PhD students across the campus of a university in Hong Kong with the hope to introduce this new reference tool to students with a high level of learning autonomy. The workshop was initiated not for this research but for helping PhD students. The two-hour session includes 1) introduction to the features and functions of different English corpora; 2) explanation and illustration of Sinclair’s five co-selection search methods, namely core, collocation, colligation, semantic prosody and semantic preference, 3) encouragement of more strategic use of corpus, and 4) hands-on experience of using corpus. After the tutor’s brief introduction to corpus types and some free online sources the participants searched the Corpus of Research Articles (CRA, [http://rcpce.engl.polyu.edu.hk/RACorpus/default.htm](http://rcpce.engl.polyu.edu.hk/RACorpus/default.htm)) which was compiled by the Department of English of Hong Kong Polytechnic University and
can be searched free on the internet. The students then moved on to the lab worksheets with practical tasks which had to be completed by reading concordance lines. Some questions are as follows:

- What can you say about evidence? (adjective + evidence)
- What can you do with a conclusion? (verb + conclusion)
- What can you do with an assumption? (verb + assumption)
- What can findings do? (findings + v)
- Can synonym utterly and completely be used interchangeably? Why not? (semantic prosody)
- What is the difference between commit and commit to? (colligation)

After the training session, student opinions were collected by two instruments: questionnaire and interview. Questionnaires are a commonly used data-gathering technique in the social sciences and many dictionary user studies used questionnaire surveys because they allow the researcher to sample a relatively large population, eliciting quantitative data (Gillham 2000). In this project the survey was used to seek student’s opinion on both dictionary and corpus use. To ensure content validity, the design of the items were discussed with colleagues who are specialized in dictionary and corpus teaching. References to published research (Chen 2010; Daskaloska 2013; Donesch-Jezo 2011; Li 2005; Liu and Jiang 2009) in the related field were also consulted to make sure that all the items are directly related to the research questions. To obtain a profile of corpus use in academic writing by postgraduate students the questionnaire had three sections:

1) Demographic information of informants, including gender, department, mother tongue and educational background.
2) Sociopsychological variables:
   - General reference sources
   - Awareness of corpora
   - Attitude towards corpora:
   - Purpose of corpus use
   - Training in corpus
   - Expectations in corpus
3) Psycholinguistic variables:
   - Motivation and interest
   - Problems in use

As stated before, the corpus training was the main purpose of the workshop and the survey was a by-product to collect student feedback. The sample size is therefore small. 70 PhD students who completed the two two-hour corpus workshops answered the questions of opinion, motivation, experience and problems in consulting corpus. The data gathered from the survey were then analysed by different statistical models to demonstrate opinion or behavior patterns of a sample group. The follow-up interviews were conducted two months after the corpus workshop to see whether corpora were actually used by PhD students in their thesis and/or journal article writing and to find out their ability to handle corpus information and their satisfaction of use. Seven students from different departments were interviewed.

Findings and discussions

This section discusses the results of the questionnaire survey and follow-up interviews. The frequencies and percentages of students’ responses were mainly used to present the statistical results of the survey. Demographic information indicates that the 70 PhD students were from 19 department including: computer science; electronic engineering; nursing; hotel management; linguistic studies and so forth. As Asia’s international city, Hong Kong
universities house many international students. Our respondents are from a variety of linguistic backgrounds: 55.7% are from Mainland China; 25.7% are Hong Kong local students; 5.7% from English-speaking countries and the rest from other parts of the world. 38 are male and 32 are female. The interviewees are from 7 departments; three are art students and four science students.

**Awareness of corpus**

Question 1 and 2 of the survey tested student knowledge or awareness of English corpus. Although language corpus has had a history for over 30 years, many scholars in other disciplines may not be aware of this reference tool for language use. Of the 70 participants, 64.7% had never heard about corpus. The 24.5% participants claiming to have used corpus in academic writing were mostly from the Faculty of Humanities doing linguistic research either in English or in Chinese. Corpus was not perplexing to students from the Department of Computer Science; they may have done a large amount of work in data retrieval both from English corpora and Chinese corpora, but they did not seem to be aware how data can help themselves directly as a reference tool in academic writing. In the interview a computer student told us his corpus experience:

_We treat corpus as data, not as a tool. We are more interested in information retrieval, data mining, machine learning, and probably statistical patterns. The work may be for other institutions such as the government and media institutions. It is amazing to see how text mining can be presented in a way directly help me myself in academic writing._

(Y2 Computer student)

**Profile of student reference tools**

Question 3 creates a profile of reference tools used by PhD students. As a long tradition and stipulated by law, English is the medium of education at universities in Hong Kong. PhD students are supposed to have a high level of English proficiency when they are admitted to the university. Yet, as many non-native English (NNE) speakers admit, the productive use of English words is a big challenge to them. To write better English especially for publication purposes, reference tools are always consulted. What are the reference tools by this advanced learner group? Figure 1 provides the answer.

![Figure 1: Students' preferred reference tools when they had lexical problems](image)

Figure 1 Students’ preferred reference tools before attending the corpus workshop.

Figure 1 shows five common reference tools used by PhD students in their academic study. Similar to the findings by other scholars (Jian et al. 2009; Li 2005; Lee 2008; Tseng 2009), the number of students using online dictionaries significantly outnumbered others. 60 out of
the 70 respondents reported that online dictionaries are their first choice to seek help when solving lexical problems. However in the interview, students admitted that online dictionaries are used for word meaning only. The limitation of online dictionaries, especially their weakness for productive use of words may have driven some students to other reference sources. The second choice of reference tools of the PhD group is Google Search. 51 (72.9%) students admitted they use Google Search very often for lexical help. With the wide coverage of Wi-Fi or 3G in society, particularly in education community, the internet is enjoyed by almost everyone, therefore it is no doubt that online resources are regarded more convenient than the conventional printed dictionaries, which has only 17% users in this group. It is not surprising to see 8.6% students use corpus for lexical problems; majority of the group had never heard about it before the workshop.

In the post-training interview, we asked interviewers to re-rank the reference tools. Among the seven interviewees five thought online dictionaries are still the most popular reference tools. Corpus is now regarded more useful than Google search:

‘Although I was used to conducting Google Search for everything: from terminology to collocation and examples, corpus search seems more relevant to me. You can find the word you wish to use from the concordance lines quickly’.

(Y2 Tourism student)

I use dictionary apps for word meaning. They are in fact online dictionaries, aren’t they? As a foreign student, you have to use dictionaries. However, I use them critically, or just for reference, because they normally give wrong information in my area. Reading corpus can help me understand words in context.

(Y3 BRE student)

Google search generates too many items, and many hits are just hyperlinks to company websites not relevant to the question I have in mind. Finding useful information may require reading a few pages and you may get distracted if something interesting hit your eye.

(Y1 EE student)

Corpus search, especially the CRA search is much faster and relevant in getting the information, compared to Google Search. I wish you had run this corpus workshop earlier in my Year One PhD study.

(Y3 ABCT student)

The students’ opinion is consistent with Davis observation: ‘Although Google (and other search engines) are useful for finding web pages on a particular topic, they are much more limited in terms of linguistically oriented searches. Indeed, it is challenging to look for anything beyond simple words and phrases’ (Davis 2012: 264).

**Perceived benefits of corpus**

Question 5 illustrates the students’ perception of the benefits of corpus, collected immediately after their first experience of working corpus data. As expected, collocation, colligation and language contexts were immediately recognized as advantageous and useful. Figure 2 illustrates the students’ judgement.
This finding is consistent with a number of studies (e.g. Aston 2001; Biber, Conrad and Reppen 1998; Conrad 2000; Liu and Jiang 2009; Odlin 1994; Yoon 2004), which believe that the collocational and colligational patterns provided in the concordance output and the integration of grammar and vocabulary i.e. lexicogrammar are helpful for language learners. According to Halliday (1994), language is a network with syntagmatic (what should go with what / colligation) and paradigmatic (what should go instead of what / collocation) relations. Language learners need to face different choices in their learning process. What corpus provides to learners is a wide range of authentic target language discourse (Yoon 2004), “some of which may be more memorable than those in the dictionary” (Aston 2001: 13), so that people can know the different choices that either go with or without a certain word. In a study carried out by Liu & Jiang (2009), 69.7% of the respondents agree that corpus can help language learners improve their command of lexicogrammatical rules and sensitivity to the context in which a word should be used. Such context-based or genre-based corpus study is found to be helpful for students to identify “common collocational frameworks within particular genres of communication” (Yoon 2004: 259).

In addition, by looking at the concordance, or KWIC, students can learn the appropriate use of a word, and improve what Hymes (1972) called “communicative competence” which involves not only linguistic appropriateness (the right choice of a word), but also cultural appropriateness (the context in which a particular word can be used). In other words, corpus can help teachers explain to students “Why something that is grammatically acceptable just doesn’t sound right” (Conrad 2000: 554), which is especially useful for students to distinguish between synonyms (Biber, Conrad and Reppen 1998; Liu and Jiang 2009).

As mentioned before, 75% of the participants never used corpus before. At the workshop, they learned and tasted Sinclair’s five co-selections in corpus study: collocation, colligation, semantic prosody and semantic preference. After the practice, 89.6% participants regard corpus is a useful tool for them in using the English language. A number of students regarded the workshop very rewarding and wished they had known the reference tool at the beginning of their PhD study; most participants were at the third year of their study.

Reliability and validity of the questionnaire were examined. The relationship between dependant variables (gender, department, language background, educational level) and independent variables (survey responses to corpus, dictionary,) were also examined by correlation analysis. Grouped by the demographic elements which strongly influence the survey results, ANOVA was conducted to check significance of differences among various items.
**Challenges in using corpus**

To understand the potential obstacles which may prevent PhD students from using corpus, Question 6 listed six items relating to the challenges of using corpus. As shown in Figure 3, more than half of the students felt somewhat frustrated when they had to analyze the data and work out themselves the language rules of using a word, which will not be provided by corpus, and may be approachable in dictionaries. The challenge of self-analyzing is also found in the study by Liu and Jiang (2009) in which the greatest concern of the respondents was how to use the data effectively and accurately to identify the lexico-grammatical rules. We were a bit surprised to find that 50% PhD respondents did not like making lexical judgement by themselves because they are supposed to be at the top level of independent learning, and should be actively learners.

Similar to the respondents in Thurston and Candlin (1998), who complained about the time and the information load, some students in the present survey also felt daunted by the time required on the processing and analyzing data (32.9%) as well as the large number of examples generated from corpus (30%).

![Figure 3. Perceived limitations of using corpus](image)

As PhD students are advanced learners less than one fifth of the respondents regarded unfamiliar words as a problem. One limitation of the corpus, as reported by an interviewee, is that corpus does not categorize data according to “specific journal or conference” reports. In other words, the search function of corpus has space for improvement.

**Determination to use corpus**

Despite of the limitations and the first contact with corpus, nearly all the students (69 out of 70) regard corpus as useful and every student claimed s/he will use corpus in their thesis writing. The seven PhD students we interviewed two months after the training informed us they have been using CRA and other corpora we suggested after the training, the frequency is at least once a week. The student who was writing the first draft of his thesis claimed using corpus every day. Some students also suggested using corpus to their NNE supervisors, not only for language but also for knowledge organization and transfer.

**Training is highly needed**

In many dictionary user studies, quite a percentage of respondents had doubts about the necessity of dictionary training in education; people seemed to take the ability of using dictionary for granted. In this study training of using corpus as a new reference tool was regarded a highly necessary; all the students except one agreed with this proposition. The lack of training was also regarded as one of the reasons that students think corpus is challenging (Liu and Jiang 2009).
The study has some limitations and implications. First the sample size was rather small. With the intention to understand the corpus training result, only 70 PhD students participated in the survey, therefore we found it difficult to describe differences caused by gender, linguistic background, year of PhD study and subject area. With the project moving on more corpus training programmes will be conducted and more survey results will be obtained. Only with comparable number of participants in each category can we confidently describe a full picture of corpus use by advanced language users. Second, the survey failed to detect what kind of lexical aids are especially useful through the use of corpus (e.g. synonyms, antonyms, learning of a new word, etc). Some studies show that students are particularly sensitive to the usefulness of corpus for learning nearly synonyms (Liu and Jiang 2009).

**Conclusion**

This paper reports PhD student opinion, motivation, experience and problems in corpus exploitation in contrast to other reference tools. The survey and interview identified the potential benefits that would facilitate language learners’ choice of using corpus as a reference tool, and the limitations that might impede corpus use. The practice was investigated not from a scholarly perspective but from student perspectives. The findings show that although various online dictionaries are available they are mainly used for receptive use for finding out word meanings in reading. Google search is regarded convenient and is mostly used for productive use but its overwhelming amount of information is often irrelevant. PhD students need corpus; they regard the application of corpus offers direct help in academic writing better than any other reference tools. Nearly all the research students claimed they would use corpus, no matter what linguistic background they have and which subject area they are pursuing. It is also evident that the use of corpus should be initiated and trained by teaching staff because corpus exploitation is a completely new learning practice and majority of students may not be aware of such resources. By understanding students’ needs, motivation and behavior in corpus use, we should be able to adapt our curricula to train students to become competent corpus users, which can not only benefit their ongoing study but also benefit their life-long language learning. The results can also shed light on the design of corpus training materials for advanced language users.

In addition to understand student opinions on corpus use, further research will be conducted into student performance analysis in corpus exploitation by using screen capture programmes. We assume that students’ critical thinking and analytical skills, and other aspects addressed in the questionnaire survey could be measured by the frequency of complex transactions captured during corpus consultation.

**Acknowledgments**

This work was supported by the Dean’s Reserve of Faculty of Humanities, Hong Kong Polytechnic University (No. 1-ZVDQ)

**References**


Using Dictionary for Learning English: a Case Study of Non-English Major Students at People’s Public Security University of China

LI Xiu-ping, WU Huang-gang
People’s Public Security University of China, China
Email: lixp9999@126.com; matt19880909@sina.com

Abstract
Dictionary use plays a pivotal part in language learning. For lexicographers, greater knowledge of dictionary use makes for better dictionary compilation; for publishers, observing users’ consultation behaviors helps them to better serve their business; for teachers, learning about learners’ dictionary use helps them to increase the success of learning English with the use of a dictionary resource. This study focuses on non-English major students’ using the dictionary for learning English at People’s Public Security University of China (PPSUC). The research is conducted from both a macro perspective and a micro perspective. A macro analysis shows what dictionary the subjects own, what dictionary they prefer when learning English, the purposes of dictionary use, etc. to help the author have a full grasp of the situation at PPSUC. The micro component of the present study is based on tests, in particular reading tests and word tests. Those tests further an understanding of the issues such as how the subjects use a dictionary to master a new word, how much they learn about a new word, what they are most interested in about each word and so forth. A detailed study on using dictionary not only helps facilitate English teaching, but contributes to lexicographers’ compilation based on English learners’ feedback.

Keywords: English Learning, dictionary use, lexicography, non-English major

1. Introduction
Dictionary is not only a reference book but also a record of the vocabulary of a language (Jackson 2002). Any language learner, whether one is a native or foreign language learner, may turn to dictionary for academic help when encountering some lexical problems (Swanepoel 2003). User’s ability to utilize a dictionary partially affects the success of tackling those lexical problems because of their devoid of lexical knowledge.

Dictionary use is commonly treated as an individual activity (Atkins and Varantola 2002). Atkins and Varantola (2002) try to answer some basic questions about dictionary use through a detailed description of 103 subjects’ look-up process (in the paper, these users are divided into two groups: the Oxford group and the Tampere group). A language user, who is called learner by Jackson (2002) may refer to a dictionary in two broad types of language tasks: one where they solve the totally unknown words in reading and listening; the other where they find the known words’ specific usages in contexts in writing and speaking. Atkins and Rundell (2008) points out that there are two types of user research (market research and academic research). A clear idea of user profiling and user research contributes to dictionary compiling and design. It is significant to know about how users consult dictionary and what information they may get in a dictionary. East (2008) calculates the lexical frequency profiles to figure out whether dictionary use influences the lexical sophistication of test takers.

Chinese scholars (Zhao 2004; Lang and Li 2003; Meng and Su 2003) have many analyses of dictionary users. Zhao (2004) investigates the strategies non-English majors use for learning English, discovering that many students look up words via online dictionary through their mobile phones. Online dictionary, with the popularity of smartphones, has been increasingly replacing the traditional paper dictionary and has become the main complementary learning tool for college students in English learning. Lang and Li (2003) have a macro investigation, such as how often users use a dictionary, when users use a dictionary and what dictionary they prefer to purchase and the like to show the case of English-major students’ dictionary
using at Beijing International Studies University and Beijing University of Technology. Meng and Su (2003) argue that more information about types of dictionary and usages of dictionary should be given by teachers. They always ignore these points or even never mention any information about dictionary. And dictionary’s compiling has to fully reflect users’ academic demand (the richness of a dictionary, explanation functions, etc.) or practical demand (convenience and price, etc.)

2. Methodology

The survey is conducted through two ways, namely questionnaire and test. A total number of 300 college students are randomly taken as the subjects of the survey. 300 subjects including 140 freshmen and 160 sophomores, major in criminal investigation, criminal technology and science and cyberspace security. Questions are divided into two parts: multiple choices and general questions. In addition, 30 students are randomly chosen from sophomores to have a test. There are five stages in the course of taking a test. First, they are given five words without any context and are required to write as much as possible the linguistic information they have learned or known about each word without any help from their dictionaries. Second, they will read part of an article including the previous words and guess each word’s specific meaning in the given context, also without any dictionaries. Third, any kind of dictionary is permitted in order to help them have a better understanding of these words they are partially familiar or completely unfamiliar with and they can record what they get from dictionaries they refer to on the sheet. Fourth, they are asked to rewrite these words’ linguistic information and give the words’ meanings in changed contexts after referring to dictionaries. Last, they will see a passage and underline the words they do not know and write the words they may look up in their dictionaries to help their reading.

3. Data collection

The data in the paper is collected from questionnaires and students’ performances in their word tests. In this survey, 300 questionnaires are sent out while 283 valid ones are used for analysis, including 134 from freshmen and 149 from sophomores. In the test, 30 test papers are all available for data analysis.

4. Results and discussions

4.1 Dictionary ownership

In terms of paper dictionary ownership, among the total number of the subjects including freshmen and sophomores, 48.1% have at least one dictionary but 51.9% have no dictionary. 50.7% of the 134 freshmen own their dictionaries, but 45.6% of the 149 sophomores have one. The ownership percentage of the freshmen is a bit higher than that of the sophomores. Supporting Yong and Peng (2007), the statistics show that the percentage of dictionary owning may be affected by the frequency of using dictionaries by the subjects. In the questionnaire, 77.7% do not frequently use dictionary, and only 8.5% frequently use dictionary in learning practice. However, in terms of the number of dictionary students own, Table 1 shows details.

<table>
<thead>
<tr>
<th>Number of dictionary</th>
<th>One dictionary</th>
<th>Two dictionaries</th>
<th>Three dictionaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of ownership</td>
<td>93.3%</td>
<td>4.4%</td>
<td>2.3%</td>
</tr>
</tbody>
</table>

Among the subjects who own paper dictionaries, 93.3% of the subjects who own dictionaries have only one dictionary, 4.4% have two dictionaries, and 2.3% have three dictionaries. The situation at PPSUC is an opposite picture of that in Yong and Peng (2007). They state that all of the subjects claim to have at least one dictionary and 68% of the subjects own at least three
dictionary but only 3% have one dictionary. The striking difference may lie in that subjects in (Yong and Peng 2007) are university students of English majors for whom dictionary is a must for learning English. Besides, different from the case before 2007, students, especially non-English major students more rely on online dictionary instead of paper or electronic dictionary.

When it comes to types of paper dictionary ownership, more details are presented in Table 2.

<table>
<thead>
<tr>
<th>Type of dictionary</th>
<th>Percentage ownership</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxford Advanced Learner’s English-Chinese Dictionary</td>
<td>63.2%</td>
</tr>
<tr>
<td>English-Chinese dictionary</td>
<td>32.4%</td>
</tr>
<tr>
<td>Longman Dictionary of English</td>
<td>8.8%</td>
</tr>
<tr>
<td>other dictionaries</td>
<td>3.7%</td>
</tr>
</tbody>
</table>

63.2% of the subjects own Oxford Advanced Learner’s English-Chinese Dictionary, 32.4% own English-Chinese Dictionary (generally by Chinese publishers), 8.8% own Longman Dictionary of English, and 3.7% own other dictionaries. (Here, note that the percentage given is based on whoever owns the type of a certain dictionary, so a certain degree of overlap exists.)

Particularly, 55.9% of the freshmen owning dictionaries have Oxford Advanced Learner’s English-Chinese Dictionaries, and 77.9% of the sophomores have one; freshmen prefer to own English-Chinese Dictionary compared with sophomores, for 42.6% freshmen have English-Chinese Dictionaries while 26.4% sophomores have one; Unfortunately, a small number of freshmen and sophomores choose Longman Dictionary of English (7.4% freshmen have Longman Dictionary of English, lower than 14.7% sophomores). Interestingly, whichever type of dictionary the subject own, bilingual dictionaries are the first choice for them to use in their English learning. Besides, Oxford version is more widely used by university students than other foreign-publishers’ dictionaries.

A question is designed to know whether it is necessary for the subjects to have one specialized dictionary based on the fact that the subjects at PPSUC major in the field of policing. The result is that 95.1% have an obvious demand for a specialized dictionary to facilitate their understanding of those terms in their major references or readings or for their communication with foreign experts or scholars on some specialized topics. Actually, the characteristics of specialized dictionary (Bowker 2003) are the main basis for their preference since general dictionary may not effectively solve their perplexing points they come across when doing reading or speaking.

4.2 Dictionary preference

Some students still choose other more convenient tools to solve their linguistic problems even though they may have paper dictionaries. In this section, the subjects were required to choose which dictionary (paper dictionary, electronic dictionary and online dictionary) they preferred to consult. Table 3 shows the percentage of the subjects who prefer which type of dictionary.

<table>
<thead>
<tr>
<th>Type of dictionary</th>
<th>Online dictionary</th>
<th>Electronic dictionary</th>
<th>Paper dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage preference</td>
<td>43.1%</td>
<td>37.5%</td>
<td>19.2%</td>
</tr>
</tbody>
</table>

43.1% of the subjects regard online dictionary as their best choice. 37.5% choose electronic dictionary and 19.2% paper dictionary. 91.9% of the subjects preferring to online dictionary are concerned about the convenience of looking up new words. The richness and examples online dictionary shows are another 2 reasons for this choice: 42.3% consider the coverage
online dictionary shows, 30.1% think more examples regarding one word seem to be more important for their choosing online dictionary. However, only 19.1% claim to prefer paper dictionary. This is supported by Zhao (2004) saying that more than 90% of the college students frequently use online dictionary installed in their mobile phones when leaning English. Many of them do not use a paper dictionary, or do not know how to use them. Quite a number of students think that online dictionary can meet their daily demand of English learning. It is true that online dictionary is available for new words users may not find in paper dictionary, which has the characteristics of open-ended and “under construction” (Cristina Gelpi 2007). Its easy access and being dynamic are also edges compared with printed dictionary.

Among the 109 subjects who like electronic dictionary, 86.2% think convenient look-up is the primary reason why they like using electronic dictionary, which has the same case as those like using online dictionary. Of course, it is also critical to conveniently carry for 55% electronic dictionary lovers. Surprisingly, 63.6% of the subjects preferring to paper dictionary use it due to their habit of using paper dictionary. 49.1% owe their preference to the content it contains. 18.1% subjects point out that their teachers recommend them some brand of paper dictionary.

A further investigation shows that only 2.5% of the 283 subjects prefer English-English dictionary (E-E) and 11% like Chinese-English dictionary (C-E), and 86.5% like English-Chinese dictionary. This point is supported by Yong and Peng (2007) that 80% subjects have a preference for bilingual dictionary even though the subjects are English majors. Table 4 shows the reason why they don’t use E-E dictionary.

Table 4 Percentage of disliking E-E dictionary

<table>
<thead>
<tr>
<th>Reasons of disliking</th>
<th>Percentage of disliking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-consuming</td>
<td>47.8%</td>
</tr>
<tr>
<td>Failing to understand</td>
<td>44%</td>
</tr>
<tr>
<td>No using habit</td>
<td>8.2%</td>
</tr>
</tbody>
</table>

8.2% of the subjects claim that they have no habit of using this kind of dictionary. 44% claim to fail to understand its English explanations very well. 47.8% think that referring to E-E dictionary is time-consuming for mastering a word and beyond it.

Apart from the reasons listed above, it cannot be denied that quite a number of students have an intermediate or relatively low level of English. In general, the percentage of ownership and the frequency of using dictionary at PPSUC are statistically low.

4.3 Purposes of dictionary use
What do dictionary users use for? Table 5 shows some details.

Table 5 Purposes of dictionary use

<table>
<thead>
<tr>
<th>Purposes of dictionary use</th>
<th>Reading</th>
<th>Translation</th>
<th>Writing</th>
<th>Speaking</th>
<th>Listening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of use purposes</td>
<td>77.7%</td>
<td>68.6%</td>
<td>55.5%</td>
<td>19.1%</td>
<td>14.1%</td>
</tr>
</tbody>
</table>

The survey suggests that 77.7% use dictionary to match their need to read, followed by that 68.6% use it for translations. 55.5% subjects may depend on dictionaries for their writing tasks. When they speak and listen, there are fewer chances to refer to it. Only 19.1% will do that when they speak and 14.1% will do when they listen. What is discovered supports Bogaards (2003) that dictionaries mainly serve users’ reading tasks. For bilingual users, translations seem urge users to open the dictionary more than writings.
When it comes to how to use paper dictionary, 72.1% of the subjects do not spend time on directions and appendices. They say these parts do not affect what they are looking for. Even though several editions of a certain kind of dictionary have come out, students seem not to care about which words have been taken in the dictionary or how many new words there are in the part of appendices. 67.9% of the freshmen and 79.9% of the sophomores do not consult the new-intake words.

It is one of the main concerns for lexicographers how users look words up in the dictionary. What information a word harbors is their focus? This can be seen in Table 6.

<table>
<thead>
<tr>
<th>Looking up information</th>
<th>Meanings</th>
<th>Spelling</th>
<th>Collocations, idioms, etc.</th>
<th>Grammatical knowledge</th>
<th>Pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage</td>
<td>57.1%</td>
<td>37.3%</td>
<td>24.8%</td>
<td>20.2%</td>
<td>20.2%</td>
</tr>
</tbody>
</table>

In this survey, 57.1% get their direct access to semantic meanings and 37.3% want to know about a word’s spelling system. 24.8% consult some pragmatic knowledge such as some collocations, idioms and so forth. Following the above are its pronunciation and grammatical information. Only 20.2% of the subjects consider its pronunciation when they look up a word in a dictionary and 20.2% concentrate on its grammatical information, such as its word class. This may be in a sense determined by what they use dictionary for. Fons Moerdijk (2003) puts that seeking for information about meanings is the main reason why people want to consult dictionary. Yong and Peng (2007) report that 77% of the subjects in the research use dictionary for meanings, less for pronunciation and spellings and the least for usage. When they encounter some new words when reading, it is a priority for them to know the unknown words’ meanings (Bogaards 2003).

4.4 Effectiveness of dictionary use

The word test mainly examines how effectively the selected subjects consult their dictionaries when reading. The first stage is designed to have a clear picture of how much they master the words before they consult the dictionary. The words are “answer”, “blunt”, “prompt”, “profile” and “rebel”. Three degrees (completely-mastering, partially-mastering and non-mastering) illustrate their command of the mentioned words. Table 7 respectively shows the degree of mastering linguistically.

<table>
<thead>
<tr>
<th>Words</th>
<th>Answer/ blunt/ prompt/ profile/ rebel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree</td>
<td>Completely mastering</td>
</tr>
<tr>
<td>Aspects</td>
<td></td>
</tr>
<tr>
<td>Pronunciation</td>
<td>40%/73.3%/66.7%</td>
</tr>
<tr>
<td></td>
<td>46.7%/43.3%</td>
</tr>
<tr>
<td>Word class</td>
<td>53.3%/3.3%/3.3%/6.7%</td>
</tr>
<tr>
<td></td>
<td>3.3%/53.3%</td>
</tr>
<tr>
<td>Meanings</td>
<td>0%/0%/0%/0%</td>
</tr>
<tr>
<td></td>
<td>0%/3.3%</td>
</tr>
</tbody>
</table>

73.3% know the correct pronunciation of the word “blunt”, followed by the word “prompt”. 66.7% of the subjects are clear about its pronunciation. The percentages of mastering “profile” and “rebel” are a little higher than that of “answer”, respectively 46.7% and 43.3%. Surprisingly, only 40% of the tested subjects can choose which item has the correct phonetic transcription is correct. That may be resulted from a
lack of systematic learning of phonetic knowledge and a lack of strong awareness of figuring out how they are pronounced. For non-English majors, no courses about phonetic knowledge are given at PPSUC. It sometimes to a large degree depends on a teacher’s teaching schedules. Some teachers may incorporate some phonetic knowledge into their routine teaching activities. For word classes of these words, 53.3% are completely clear about word classes of the word which may have in all contexts. 70% partially know about the case of “blunt”. Half of the subjects can partially master the other words’ word classes. Compared with these words’ pronunciation and grammars, their knowledge of words’ meanings is extremely little. Only 10% partially or totally know the other words’ meanings except for the word “answer” they are largely familiar with. Further, the words’ meanings are confined to those frequently used by the subjects. It may reflect that some subjects are actually unfamiliar with the words’ pronunciation, only playing a guessing game.

Table 8 presents the accuracy of each word’s meaning they can give out in the second stage where they are required to write what each word means in specific contexts before consulting a dictionary.

Table 8 Meanings of the tested words before consulting a dictionary

<table>
<thead>
<tr>
<th>Tested words</th>
<th>Answer</th>
<th>Blunt</th>
<th>Prompt</th>
<th>Profile</th>
<th>Rebel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>33.3%</td>
<td>60%</td>
<td>16.7%</td>
<td>66.7%</td>
<td>33.3%</td>
</tr>
</tbody>
</table>

More than 60% can write down the correct meanings of “blunt” and “profile”. 33.3% can give the correct meanings of “answer” and “rebel”. Only 16.7% can write down the correct meaning of “prompt”. After knowing their degrees of mastering, they are permitted to learn more about these words by looking up their dictionaries. According to their test feedback, it is obvious that semantic meanings come first especially when they read some materials compared with phonetic, pragmatic or grammatical information even though they are supposed to master. 73.7% amongst the subjects only give the semantic meanings, and 26.7% write down grammatical information apart from semantic information. Only one subject apparently write out these words’ phonetic transcriptions but that doesn’t mean others are not clear about how these words are pronounced.

The third stage shows a picture how much they know after referring to their own dictionary, compared with that before using a dictionary. The details are shown in Table 9 after consulting a dictionary.

Table 9 Degree of mastering each word linguistically after dictionary use

<table>
<thead>
<tr>
<th>Words</th>
<th>Degree</th>
<th>Aspects</th>
<th>Completely mastering</th>
<th>Partially mastering</th>
<th>Non-mastering</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pronunciation</td>
<td>43.3%/76.7%/66.7%</td>
<td>0%/0%/33.7%</td>
<td>56.7%/23.3%/33.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Word class</td>
<td>76.7%/16.7%/6.7%</td>
<td>23.3%/80%/90%</td>
<td>0%/3.3%/3.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Meanings</td>
<td>3.3%/10%/3.3%</td>
<td>96.7%/86.7%/73.3%</td>
<td>0%/3.3%/23.3%</td>
</tr>
</tbody>
</table>

90% or more of the subjects have a better command of the words’ word classes. Generally speaking, their performance in phonetic knowledge is a little better than the first stage, but their knowledge of the word “rebel” is worse than previous one. It to some extent proves the previous statement that some of the subjects actually have no idea of how it is pronounced but they choose the right item occasionally. More than 90% of the subjects can give out at least one meaning of the words, which has a striking distinction compared with the case in the first stage. Among them, 96.7% give out the meanings of “blunt” and “rebel” contextually. 93.3%
write what the word “profile” means. Nonetheless, according to statistical analysis, 70% of the subjects only write down one semantic meaning each word has, and it’s just what it means in the article. They use dictionary just for instant satisfaction (Bogaards 2003) not for long-term satisfaction. This can be seen from the difference between the second stage and the fourth.

They are already familiar with the tested meanings, but when required to explain what the five words mean in changed contexts, the picture is different in Table 10.

Table 10 Meanings of the tested words in changed contexts

<table>
<thead>
<tr>
<th>Changed context</th>
<th>Answer (for)</th>
<th>Blunt</th>
<th>Prompt</th>
<th>Profile</th>
<th>Rebel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>3.3%</td>
<td>50%</td>
<td>86.7%</td>
<td>60%</td>
<td>70%</td>
</tr>
</tbody>
</table>

Only 1 subject’s meaning is completely accurate and 26.7% of the subjects have comparatively relevant explanations with regard to “answer for”, a collocation of the tested word “answer”. Thus, it is obvious that they always ignore words’ pragmatic information in which they do not care about whether it has idioms or collocations for specific meanings. 50% of the subjects grasp the word “blunt” contextually; 86.7% of the subjects are clear about what the word “prompt” means and 60% and 70% can guess the meanings of the words “profile” and “rebel” in respective. But the case is a little worse than that in the 3rd stage. The statistics prove that the subjects may be concerned about what a word means in a certain context instead of having a full knowledge of each word. Their mastering a word gradually varies from different contexts. It also illustrates that the effectiveness of dictionary use is relatively low, only limited to what they are exposed to. Except semantic meaning, they have a weak awareness of actively capturing other linguistic information.

After a clear picture of how the subjects look up a word in a dictionary, here arises another question: what words do they have to look up in a dictionary? In the fifth stage, the subjects are told to underline any new word they may encounter and write down which words they have the intention to consult their dictionaries. Among the subjects, only 46.7% underlined the proper names of persons or places, such as “Alejandro Gonzalez Inarritu” and “Leicester”, “Loughborough” and so forth apart from some lexical words like “ubiquitous”, “hilarious” and “pundit”. But when deciding which words they intend to look up, these proper names did not appear in their word lists no matter how much the subject’s vocabulary is. The reason why they do not look up these words is that these terms are not that important for their comprehension of the text even though they may not correctly pronounce them. The findings also prove that categories of words may influence users’ looking-up behaviors (Bogaards 2003).

4.5 Satisfaction of dictionary use

Because of their simple motivation of dictionary use-mainly seeking for words’ contextual meanings, they always have a high level of satisfaction of dictionary use. 90.1% of the 283 subjects feel satisfied with what they have got from a dictionary. However, the subjects who are not satisfied with their searching list some reasons, such as “a small coverage”, “unclear explanations”, “no updating” and the like.

5. Conclusion

The survey mainly investigates the case of non-English major students’ dictionary use at PPSUC. It is a tendency that students will consult online dictionary more than paper dictionary. Features like “convenience” and “coverage” obviously affects students’ dictionary preference. At the same time, a specialized dictionary is of high demand for their major study.
Finishing reading tasks is the strongest motivation for students to consult a dictionary, compared with translation, writing, speaking and listening practice. When referring to a dictionary, they are always concerned about direct satisfaction, finding the meaning they need most in their reading difficulties not the information such as spelling, pronunciation, grammar and collocations. However, they only try to solve some specific contextual meanings instead of fully mastering a word linguistically. This quick satisfaction results in their short-time concentration on a new word. When reading, any word which does not hinder their general understanding of a certain context will not enter their word lists and further their eagerness to learn a word. From the above analysis, there is a low effectiveness of dictionary use. So when asked whether it is indispensable to teach how to use a dictionary, 84.5% among the 283 subjects feel teaching dictionary use necessary.

References

Shigeru Yamada
Waseda University, Japan
shyamada@waseda.jp

Abstract
EFL dictionaries are difficult to use for novice users. There are several types of material to teach users how to utilize those dictionaries. Among them is a booklet which accompanies a dictionary and is written in the user’s L1. Such guides are useful – accessible and specifically designed for a target user group – especially where there has not been enough training in dictionary use. This paper examines those booklets in the context of the Japanese EFL situation with special reference to the one for "Oxford Wordpower Dictionary (4th ed., 2012). The usage guide is devised to familiarize users with the intermediate dictionary without overtaxing them. A unique feature is in the section dealing with the dictionary use for comprehension. The section breaks down the look-up process into seven steps, based on Hartmann’s model (2001: 89-92), and focuses each step with exercises. Special attention is given to Step 6 “Extracting relevant data,” which constitutes a key to successful consultation, involving correct interpretation of the L2 definition. The paper also considers the possibilities of the booklets in view of their transfer to the Internet.

Keywords: consultation process, dictionary usage guide, EFL dictionary, reference skills, teaching dictionary use

1 Introduction
An EFL dictionary is a useful yet difficult-to-use tool for novice users. Not all of them will find the built-in L2 guide to the use of the dictionary helpful. Where there has not been enough training in the use of dictionaries, an L1 booklet on the use of an EFL dictionary comes in handy. This paper examines such booklets in the context of the Japanese EFL situation with special reference to the one for "Oxford Wordpower Dictionary (4th ed., OWD4, 2012), which I authored.

2 Types of guides available
There are two broad categories of guides to the use of dictionaries: general and dictionary-specific. The latter can be roughly divided into three categories: 1) a booklet sold separately from the dictionary, 2) a free separate booklet (an accompaniment to the dictionary or a promotional material), and 3) a guide incorporated in the dictionary. Fewer and fewer of these print materials have been produced. Some web-based dictionaries have the accompanying onsite guide (e.g., “How to Use” on Longman Dictionary of Contemporary English Online, “Using OALD” on Oxford Advanced Learner’s Dictionary Online) and Longman Dictionary of Contemporary English on DVD (5th ed., 2009) offers a guided tour.

3 The L1 Booklet on the use of an EFL dictionary: the Japanese situation
In Japan, EFL dictionaries are usually sold in two ways: through a local distributor or by a foreign publisher directly. In the former case, the dictionary typically comes in a hard-cover case or a jacket with Japanese explanatory notes and with a user’s guide in Japanese. The booklet is usually written by an academic. While such booklets help to make the dictionary accessible to Japanese learners of English, there are two disadvantages: extra money and time. Recently, usage booklets have been made available online (e.g., the Japanese guide to the use of OALD8 [2010], that to OWD4 [2012]). This shift seems to mediate both problems mentioned above.

I was asked by Obunsha Co., a major Japanese dictionary publisher, to write the Japanese guide to *OWD4*. Obunsha has been distributing Oxford’s EFL dictionaries for the past several years. Anyone (non-purchasers, too) can download the PDF file of the 20-page guide free of charge from the publisher’s site. The guide is intended not only for the users of *OWD4* but also as an introduction to the use of EFL dictionaries, more generally. This is because *OWD*, an intermediate EFL dictionary, can be the first monolingual English dictionary to be used by EFL learners. The contents of the guide are as follows:

1 Introduction  
2 Why to Use EFL Dictionaries?  
3 Characteristics of *OWD4*  
4 How to Use *OWD4*  
5 Abbreviations and Labels Used  
6 Let’s Use *OWD4*  
7 Four Tips  
8 Concluding Remarks

What makes my guide unique is Section 6 “Let’s Use *OWD4*” because the first half of this section dealing with the dictionary use for comprehension is based on the seven components of the consultation process identified by Hartmann (2001: 89-92):

1 Activity problem  
2 Determining problem word  
3 Selecting dictionary  
4 External search (macrostructure)  
5 Internal search (microstructure)  
6 Extracting relevant data  
7 Integrating information

For a successful dictionary consultation, the user is supposed to clear all these steps without fail (though some trial and error may be involved). However, it will be very difficult for the inexperienced learner to go through the whole process with an EFL dictionary (all in L2) on their own. Therefore, I divided the process into manageable bits, explaining each and providing exercises (I treated Steps 6 and 7 sequentially as well). By reading through the subsections and doing the exercises, the student can familiarize themselves with the EFL dictionary consultation process step by step. In the case of an EFL dictionary being used by a non-native learner in particular, Step 6 “Extracting relevant data” can be a key and at the same time a major obstacle to a successful consultation. Unlike the consultation of an L1 or L2-L1 dictionary, the user of a monolingual EFL dictionary has to interpret the L2 texts correctly. The interpretation of definitions can be a hindrance because they may be presented in succinct style (while example sentences are written in natural prose). I paid special attention to this step by 1) covering words frequently occurring in the definition (e.g., “quality,” “state,” “particular”) and by 2) elucidating a definition in complex structure.

5 Concluding remarks

There are a number of approaches to the education on the use of EFL dictionaries: dictionary-specific, user-specific, genre-specific, purpose-specific, media-specific, etc. Being dictionary-specific and user-specific, an accessible L1 booklet is useful. The online version is much more promising because multi-media components can be incorporated.
The circulation of the booklet is limited to a small circle of purchasers of the target linguistic community. However, the transfer of the booklet to the Internet gives the material much more exposure, invites feedback, and makes possible a comparison of similar materials for further improvement. Ultimately, it may lead the way to an internationally concerted effort to compile a universal dictionary usage guide (involving the editor) to be made available to speakers of different languages. Such initiative would meet the needs of learners in various linguistically-specific contexts.

References

1 Dictionaries cited

2 Other works:
Meeting More Readers’ Needs: the YES-CEDICT Chinese Dictionary

Xiaoheng ZHANG
Hong Kong Polytechnic University, China
ctxzhang@polyu.edu.hk

Xiaotong LI
Shenzhen University, China
sharklxt@aliyun.com

Abstract
The YES-CEDICT Chinese Dictionary is designed to meet the extensive needs of international learners. Chinese native students will also find it helpful for learning English learning. Currently, there are 114,300 word entries in the dictionary, covering: all the 5,000 words of the new word list of HSK; all the 8,000 words of the new word list of TOCFL; all 112,178 headwords in CC-CEDICT; and all 13,000 plus characters in the latest version of Xinhua Dictionary and Contemporary Chinese Dictionary. There are 10 fields of information for the word entries: the simplified form of the headword; the traditional form of the headword; Putonghua Pinyin; Cantonese Jyutping; English definition; HSK level; TOCFL level; frequency order; number of strokes; and Unicode. The dictionary is arranged in YES order, which is similar to alphabetic order, if we consider the stroke sequence of a Chinese character as the letter sequence of an English word. To help readers unsure of the standard stroke order of some Chinese characters, a Pinyin-stroke order index is provided for all single-character words in the dictionary.

Keywords: sorting method, Chinese dictionary, CC-CEDICT, HSK, TOCFL

1. Introduction

CC-CEDICT (formerly CEDICT) is a large Chinese-English dictionary downloadable from the Web (MDBG, 2015). There are four fields in a word entry, including the word’s form in simplified Chinese, form in traditional Chinese, Mandarin pronunciation in Pinyin and definition in English. CC-CEDICT is a database text file, which has been the knowledge-base of quite a number of interactive electronic dictionaries. No PDF or print version of the dictionary has been reported.

YES is a Chinese sorting method which is simpler and more effective than the traditional methods (Zhang and Li, 2013).

YES-CEDICT is a PDF version of CC-CEDICT arranged in YES. It allows automatic and traditional lookup on soft and hard copy. The user can have a flexible view of more than one entry. Other enhancements made to the dictionary include:
- Adding more words and characters for Chinese learners,
- Adding Cantonese pronunciation and other information for the words,
- Eight appendices,
- A pinyin index for simplified and traditional Chinese characters,
- Two versions of simplified and traditional Chinese.

The YES-CEDICT Chinese Dictionary (一二三汉英大词典|一二三漢英大詞典) is compiled for international learners. Chinese native students will find it helpful to English learning as well.

2. Coverage of headwords
The headwords of YES-CEDICT covers the following (except a small number of characters not displayable on the computer):

- All the 5,000 words of the new word list of HSK (Hanyu Shuipin Kaoshi), the official Chinese Proficiency Test of China mainland (Hanban, 2012);
- All the 8,000 words of the new word list of TOCFL (Test of Chinese as a Foreign Language), the official Chinese Proficiency Test of Taiwan (National Taiwan Normal University, 2013);
- All 112,178 headwords in CC-CEDICT, the most popular downloadable Chinese dictionary on the Web (MDBG, 2015);
- All 11,408 characters in the Complete List of Chinese Characters Used by the Media in 2013 (2013 年度媒体用字总表, National Language Commission of China, 2014);
- All 13,000 plus characters in the latest version of Xinhua Dictionary (Linguistic Institute of the Chinese Academy of Social Sciences, 2011);
- All 13,000 plus characters in the latest version of Contemporary Chinese Dictionary (Linguistic Institute of the Chinese Academy of Social Sciences, 2012).
- First characters of all headwords in the dictionary.

There are altogether 114,300 word entries (including 15,103 single-character words).

3. Components of an entry

The information of each headword is presented in the format of an entry. There are 10 fields for mono-character words and 8 fields for multi-character words, for example

愛 (愛) ài [ai3] /to love /affection /to be fond of /to like / HSK1 TOCFL 入門級
FO323 U7231(U611B) Stroke(s)10(13) (in S version)
愛 (愛) ài [ai3] /to love /affection /to be fond of /to like / HSK1 TOCFL 入門級
FO323 U611B(U7231) Stroke(s)13(10)  (in T version)
學生 (學生) xué sheng [hok6sang1] /student /schoolchild / HSK1 TOCFL 入門級
FO148
學生 (學生) xué sheng [hok6sang1] /student /schoolchild / HSK1 TOCFL 入門級
FO148

Each entry starts with the simplified form (in S version) or traditional form (in T version) of the headword, followed by the other form in brackets if the two forms are different. The third and fourth fields are Putonghua Pinyin and Cantonese Jyutping in square brackets. In addition to representing sound, Pinyin and Jyutping can be used to input the Chinese word into a computer or to look it up in a dictionary. The fifth field is English definition (mostly from CC-CEDICT). The number of strokes appears at the end of the entry of a single-character word. If the character has different forms in simplified and traditional Chinese, then both stroke numbers are presented. The other fields are marked by capital letters with the following meanings:

HSK: HSK level (1, 2, 3, 4, 5, 6) of a headword which also appears in the HSK word list.

TOCFL: TOCFL level (入門級, 基礎級, 進階級, 高階級, 流利級) of a headword appearing in the TOCFL word list. To make the dictionary more compact, words appearing more than once at different levels in the TOCFL list appears only once in the dictionary with the earliest level. For example, 愛 ài appears twice at 入門級 and 進階級 in the list, and only appears once marked with level 入門級 in the dictionary.


U: Unicode for simplified and traditional characters. You can type the hexadecimal Unicode on MS Word and press Alt+X to input the Chinese character to the computer.

At the present state, all word entries in the dictionary have contents for Simplified Chinese, Traditional Chinese and Putonghua Pinyin. About 79% of the words have been annotated with Cantonese Jyutping, and 99% of them have English definitions. The raw data of Jyutping came from the dictionary of the Allbanced Chinese Input System.
4. Appendices

At the end of the dictionary, there are eight appendices,

1. Standard (New) and Variant (Old) Forms of Chinese Characters,
2. Rules of Stroke Order,
3. Stroke Order of Chinese Character Components,
4. Components with Different Stroke Orders between Mainland and Taiwan,
5. The Pinyin Scheme,
6. The Jyutping Scheme,
7. Province-level Divisions of China,
8. Pinyin-Stroke Order Index.

Entries in the main text of this dictionary are arranged in YES order according to the standard font and stroke order of the Mainland. Appendix 1 to 4 will improve the student’s command of the standard font and stroke order.

The student can also find the standard stroke order of a Chinese character by the Pinyin-stroke order index. This index includes all the 17,181 single-character words of simplified and traditional Chinese in the dictionary (including the cross-reference characters introduced in the next section). The characters are firstly sorted by Pinyin alphabetically, then by number of strokes ascendingly, and finally by stroke order in YES. To the right of a Chinese character, you can find its standard stroke order, which can be used to look up a word starting with that character in the dictionary. For convenience of lookup, a character with more than one form and/or pronunciation appears for each form and sound. For instance, there are four entries for character 长/長 in simplified and traditional forms, including 長/長, 長 (長), 長 and 長 (長).

5. Two versions

There are two versions for this dictionary: Version S and Version T. Version S is sorted by the simplified form of the headwords. To enable lookup by traditional Chinese as well, traditional characters different from their simplified counterparts are added to the dictionary and sorted as independent entries, pointing to the simplified characters and words beginning with them.

For example, it is easy to lookup a simplified Chinese word beginning with character 长 in version S because word entries are sorted by simplified Chinese in this version. It is also possible to find the entries by traditional Chinese 長 here. First we look up cross-reference entry “(長) See 長” by the form of 長. Then we follow the guidance to find the full entries beginning with simplified Chinese 長:

長 (長) cháng [coeng4] /length /long /forever /always /constantly /HSK3 TOCFL 入門級 FO1368 U957F(U9577) Stroke(s)4(8)
長 (長) zhǎng [zoeng2] /chief /head /elder /to grow /to develop /to increase /to enhance /HSK3 TOCFL 進階級 FO1533 U957F(U9577) Stroke(s)4(8)
长三角 (長三角) chángsānjiǎo [coeng4saam1gok3] /Yangtze River Delta /abbr. for 長江三角洲/长江三角洲 /
6. The YES Chinese sorting method

Arranging Chinese in YES order is similar to arranging English in alphabetic order, if we consider the stroke sequence of a Chinese character as the letter sequence of an English word. Two Chinese characters are sorted by their first stroke positions in the stroke alphabet. If the first strokes are the same, then check the second strokes, and so on. For example, the different characters in 一二三排检法|一二三排検法 (the YES Sorting Method) are sorted as:

一 (一)
二 (二一)
三 (三一)
検 (検)
排 (排)
法 (法)

In the rare cases of more than one glyph or stroke order for a Chinese character, YES follows the standards issued by the National Language Commission of China (1997, 1999, 2013).

Words of multiple characters are sorted by their first characters in YES order. If the first characters are the same, then check the second characters, and so on. Non-Chinese characters appear after Chinese characters in alphabetical/Unicode order. For example,

覚
覚醒
覚
覚醒
覚悟
B超
T恤.

The stroke alphabet is based on the Standard Bending Strokes of GB13000.1 Character Set (National Language Commission of China, 2001) and the Unicode CJK Strokes (The Unicode Consortium, 2014). There are totally 30 strokes, sorted by the standard basic strokes order of “横 (一) 提 (-) 竖 ( | ) 撇 (丿) 点 (丶) 掠 (亅)” and bending points order of “折 弯 钩”.

The Chinese name of the Sorting Method, i.e. 一二三，are the first three Chinese characters in YES order. The English name YES is the abbreviation of Pinyin “Yi Er San” of the Chinese name.

<table>
<thead>
<tr>
<th>Stroke</th>
<th>Stroke Name</th>
<th>Example Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>横</td>
<td>十/七</td>
<td></td>
</tr>
<tr>
<td>横折竖</td>
<td>亅 达 贯/鼓 為</td>
<td></td>
</tr>
</tbody>
</table>
### 7. Conclusion

According to our survey, YES-CEDICT Chinese Dictionary is one of the seven Chinese-English dictionaries with over 100,000 word entries. The others are CC CEDICT (MDBG, 1997 -), New Age Chinese-English Dictionary (Wu and Cheng, 2001), ABC Chinese-English Comprehensive Dictionary (DeFrancis, 2003), A New Century Chinese-English Dictionary (Hui, 2003), The Chinese-English Dictionary (Wu, 2010) and the Oxford Chinese Dictionary (Kleeman and Yu, 2010). Among these big seven, only YES-CEDICT provides information of Cantonese pronunciation, frequency of use, levels of HSK and TOCFL, Unicode and stroke numbers.

As for further improvement, there are still over 20,000 words not annotated with Cantonese Jyutping, and about 1000 words still lack English definition. In addition, example sentences should be added to the entries.
Acknowledgments
The project was partially supported by a University grant (A-PK14). And we are grateful to Dr. Cheung Kwan-hin of the Hong Kong Polytechnic University and Dr. Lun Suen of Hong Kong City University for their valuable advice.

References

Lexicographic Needs in Turkic Republics

Baktygul KULAMSHAEVA
Specialist in Modern Turkic Languages and Literatures, Turkey
b.kulamshaeva@gmail.com
b.kulamshaeva@nevsehir.edu.tr

Abstract
After the collapse of Soviet Union, Turkic communities, who share a common history, gain their independence. This paved the way to many changes and developments in those communities. Language studies also gained speed after the attainment of independence. The fact of having their mother tongue as an official language in Kyrgyzstan and Kazakhstan, being in the first place, and many other Turkic Republics gave rise to need for number of new Turkic words. This, as much as varying according to the republics, led to the revival of old Turkic words or derivation of new words. On the other side, socio-cultural and sociolinguistic changes within the Turkic societies brought to "communication" problems in their relations with outer world. Socio-political, economic, educational and trade relations between Turkic Republics, gaining speed after the independence, gave Turkic languages prominence and created a need for elementary language course books and bilingual dictionaries of these languages, as well.

In this regard, many bilingual dictionaries of Modern Turkic Languages, especially in Turkish as being in the first place, have been prepared and presented to the learners. However, these dictionaries, being a significant step in communication, still lack of lexicographic works, i.e. they are not based on modern corpus database or any statistical study. Thus they contain many archaic, or non-frequent words and out of date words. Moreover, for many of them are translation dictionaries from the ready-made bilingual dictionaries prepared in Russian, there have many mistranslate and wrong usage words in them. For example, in the Kyrgyz-Turkish dictionary you may notice that translations of the collocations were directly affected by Russian or Turkish collocations and as a result of this, new collocations, that does not really exist in Kyrgyz, were invented.

This study will analyze the present state of dictionaries in Modern Turkic Languages, and try to find common solutions to prepare dictionaries based on lexicographic studies. Thus, we hope that this study may be an aswer to the lexicographic needs around the Turkic World.

Keywords: lexicography, corpus database, bilingual dictionaries, collocations, Turkic Languages

1. Introduction

Dictionaries have a very important role in language teaching and research since they are one of the sources that display the vocabulary of concerned language. Maintaining the words which make a society a nation, they “are valuable works reflecting nations’ cultural level and their way of life” (Ilhan, 2009: 535). At the same time “they are like treasures for languages that protect cultural heritage and are transmitted to the future generations” (p. 535). Although Turkic Languages which are originated from the same language family share much common vocabulary, there is still a need for bilingual or multilingual dictionaries between dialects of Turkic Languages.

Turkic communities who nowadays are spread over quite a large area and speak common language had started to become distant by the socio-economic and political developments. As time went by there appeared many changes in their vocabulary as a result of language relations and interactions, thus many of these dialects, in particular the ones which are territorially apart from each other, as Cumakunova (2009) states had experienced such changes that it is impossible to communicate without the help of dictionaries. Bilingual dictionaries prepared in these languages will provide the opportunity to determine the magnitude of those differences and identify the common features in those languages.

Here, it would be helpful to refer to the historical development of the previous dictionary studies.
2. History of lexicographic studies in Turkic communities

The efforts and struggle of Turkic Communities to become independent republics succeeded quite late. Before Soviet Union besides some prior Turkic governments, most of Turkic communities had been living in small tribes in Central Asia and Siberia. It can be said that Soviet Union played a great role in the establishment of these societies as governments. The expansionist policy of Soviet Union dates back to Tsarist Russia.

Russians started to be interested in Turkic territories after they conquer Kazan Khanate in 1552 and Astrahan Khanate in 1556. "The interest of Russians towards Turks and Turkic languages, had increased in the period of Tsar Peter as much, depending on political developments. Establishment of Academy of Sciences by Tsar Peter I (1724-1725), enabled Oriental Sciences and in particular Turkology to be executed in accordance with tsarist policy (Kirişçioglu, 2004: 38). Especially since XVIII. Century when the influence of tsardom on Central Asia had been increased lexicographic studies began to develop. Therefore as Bayniyazovs (2014) noted, bilingual dictionaries had emerged in order to familiarize themselves with local people and to carry out the missionary activities.

According to Suleyman Kaan Yalcin (2008), Russians with the aim of Russification and Christianization tried to divide many nations, especially the Turks, the most populous region after Russia, into micro-nations and to marginalize them. For this case, language was the most important weapon; because the language is the prior tool in identity creation (p. 666). Professor Orientalist Ilminskiy, has been a key person in the creation of a separate languages and separate alphabets. Names such as V.V. Grigoryev, D.A. Tolstoy, N.I. Ilminskiy (Remnev, 2012: 17-19) tried to entitle the statuses of separate languages to Turkic Languages based on minor phonetic and dialectic differences in those dialects.

Important works published in this process are: “Comparative Dictionary of all Languages and Dialects” (1790-91), prepared by P. S. Pallas (1711-1811); “Tatar-Turkish Comparative Dictionary” (1869) by L.Z. Budagov; “Turkish-Tatar-Russian Dictionary” (1864) by L. Lazarev; “Chagatay –Turkish Dictionary” (1868) by Velyaminov-Zernov; “Trial of Turkic Dialects Dictionary (Opt Slovary Tyurkskih Narechiy)” published by Radloff in 1893-1911; “Dictionary of Altay – Aladag Dialects of Turkic Language” (1884) by I. Verbitsky; “Russian-Kyrgyz (Kazakh) and Kyrgyz-Russian Dictionary” (1883) by Ishmuhammed Bukin; “Kyrgyz-Russian Short Dictionary” (1895) and “Kyrgyz-Russian Dictionary” (1897) by V.V. Katarinsky and “Russian- Turkmen Dictionary” (1913) by I.A. Belyaev. The common point in these dictionaries, is that generally in each case they were prepared based on Russian.

According to Bayniyazov (2014: 86), Abdullayev (2007: 79) and others, lexicography in Turkic communities, followed the Arabic-Farsi-Turkish tradition before the Russian Invasion, but Russian Invasion broke all connections with the tradition of Arabic-Farsi and Turkish lexicographic studies. Therefore from the 18th century up to now most of the dictionaries were “prepared usually to meet the needs of Tsarist Russia and its language policy” (Bayniyazov, 2014: 86). Dictionaries published in this era, prepared in “Russian -Turkic Language” format, usually in one way, and more focusing on Russian.

By XX. Century the separation policy founded by Tsarist Russia succeeded its goal with the help of Soviet Union. They created separate alphabets for each of the Turkic Communities and even the languages of minorities were acknowledged as separate languages. In this way, Turkish communities were prevented the opportunity to join in a body in order to avoid possible threat against Soviet Union and they were marginalized as much as possible.

Thus, as Cumakunova (2009) noted that the languages of the Turkic communities were narrowed, their usage were limited and pushed into background. Moreover the relationship between these languages was loosened. Thereby all these factors hindered the development of these languages. As a result of having separate alphabets, different orthographies and different terms, these facts made their literary language notably distinct from each other’s and thus they had to communicate in Russian (p. 4-5). In addition, the process of separation as separate `languages` reflected in their vocabulary so much that
instead of using common words they tried to create new words or using foreign versions as they are. Russian bilingual dictionaries played a great role during this processes. Cumakunova (2009: 5) thinks that Turkic communities hold a negligent attitude towards lexicography and regretfully informs that while there are many bilingual dictionaries of Turkic Languages in Russian, German, French and etc. for some reason they do not show this enthusiastic desire in developing a common dictionary among Turkish communities. One of the most important reasons for this is thought to be derived from the fact that the common communication language among these languages is Russian, although all of these Turkic Languages are just a dialects of Turkic.

As a result, we can still see the influence of Russian in today’s lexicographic studies of Turkic Republics.


After the collapse of Soviet Union, gaining of independence of Turkic Republics which shares a common history has added a substantial speed to the developments within the language as well. Kyrgyzstan and Kazakhstan being in the first place, becoming of other Turkic Republic’s native language as official languages gave rise to the need for many new Turkic words especially in social and political fields. This need, although it shows differences depending on the communities, has been tried to be solved either by revitalization of the Turkic words used formerly or by way of derivation of new words in general. During these periods, as it was in Ataturk’s period, studies for decontamination of the language from foreign words have been conducted, for example, Month names in the old Kyrgyz and Kazakh have been brought in to use as below.

<table>
<thead>
<tr>
<th>Turkish</th>
<th>Kazakh</th>
<th>Kyrgyz</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ocak</td>
<td>Қантар (Qantar)</td>
<td>Учтун айы (Ucubuntu ayi)</td>
<td>January</td>
</tr>
<tr>
<td>Şubat</td>
<td>Акпан (Akpam)</td>
<td>Бирдин айы (Birdin ayi)</td>
<td>February</td>
</tr>
<tr>
<td>Mart</td>
<td>Наурыз (Navrız)</td>
<td>Жалган Куртан (Cgalgan Kuran)</td>
<td>March</td>
</tr>
<tr>
<td>Nisan</td>
<td>Көкек (Kökkek)</td>
<td>Чын Куртан (Çın Kuran)</td>
<td>April</td>
</tr>
<tr>
<td>Mayis</td>
<td>Мамыр (Mamır)</td>
<td>Бугу (Bugu)</td>
<td>May</td>
</tr>
<tr>
<td>Haziran</td>
<td>Мусым (Mavsim)</td>
<td>Кулжа (Kulca)</td>
<td>June</td>
</tr>
<tr>
<td>Temmuz</td>
<td>Шілде (Şilde)</td>
<td>Теке (Teke)</td>
<td>July</td>
</tr>
<tr>
<td>Ağustos</td>
<td>Тамыз (Tamız)</td>
<td>Баш Оона (Baş Oona)</td>
<td>August</td>
</tr>
<tr>
<td>Eylül</td>
<td>Кыркүйек (Kırküyek)</td>
<td>Аяк Оона (Ayak Oona)</td>
<td>September</td>
</tr>
<tr>
<td>Ekim</td>
<td>Казан (Kazan)</td>
<td>Тогузун айы (Toguzun Ayı)</td>
<td>October</td>
</tr>
<tr>
<td>Kasım</td>
<td>Караша (Karasha)</td>
<td>Жетинин айы (Cetinein Ayı)</td>
<td>November</td>
</tr>
<tr>
<td>Aralık</td>
<td>Желтоқсан (Jeltoksan)</td>
<td>Бештин айы (Beştin Ayı)</td>
<td>December</td>
</tr>
</tbody>
</table>

or new words have been derived by taking advantages of internal source of the language.

<table>
<thead>
<tr>
<th>Türkis</th>
<th>Kazakh</th>
<th>Kyrgyz</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>televizyon</td>
<td>теледидар (teledidar)</td>
<td>сыналгы (sinalgı)</td>
<td>television</td>
</tr>
<tr>
<td>radyo</td>
<td>радио (radio)</td>
<td>уналгы (unalgı)</td>
<td>radio</td>
</tr>
<tr>
<td>diktafort</td>
<td>диктатор (diktafon)</td>
<td>üн алгыч (ün algıç)</td>
<td>dictaphone</td>
</tr>
<tr>
<td>ziyalet</td>
<td>зиялет (ziyalı)</td>
<td>зиялы (ziyalı)</td>
<td>intelligent</td>
</tr>
<tr>
<td>cumhuriyett</td>
<td>республика (respublika)</td>
<td>жумурият (cumuriyat)</td>
<td>republic</td>
</tr>
</tbody>
</table>

However, language is a living and natural creature. The practices imposed to the language sometimes have been accepted, sometimes remained as a meta-language or has been rejected completely. The first three of the above examples did not survive in the Kyrgyz Turkish contrary to Kazakh Turkish, and like other 3 words, many words also remains as meta-language today and not embraced by the public completely.
After the independence, Turkey being in the first place, the relations in the field of commerce, economy, social and education field in between the Turkic Republics and other countries have gained great momentum. And this added speed to learning them as a foreign language by causing to the prominence of the Turkic languages, and as a result of it the need for basic level of language learning books and for bilingual dictionaries has arisen. In this context, Turkish Language being in the first place, in other Turkic Languages bilingual dictionaries with many Turkic reciprocal have been prepared and presented to the readers. After the independence, Turkey has pioneered in the studies for dictionaries with Turkish reciprocals. At the beginning this developed as “translated dictionaries”¹. At this point it’s been thought that the border of “translated dictionary” concept should be drawn with precise lines

In Hartman (1998: 146) translated dictionary is defined as “a dictionary which is the result of a translation of another dictionary”. However, especially in Turkic communities since previously they were prepared by translation of Russian part of dictionaries prepared as bilingual with Russian into a third language, it is being thought that this Hartman’s definition will be insufficient to explain the current lexicographical state in Turkic communities. Accordingly, in this study these types of dictionaries will be called as “Third Language Translated Dictionaries (TLTD)”.

In this regard, Kudashev (2007) defends that the “perevodcheskie slovari/translation dictionaries” and “perevodnya slovari/translated dictionaries” terms which are referring to the same concept in Russian should be distinguished from each other (Kudashev, 2007: 7).

Hikmet Koraş (2009: 765), used “(çeviri sözlük) translation dictionary” concept as “Third Language Translated Dictionaries (TLTD)”. And he states that first “originally Non-Third Language Translated Dictionary” was Hayati Develi’s “Azeri Türkçesi Lügati /Azeri Turkish Dictionary” (1993) which is prepared very late.

The first dictionary studies carried out in Turkey related to Turkic languages, as we stated above, are the second translation of bilingual or monolingual dictionaries usually prepared in other languages, in other words they are “Third Language Translated Dictionaries”. This kind of translation dictionaries which were frequently prepared during Ataturk’s period,² as we stated above, has been commonly used method in order to ensure communication between the Turkic Languages in the early years of independence.

The translation of K.K. Yudahin’s dictionary which was first published in the year of 1940 by Abdullah Taymas (1945) is an example of such kind of work. This dictionary which descriptions of several words and phrases were not given due to a number of objective and subjective reasons, and many outdated and obsolete words, phrases and proverbs took part, reworked by Yudahin in 1965, this new edition have been tried to be free from such errors as much as possible, and also it has been enriched by adding many new words. However, unfortunately this 1940 edition which has “missings” has been translated into Turkey’s Turkish by Abdullah Taymas and in the years of 1989, 1994, 1998 and 2011 with some adding this old edition has been published again.

¹ Many translated dictionaries have been published, such as: N. A. Baskakov, vd, Gagauz Türkçesinin Sözlüğü [Dictionary of Gagauz Turkish] (Translated by Abdülmecit Doğru-İsmail Kaynak), KB Yayımları, Ankara 1991, 279 s.; Gyula Nemeth, Kumuk ve Balkar Lehçeleri Sözlüğü [Dictionary of Kumuk and Balkar Dialects] (Translated by Kemal Aytacı), KB Yayımları, Ankara 1990, 60 s. etc.
² Many Turkish dialect dictionaries are translated into Turkey’s Turkish, such as Pekarski’s “Yakut Dili Sözlüğü [Yakut Language Dictionary]” (1945), Yudahin’s “Kırız Sözlüğü [Kyrgyz Dictionary]” (please see Karahan, 1999; Koraş, 2009).
Some of the deficiencies seen in Yudahin’s dictionary translated by Abdullah Taymas (2011 edition):

<table>
<thead>
<tr>
<th>Kyrgyz</th>
<th>Turkish</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Жип/çip (p. 255) * Ala çıptı atta-</td>
<td>İp (rope) * Alaça ip üzerinden atlamak (over reddish rope)</td>
<td>• to cheat; to betray</td>
</tr>
<tr>
<td>Майыла-2/mayla-2 ve</td>
<td>2. (başlica, binek hayvanı) kirali mec. kafese koymak, dolandırmak [2. renting (primarily, metaphorically) to put in a cage, (p. 558)]</td>
<td>These meanings are not in use in Kyrgyz.</td>
</tr>
<tr>
<td>Чанқылда-/çankılda-</td>
<td>1. To bark out (bawl out) kid, woman; 2. barking with bitt (about a dog)</td>
<td>“About” is a miss use; instead “can” be used or nothing can be used</td>
</tr>
</tbody>
</table>

As it is shown in the samples some words which are being used in both Kyrgyz and Turkish parts are not up to date. Only, those words used orally, and does not appeal to large masses, or those outdated, obsolete words included frequently. Since these types of words were also converted into Turkey’s Turkish from Russian, it undermines the credibility of the dictionary and also negatively affects its usability:

<table>
<thead>
<tr>
<th>Kyrgyz</th>
<th>Ottoman Turkish</th>
<th>Turkish</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Эрк (erk)</td>
<td>irade; hürriyet</td>
<td>erk</td>
<td>freedom; independence</td>
</tr>
<tr>
<td>Комолдурак (komoldurak)</td>
<td>sinebent</td>
<td>gömüldürük</td>
<td>breastband</td>
</tr>
<tr>
<td>Буу 2 (buu 2)</td>
<td>2. maden mahlülü</td>
<td>buğu</td>
<td>mineral solution</td>
</tr>
<tr>
<td>Жалаан (calañ)</td>
<td>münhasıran</td>
<td>yalin</td>
<td>exclusively; simple</td>
</tr>
</tbody>
</table>

In the figure above, it can be said that while there are hundreds of, thousands of common words are present in between these two languages, the usage of Ottoman Turkish words instead of using them were due to the translation of the dictionary was made in a period when the influence of the Ottoman Turkish on Turkey’s Turkish not exactly cleansed and Turkisisation has not been fully completed.

Accordingly, as Cumakunova (2009) stated , both the Yudahin’s first edition published in 1940 as well as the subsequent new editions are not the dictionaries which are fully reflecting the practical use of modern Kirghiz and leads to increase the functional capabilities of it. In the dictionary, as Yudahin’s himself stated above, constant word sequences related to lexical entry of contemporary Kyrgyz words and phrasal patterns were not taken part completely. Instead, examples compiled from a variety of epics, reflects the status of Kyrgyz in the historical process rather than current status.

Besides this, similar errors are encountered in the dictionaries which were originally prepared and non-TLTD ones. Selahattin Çankaya’s “Kırgız Sözlüğü [Kırgız Dictionary]” which is a such type of dictionary gave rise to the emergence of collocations and words in Kirghiz which are not present in the language, through preparing collocations and words in Kyrgyz under the influence of mother tongue Turkish or through translating from another languages word for word.

Although this dictionary which was prepared for practical purposed and contains 50,000 words and phrases, is a user-oriented dictionary by giving Kyrgyz words in Latin transcribes next to the Cyrillic alphabet, it cannot be said that it is a dictionary prepared in accordance with the methods and standards of lexicology. For example,

a) *Aba* (air) word which was given as lexical entry and phrases where this word is being used: *aba ber* (give/provide air); *aba al* (take air); *aba cut* (swallow air); *aba kirdüülügü* (air pollution); *aba irayı bocomolu* (weather forecast); *aba irayı*
maaǐmatı (weather forecasting report); aba irayı raportu (air report); cañ word and phrases... etc. (These would be given under the “aba” lexical entry).

b) The words and phrases which are not present in Kırğız Turkish and being created through translation word by word from Turkey’s Turkish or from another language: aba maydanı (air field); aba mekendik (air space); cañ kelin (new fashion); cañ çeri (janissary); cañ çeri komandiri (janissary commander); cañ çeri korpusu (janissary corps) etc.

c) Denotative and connotative meanings of the words, polysemy or homonym distinctions of them were not performed; as homonym words were given under the same entry and polysemes were given under different entries: aarçuu-I (clear, transparent, pure, clean); aarçuu-II (bleaching, cleaning)* these two words are the gerund and infinitive form of the verb given after them aarç-; Also the verb aarç-; aarç- I; aarç- II or aarç- III; kooz I, kooz II, does not have kooz III usage in Kırğız Turkish.

d) Some wrong translation took place: abjora* (*in Russian objora – greedy, gourmand) (smart, resourceful, cunning) etc.

e) Phonetic doublets (doublet) are also included as a separate lexicon entries frequently: aygayla- (shout, scream); aykayla- (shout, scream); aylager (deceitful, cunning); aylaker (deceitful, cunning); aylampa (whirlpool, vorticity, whirlwind, vortex, circulation); aylanpa (whirlpool, whirlwind, swirl) etc.

Apart from these, inconsistencies in punctuation, too many orthographic errors, incorrect translations or taking place of frequent wrong phrases, and Inconsistencies in the dictionary’s macro- and micro-structure undermines the credibility of this dictionary.

It can be thought that the common shortfalls which can be found in general at dictionaries prepared by a single author rather than not being a collective work prepared by dictionary editors who are native speakers of both languages, can be caused by author’s lack of full knowledge of words or collocations of that language although its dominance on that language or due to preparing of the dictionary by translation and most importantly, due to inadequacy of the theoretical work in the field of Turkish lexicography and lack of database.

Other Turkic languages just as Turkish and Kyrıgz, although they are in the same language family and their basic vocabularies are same, the differences can be found in word phrases and collocations. For example, the phrase “adamın dışkı” (man’s feces)” given as “adamdın tıskısı” in S. Çankaya’s dictionary, however such use is not available in Kyrıgz language, because, the word “tıskı” does not have meaning of “feces, excrement”; likewise the phrase “bazi şartlar altında (under certain conditions)” in Turkey’s Turkish does not has a usage just as “ayrim şarttardın aldında” in Kyrıgz language.

As we stated above, phonetic differences in all Turkish languages, although they are same externally the phrasal differences are present: For example, the verb doğmak (rise/born): TT. Doğ; Kır. tuu-; Kazakh tuu-. Uzb. tug- etc. TT, with collocations Güneş doğdu (Sun has risen); Kyrıgz kın (not tuudu) çekti/köterildi; Kazakh kın şktı/köterildi; Uzbek Kın/kuyuș çıktı/kütarildı. In the usage of the doğur- verb which is the causative form of the same verb differences are available in Kyrıgz Turkish. The verb tuu- has a synonym in the form of törö- in Kyrıgz Turkish, so while the tuu- verb has a common usage against both human and animal and other situations, the usage area of the törö- verb so narrow and only can be used for humans. Because of that the phrases in the form of “it törödü”, “koy törödü” for animals do not exist.

4. Conclusion
Today when the relations in between Turkic Republics and Turkic communities have developed ambidextrously, the preparation of bilingual and multilingual dictionaries for various purposes and forms is much needed. As Cumakunova (2006: 24) rightfully stated, due to needs such as opening of Contemporary Turkic Language and Literature Departments in Turkey’s universities especially, and opening schools and universities in Turkic Republics where the education and training language is Turkish, training staff for these educational institutions, preparation of scientific and pedagogical publications, the need for such dictionaries is increasing with each passing day.

Bilingual dictionaries should be prepared by a group of the specialists and native speakers in these languages (Kolukisa 2015: 159). Otherwise, these types of errors in bilingual dictionaries are inevitable, if they are prepared by only one person.

As a result, the things need to be done in this matter can be summarized as follows:

1. Due to Russian was common language in all Turkic Republics and communities under the Soviet rule, the lexicon entries shall be taken from Russian, and a common database shall be formed where their counterparts and definitions in those languages along with their phrases are present.

2. And Russian words should be a tool that can be used while indexing. Formation of such database would reveal not only common and different points in using the words between these dialects through inter-dialect collocation dictionaries but also would reveal social a cultural commonalities and differences indirectly. The preparation of such will provide a major contribution to the inter-dialect machine translation at later stages (without a translation requiring a third language such as English).

3. First of all, theoretical works on the lexicography shall be made firmly and then it should be put into practice.

4. The dictionaries where common vocabulary of Turkish languages and collocation consisting of basic words are present. Such a study will also demonstrate social-lingual differences of Turkic communities who are living in the same geography and sharing a common historical background by putting these differences into the middle.

Collocation dictionaries represents a major sources for those who are learning these languages with regard to provide information about the use of words in that language, to contain important information about in which words it is coupled with and in which grammatical forms they are being used. In terms of answering to the current needs today, many new dictionaries are being published in Turkic state languages. However, since the theoretical lexicography works in these countries are not fully developed, they are not formed depending on the particular system and unfortunately since many of the studies are not based on concrete data such as database, they are not meeting expectations by reaching to desired quality.

This study has been prepared to share our thoughts on what needs to be done to base dictionary works on concrete scientific data in Turkic state languages which shares the same common geography and having the same historical background, and because of that, along with expectations to open a new era in lexicography works in Turkic states, the feasibility of application of a similar system into different languages also revealed.

References
A Contrastive Study of Phraseological Units among Andalusians and the Fluminense people

Rita Giovana M. Ferraro
University of New South Wales, Australia
rita.ferraro@unsw.edu.au

Abstract
This paper is the result of qualitative field research conducted over 3 years of stay in Andalusia. The original corpus collected contains one thousand expressions spoken in the southern region of Spain. It is often said that the Andalusian people produce units of speech that greatly differ from the ones constructed by Spaniards from other regions. Indeed, this study shows that the block of recurrent expressions used in Andalusia has one or more of its forming elements modified or suppressed in comparison to what is spoken in other parts of Spain. This paper provides examples to illustrate it. However, this unit act which characterises the Andalusians speech goes beyond the boundaries of orality. The collection of speech samples came through 1996 to 1999 and was obtained through recordings of Andalusian people in spontaneous interactions as well as in controlled interviews and questionnaires. In order to validate the recorded phraseological units spoken by a particular generation, diaphasic, diastratic and diatopic criteria were adopted. The research provides support for the incidence of use (identified in the research through asterisks) of phraseological units in Andalusia. The incidence of use does not seem to be dependent on the speakers’ socio cultural status, but rather related to age reasons.

Keywords: lexicography, phraseology, phraseological units, Spanish/Portuguese

1. Introduction
This article concerns the contrastive study of some phraseological units used in speech in Andalusia by the “twenty-something” generation and their equivalent uses in Rio de Janeiro. We chose to use “phraseological units” (PUs) as the general term encompassing expressions, idioms, sayings, figures of speech, set phrases and interjections. We have also included some widely-used vulgarisms, many of which are formed by somatisms, that is, expressions referring to parts of the body, either human or animal.

The corpus that has been compiled includes approximately one thousand expressions in alphabetical order from A to Z and their equivalents in Portuguese. In this article only a few have been included for reasons of space and also because they are currently in press.

This corpus was collected over the three years of my stay in Spain, from 1996 to 1999, in public places such as discos, cafés, restaurants, supermarkets, bus-stops, beaches and friends’ houses. We only gathered examples of speech from Andalusians and from people born in the state of Rio de Janeiro, aged from 19 to 35.

The speech sample collected comes directly from source. All the expressions have been faithfully transcribed respecting the original utterances. Thus we took note of speech acts by acquaintances and strangers in public places as well as the speech of fellow students at the University of Cádiz (UCA) and that of our circle of friends.

We also enlisted the help of those who became regular collaborators, such as the waiters in the restaurant of Cadiz Hospital, as well as shopkeepers and waiting staff of neighbourhood bars. Besides that, we received the assistance of the shop assistants of La Gloria, a chain of cake shops situated in the Old Quarter and in the new part of the city of Cádiz and of the staff
of the Traca Playa restaurant and of the local “veinte duros” shops, who, on countless occasions, generously gave us their help, not only by contributing with their own voices, but also correcting and confirming the units that we had collected, (It should be explained that “veinte duros” shops, which later became 1 euro shops, used to sell extremely cheap goods, similar to “2 dollar shops” in Australia and some other English-speaking countries).

All those involved in the study were from Cadiz, Jerez de la Frontera, Sevilla, Huelva, Jaen, Malaga, Cordoba, Granada and other parts of Andalusia.

Our other reference works were the Diccionario Fraseológico del Español Moderno, by F. Varela and H. Kubarth (1994); El Español Idiomático by P. Domínguez et al (1988); the Diccionario Esencial de la Editorial Santillana (1991); the Diccionario de Uso del Español de María Moliner (1982); the Diccionario Práctico de Locuciones y Frases Hechas, Everest (1998).

Subsequently, in 1999, we took the Spanish corpus to the UCA in order to discover if the phraseological units used by the general public were also used in university circles among people of the same age.

2. Methodology

In the aforementioned university the research was carried out through sociolinguistic surveys aimed at students in the faculties of Arts, Business Studies, Marine Science, Information Technology and Medicine.

The general conditions of eligibility were, among others:

1) People born in Andalusia or those who had been living there for no less than 10 years;
2) People between 19 and 35 years old;

With regard to family incomes, the data obtained came from the following categories:

a) Less than 100,000 pesetas
b) 100,000 to 200,000 pesetas
c) 200,000 to 300,000 pesetas
d) Over 300,000 pesetas

Most of the participants belonged to group c), which represented an average family salary.

As for housing conditions, we observed that most of them lived in modest houses or flats (that is “normal dwellings”). This was ascertained by suggesting the following types:

a) Dwellings without sanitary installations and with difficult access;
b) Modest houses or flats/normal dwellings;
c) Spacious and elegant houses or flats, with all modern conveniences.

The surveys lasted for several months and each student received thirty expressions per session. In order to facilitate the classification of these units, they were typed in bold, numbered and ordered alphabetically. Each expression was followed by its meaning, given in brackets, and examples were given below in inverted commas in one or two sentences taken from our compilation. Participants were also asked to replace those expressions that were unfamiliar or never used with an equivalent. Students were required to indicate the frequency of each expression using asterisks. Thus, unfamiliar or unused expressions should be marked with a single asterisk (*), infrequently used ones with two (**) and those frequently used should receive three asterisks (***)

To illustrate this, we have included part of the survey below. The asterisks and replacement expressions were provided by the students.
1) **ser de boca** (algo)  (to be said insincerely)
   “Ya veo que tus promesas son de boca. Hace tiempo que me prometiste un puchero y hasta ahora nada.”
   (*)

   Replacements:  
   * ser de boquilla  
   * ser mentira  
   * mucho lirili, poco lerele  
   * no valer ni un duro  
   * ser de boca p’a fuera

2) **tener guasa**  (to be funny, often ironically)
   “Tus amigos tienen guasa.”
   (***)

   Replacements:  
   X

Likewise, we took some similar expressions to the Universidad Federal Fluminense (UFF) in Niterói, RJ, Brazil. In the UFF the task was carried out only in the Arts and Business Studies Faculties because final exams were taking place at the time.

We kept the same Spanish expressions for the Brazilian students although we provided the meanings in Portuguese. We translated the expressions as well as giving the semantic meaning in every case and the communicative-pragmatic one when it was considered necessary. The Brazilian students confirmed the use of the items in Portuguese as well as making new contributions which have been added in italics.

3. Entries

Whenever two verbs appear together, the first takes precedence. For example, in the case of “hacerse de rogar” the reader will find the expression under the letter H.

If an expression can be used alone or with one or more accompanying verbs, it will appear under the letter corresponding to the best-known part of the expression. Thus, the expression “de puta madre” is included under the letter D and the accompanying verbs are included in brackets. For example, “(estar) (algo) de puta madre”.

When the verb or verbs form an integral part of the expression in such a way that it is impossible to detach them, then the expression will be classified according to these verbs. Thus, “decir/soltar cuatro frescas” the verbs “decir” and “soltar” will be included as sub-entries.

Square brackets are used to indicate an equivalent expression in Portuguese, together with an illustrative example of use. Should an identical expression appear with a different meaning, it will be pointed out in a note.

In this way, the following results were obtained:

**A**

**(vivir)** **a cuerpo de rey**  (to have an easy, excessively comfortable life)

“Mi vecino dijo que su sueldo no cubre los gastos, sin embargo, vive a cuerpo de rey.”

(*)

Replacements:  
* de puta madre  
* como un marajá  
* como un rajá  
* como un rey  
* como un fraile  
* de escándalo
de muerte

[‘viver como um marajá/ um rei’, ‘ter a vida que pediu a Deus’.]
“Meu vizinho disse que com o seu salário mal dá para sobreviver mas, no entanto, vive como um marajá/ vive como um rei/ tem a vida que pediu a Deus.”]
(…)

B
beber los vientos por alguien  (adore somebody)
“Ramón bebe los vientos por su novia.”

(*)
Replacements /Substitutes:  estar enchochado
                   estar loco-a por alguien
                   estar loquito-a

[‘ser/estar louco-a/doido-a (por alguém)’, ‘ser/estar amarrado-a/ fissurado-a/gamado-a/ vidrado-a (em alguém)’.
“Ramón é/está louco pela noiva”. “Ramón é/está amarrado/fissurado/gamado/vidrado na noiva dele.”]
(…)

C
contra viento y marea  (facing obstacles and difficulties)
“Para que tu novio y tú estéis juntos tendréis que luchar contra viento y marea.”
(***)
[‘contra vento e tempestade’, ‘contra tudo e contra todos’.
“Para que você e o seu noivo fiquem juntos terão que lutar contra vento e tempestade / contra tudo e contra todos.”]
(…)

CH
¡chúpate ésa!  (Exclamation of approval or applause when the speaker answers somebody sharply or wittily: exclamation indicating satisfaction at somebody else’s misfortune)
“¡Chúpate ésa! A mí me han concedido permiso para salir el jueves y a ti te obligan a quedarte en la oficina”.

(*)
Substitutes:     ¡Toma!
                   ¡Toma ya!
                   ¡Cómete esa!
Vulgarism:       ¡Para que te jodas!
[‘escuta/ouve essa ...’, ‘sente/olha só ...’, ‘aqui...’
Note: these expressions are normally followed by the argument or comment.
(…)

4. Criteria used for the organisation of phraseological units

According to some theorists, phraseological units play an important role in the acquisition of both the mother tongue and foreign languages. In the mother tongue, phraseological units, acquired automatically and unconsciously, are used by adults to save words and to facilitate “speed in language processing” (G.Corpas: 1996: 15). Phraseological units, which are so widely used in all languages, likewise facilitate the rapid assimilation of a second language, increasing the student’s possibilities of expression and communicative capacity. However, the formal aspects, such as fixedness and idiomaticity, and the pragmatic content can complicate the learning of these units. Therefore the L2 teacher must consider these aspects when introducing phraseological units and select the units to be worked on according to the
students’ level. This task requires a great deal of time and dedication since there are few studies of this field. (I.Penadés, 1999: 35 – 37). Currently, there is a great choice of didactic materials which, as well as providing the meanings and uses of phraseological units, also add indications of formal or informal register and restrictions on their use. Others offer a wide range of relevant exercises. These materials include the book El Español Idiomatico by P. Dominguez González et al (1988), the Diccionario fraseológico by F. Varela and H. Kubarth (1994), the Diccionario Espasa by A. Buitrago Jiménez (1995); María J. Beltrán and E. Yánez Tortosa (1996) (apud, Penadés). Some authors suggest that the expressions be presented in onomasiological or analogical order (when the expressions are grouped according to their semantic field) in order to simplify the process of locating them. This is the system used in ideological dictionaries and in this way the words are distributed according to a unifying criterion using a base word. The following idioms illustrate this:

<table>
<thead>
<tr>
<th>Entry</th>
<th>ojo/ojo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>“cuestar el ojo de la cara”</td>
</tr>
<tr>
<td></td>
<td>“no pegar ojo”</td>
</tr>
<tr>
<td></td>
<td>“a ojo de buen cubero”</td>
</tr>
<tr>
<td></td>
<td>“entrar/meterse algo por los ojos”</td>
</tr>
<tr>
<td></td>
<td>“en un abrir y cerrar de ojos”</td>
</tr>
<tr>
<td></td>
<td>“mirar con buenos ojos”</td>
</tr>
</tbody>
</table>

The same criterion can be used in the sample of phraseological units starting from a single word to refer to homogeneous groups of notions, such as parts of the body (con la boca abierta; boca arriba; boca abajo, se me hace la boca agua), feelings (to express anger: estar hasta + adjetivo = el moño, el coñ o, los cojones, los huevos, la coronilla, las narices, etc.), military hierarchies, etc.

The choice of an adequate criterion to use in ordering phraseological units is not a simple task. As there are no strict norms regarding which criterion to use, some scholars on the subject choose whichever is most suitable for them. However, problems may arise with the choice of certain criteria when the type of target reader and their possible difficulty in consulting phraseological units are not taken into account. Bearing this in mind, I. Penadés (1999) has presented eight recent dictionaries of Spanish phraseology of which we will use the six that are most useful in our study. In Penadés’ work she contemplates the usefulness of ordering the phraseological units according to the target user: the teacher, the researcher, the student, with or without a great knowledge of grammar, as well as the foreign user.

According to this author, one of the most useful criteria for foreigners and students with no great grammatical knowledge is that which offers the first word of the phraseological unit, followed by the second and so on. Thus, in the case of units like “en un plis plas”, “a rajatabla” or “tomar las de Villadiego”, it would be easiest to find them in in a dictionary under “en”, “a” and “tomar”, that is, under the first element of each unit. This type of ordering can be found in the “Diccionario Espasa de dichos y frases hechas”.

The most usual criterion would seem to be presenting the UFS according to a key word, which is usually a substantive noun. The problem with this is that the reasons for choosing a noun are not clear. In our opinion the choice of an element of the structure as the base form should be made bearing in mind its formal aspect, as occurs with compound nouns (this means that if only the first element of the lexicalised syntagms were considered, idioms would constitute complete entries instead of sub-entries with two base elements). The dictionaries in question are aimed at both teachers and students, although the latter could have difficulties in using them. One of the dictionaries using this criterion is the Larousse.
M. Martín Sánchez (1997, apud Penadés) includes in his dictionary colloquialisms and phraseological units grouped according to lexical fields, such as “The Human Body”, “Religion”, etc. Although some units are outdated, the book can prove very useful.

The RAE dictionary (1992) includes phraseological units in the following order of preference: substantive, verb, adjective, pronoun and adverb.

F. Varela and H. Kubarth (1994) begin their phraseological classification using 2000 key words, although they do not include sayings and proverbs. The words are arranged alphabetically, with examples of use in the following order: proper nouns, substantives, main verbs, pronouns, numbers, and the verb SER (to be).

In this case, we prefer to introduce the expression using the first word, and this element would appear in only one entry in the dictionary, a criterion also suggested by J. Martínez Sousa (1995). This type of classification would save space and time in the compilation of the dictionary which would be aimed at Brazilian students or foreigners with a reasonable knowledge of the language. In this way, there would be no need to include all the elements of the expression in their respective alphabetical categories, as occurs, for example, in the Espasa (although we do understand that this method makes it easy to find any expression). Thus, for an expression like “por si las moscas” the only entry would be under “por”. Exceptions would be made in the case of expressions which could be the complement of two or more verbs. In this case, an expression like “estar/ ponerse/ quedarse boca abajo (algo) (alguien)” would appear under the entries of the respective verbs as separate headings for a single unit.

The book by P. Domínguez González (1988) contains an Índice General de Modismos with their respective lexical definitions and approximately 1200 items. All the expressions mentioned are in general use in the 50 Spanish provinces but there is no index of frequency.

5. Survey

In 1999 we conducted a survey among ten students from Cádiz and ten from Asturias, showing them samples of expressions from the aforementioned book. Of the 244 expressions, only 91 were used in speech by the Asturian students, or were at least familiar to them: 25 were used only by the Andalusian students, 78 were used by both groups, 45 were unfamiliar and five were expressed in a different way.
Although many of the idioms and set phrases are dated or unfamiliar, P. Domínguez’ work is still very valuable in the teaching of phraseological units. The exercises given in the two sections of the book are useful tools to check the understanding and correct use of Spanish idioms.

There is one further problem that should be mentioned regarding the ordering of phraseological units: the arbitrary use of the words *le, uno* and *alguien*. Some units appear in didactic texts and dictionaries as follows: *irsele a alguien el santo al cielo, irse uno con viento fresco, dejarle a alguien plantado.*

We have followed the M. Marin’s suggestion of replacing the words *le* and *uno* by *alguien, a alguien.*

Another possible solution in this case is to use a different font, not italics or bold, bracketing the accompanying pronouns. The results would be as follows: *irse (a alguien) el santo al cielo, irse (alguien) con viento fresco, irse con viento fresco (alguien), dejar (a alguien) plantado.*

6. Conclusion

Andalusians have a different way of speaking from that of Spaniards from other regions. We have observed that in the group of commonly used expressions in Andalusia, when they are formulated in other parts of Spain, some elements may be modified or even omitted. However, we have also discovered that the characteristic unity of speech acts in Andalusia goes beyond orality. Using the samples collected from individuals of 19 – 35 years of age, with secondary or higher education and coming from different social strata, we could observe that the same phraseological units were used throughout the group. Thus, our research has shown that the incidence in the use of phraseological units observed in the “twenty – something” generation of Andalusians does not depend on the socio cultural level of the speakers, but, rather, on age reasons.
References


376
Phraseology and Terminology in LSP: Implications for lexicography

Jirapa Vitayapirak
King Mongkut’s Institute of Technology Ladkrabang, Thailand
kvjirapa@kmitl.ac.th

Abstract
Phraseology is a study of word combination which is one of the important linguistic production. It is also referred to as collocation, fixed expression, multi-word lexical unit, cluster, chunk or lexical bundle. This paper explores the distinctive phraseological patterns which show the meaningful linguistic behavior of the high frequency words using examples from specific corpora at King Mongkut’s Institute of Technology Ladkrabang. Data for the study consist of three electronic corpora, i.e. Food Technology Corpus, Cosmetic Science Abstract Corpus and Environmental Science Research Articles Corpus. An analysis of the co-occurrence of lexical items, concordance data and their frequency exhibit both phraseological and terminological tendency. The forms, structures and functions of lexical bundles show the distinctive variations. Many phrases show highly specific terminological meanings with domain specific implications. The particular examples of terminology were recorded and selected as a raw material for compiling LSP dictionary and teaching English for Specific Purposes.

Keywords: phraseology, terminology, LSP, corpora

1. Introduction

Specialized or technical dictionaries are devoted to the description of the technical language of a specialized subject or discipline. The corpus-based methodology is useful for ESP lexicography. As McEnery and Wilson (1996: 166) and Bergenholtz and Tarp (1995: 178) suggested that a finite lexicon in ESP can be checked using word frequencies searching for patterns of word combinations, and looking for examples of usage of particular words in order to make the information in LSP dictionaries authoritative and backed up by hard evidence. Moreover, an important component of fluent linguistic production in ESP is the multiword expressions referred to as “clusters”, “chunks” or “lexical bundles” (Hyland, 2008). These are collocations which appear more frequently than expected by chance, helping to shape meanings and contributing to our sense of coherence in a text. As Hyland (2008) pointed out that clusters or collocations seem to present considerable challenges to EFL students.

The present paper describes the systematic use of corpus-based information in a technical dictionary designed particularly for LSP students. It explores the distinctive phraseological patterns which show the meaningful linguistic behavior of the high frequency words using examples from specific corpora. They offer insights from an actual LSP situation, and show how future reference tools can be oriented to meet the lexical needs of the LSP learner. It is important that LSP lexicographers should make good use of computer technology in order to make better specialized dictionaries.

2. Corpora

Specific corpora become central to the fulfillment of this research. In order to analyze the phraseological tendency and terminology, the co-occurrence of lexical items derived from concordance data and their frequency can exhibit both of them. Data for the study thus consisted of three electronic corpora at King Mongkut’s Institute of Technology Ladkrabang, i.e. Food Technology Corpus, Cosmetic Science Abstract Corpus and Environmental Science Research Articles Corpus.
Research Articles Corpus. These corpora are specialist corpora writing in English. They were designed and compiled to be specific and to represent the range of subjects in the text types in the target population using the texts written by native speakers. The structures of three corpora are briefly detailed in Table 1 below:

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Food Technology Corpus</th>
<th>Cosmetic Science Abstract Corpus</th>
<th>Environmental Science Research Articles Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (tokens)</td>
<td>2,304,640</td>
<td>625,579</td>
<td>1,883,197</td>
</tr>
<tr>
<td>Text type</td>
<td>Food Technology textbooks (90%) and journal articles (10%)</td>
<td>Cosmetic science abstracts from International Journal of Cosmetic Science</td>
<td>research articles in SciDirect database</td>
</tr>
</tbody>
</table>

The three corpora are specialist corpora of writing in English. The corpora consist of 2,304,640, 625,579, and 1,883,197 tokens of running text respectively. The purpose of the corpora were to collect a well-balanced database of texts. All texts were then selected and entered into the database by re-using existing electronic texts.

3. Methodology

The corpora were compiled and then analysed by a concordance program (WordSmith Tools Version 6). It provided lists of words or word-clusters in texts set out in alphabetical or frequency order, while the concordancer, Concord, offered key-word-in-context (KWIC) concordance displays. Basic statistics such as the number of types, tokens and hapax legomena were also examined.

4. Results and discussion

(1) The Top Twenty Nouns

For LSP dictionary, function words such as the, of, and are disregarded, while high frequency nouns as lexical words become the focus. For example, the top twenty nouns in the food technology corpus are:

- temperature, water, heat, process, product,
- food, rate, liquid, pressure, air, time,
- drying, surface, transfer, moisture,
- quality, equation, data, design, number

Figure 1 The top twenty nouns in the food technology corpus

(2) Technical vs. sub-technical vocabulary

Generally, it might have been predicted that many of the lexical items particular to the specialized corpus would be technical words. However, this is not the case. This finding confirms Béjoint’s (1988: 354) study that scientific and technical discourse makes use of words that are not specialized at all. For example, nouns like temperature, water, heat, process, product are not technical terms, but rather can be classed as sub-technical or semi-technical vocabulary (Yang 1986: 98, Robinson 1991: 28, Flowerdew 2001: 77). In other
words, these words have a special meaning within the food technological area, and are used across many sub-fields.

(3) Phraseology and terminology
A problem in categorization is the fact that such categories of technical vocabulary in LSP texts overlap considerably, due to their use in general language as well as for specific-subject purposes. Examples of this are: temperature, water, heat, process in food technology. The polysemy of sub technical words in different subject domains underscores the need for an objective means of recognizing the semantic functions of terms in text. Pearson (1998: 26) points out that the communicative setting helps to identify meaning and to determine whether an expression will be behaving ‘terminologically’ or ‘normally’. In fact, technical terms often pair up to form larger meaningful groups within the contexts, i.e. compounds or lexical phrases that always appear in the same form. For instance, heat collocates with transfer in heat transfer. These collocations are potentially problematic for EFL readers. Technical terms and collocations can be extracted and identified in context using corpus data in the form of concordances (see Figure 2):

sealer. A metal wire is heated to red heat simultaneously to form a bead straining temperature and convective heat transfer coefficient. (4) Air impingement accomplished with the same type of heat exchangers as used for product comparison with the rate of sensible heat flow by conduction and convective plate with an auxiliary blower, the heat transfer coefficient increased with decreased temperature. If heat removal rate is slow and the product is comparable with the latter as a heat transfer medium. Superheated s

Figure 2 KWIC concordance data for the word ‘heat’

The concordance data for the word ‘heat’ shows its collocation behavior in technical texts in the food technology corpus. It appears in numerous compound and multiword units, which show up in concordance data, e.g. convective heat transfer coefficient, heat exchangers, heat flow, heat removal rate, heat transfer medium, heat-induced flavor. As Sinclair (1995: viii) says, we only notice compounds and multiword units when we see many examples of them set together. Without the corpus we might not have found any of them.

Another example, nouns appear at the high frequency in the corpus of cosmetic science abstracts concerning with either the topic of cosmetic science such as ‘skin’ (409 times), ‘hair’ (343 times). As ‘skin’ occurs at the highest rank in the corpus, the concordance lines of this word were selected to show its circumstance. In terms of adjectives, it is found that ‘cosmetic’ appear at the highest rank (103 times) in the corpus of cosmetic science abstracts. Concordance lines of this word are selected to show its circumstance as illustrated in Figure 3 and 4:

Figure 3 Concordances of ‘skin’ in the Cosmetic Science Abstract Corpus
From Figure 3 and 4, the concordance lines show phraseological terms such as ‘intact axillary skin’ or ‘cosmetic formulation’. The sample sentences can be drawn as for example:

- Collagen is an important component for cosmetic formulation, where it is an effective natural humectant with high substantivity.
- Integrating toxicity assessment of new cosmetic ingredients early in the R&D cycle would help avoid developing ingredients with limited or no market potential due to toxicity concerns.

The next example is from Environmental Research Article corpus, the highest frequency of content word was “environmental”. The sample collocations of this word showed up clearly in KWIC concordances. Concordances can highlight the fact that they are not only compounds but have more parts of multiword lexical units as shown in Figure 5 below:

Top five of ‘environmental’ collocations in the Corpus of ESRAs were environmental management (Freq. 263), environmental performance (Freq. 137), environmental management system(s) (Freq. 162), environmental management practices (Freq. 90), and environmental protection agency (Freq. 58). We can see that frequency data of the phraseological terms can be a powerful tool in the hands of technical lexicographers since each level of frequency offers a potential cut-off point for headword selection and grading of items. The potential headword list for technical dictionary starts with those terms with middle to high frequencies in the corpus. The terms included in the dictionary are examined to make sure that they have technical meanings in the domain of food technology. Thus, this terminological dictionary is based on corpus frequency information, rather than on introspective intuition or a traditional inventory of words.

As Svensén (1993: 98) defines, the compounds are two or more words that always appear together, in the same form, with a fixed meaning. They consist purely of lexical words such as...
environmental management system or heat transfer. The fact that the lexical strings mostly consist of compounds and multiword lexical units, a close inspection revealed that the collocations in the corpora have a meaning exclusively their own and their meaning cannot be deduced by breaking their parts. A term consisting of two or three more words, e.g. proactive environmental management, is almost impossible to decode. For the EFL students, the length of the words makes them impenetrable. These collocations are difficult for EFL students. Therefore, in order to help them learn how to use properly, more evidence based instructional practices from the corpus in advance writing context seeks to address.

Qualitative observations revealed that some words have specific or technical meanings related to the specific genre. The LSP dictionary should be compiled basing solely on frequency criteria from the corpus, showing the separate senses in context, and examples of usage from the concordance in each dictionary entry. The focus would help students understand and encourage them to use appropriate vocabulary when writing research articles or completing academic assignments (Lewis, 2001).

5. Conclusion

The corpus frequencies and concordance data are evidently valuable in showing the linguistic contexts of terms. It illuminates the role of terms and their collocations in terms of compounds and multiword lexical units. The findings thus have applications for LSP lexicography. The evidence from the corpus suggests which vocabulary items are likely to be encountered by language users, and which therefore deserve to be headwords in dictionary. The forms, structures and functions of lexical bundles show the distinctive variations. Many phrases show highly specific terminological meanings with domain specific implications. The particular examples of terminology were recorded and selected as a raw material for compiling LSP dictionary and teaching English for Specific Purposes.

References
Chen, Qi, and Guangchun, Ge. 2007. A corpus-based lexical study on frequency and distribution of Coxhead’s AWL word families in medical research articles (RAs). English for Specific Purposes 26: 502-514.
Equipping Students with Lexicographical Means to Tackle Collocation Problems

Liu Liming
The Hong Kong Polytechnic University, Hong Kong
liming-leo.liu@connect.polyu.hk

Abstract
Collocation is a crucial aspect of the idiomaticity of language and a notoriously difficult area for second language development. Students who expect to achieve an advanced level of English academic writing have to be able to use collocations acceptable to the international academia. The most challenging aspect of the difficulty of collocations is the restrictive co-selection of collocating lexical items, making the production of collocations subject to prefabrication and not amenable to creativity as in the case of free combinations. This paper aims to pedagogically equip advanced L2 research students with lexicographical and corpus linguistics means to tackle their collocation problems, allowing students to bring initiatives to autonomous data-driven learning. The analysis identifies the relative strengths and weaknesses of an online academic corpus and a collocation dictionary for students to look for important lexicogrammatical information on collocations.

Keywords: collocation, academic writing, pedagogical facilitation, corpus-assisted learning

1. Introduction

The use of collocations contributes greatly to the idiomaticity of written production and has been regarded as a crucial component of second language (L2) vocabulary depth knowledge (Laufer and Waldman 2011; Read 2000). It has been agreed that L2 learners’ difficulty with collocations is not concerned with recognizing and understanding them in meaningful contexts as much as using them in free written and spoken production. The development of collocation production has been found to lag behind other aspects of L2 acquisition. Along the continuum of phraseological constructions, what distinguishes collocations from free combinations and idioms hinges upon the restricted co-selection of combinatory elements and relative semantic transparency (Laufer and Waldman 2011). Therefore, it can be inferred that it is the restriction on lexical selection that problematizes the use of collocations, since semantic transparency would not pose any great challenge for selecting co-occurring items.

The production of verb-noun collocations by L2 learners has been the most widely studied among research on lexical collocations, given the widespread presence of transitive verbs in finite and non-finite clauses (Biber et al. 1999). The principle of restricted lexical co-selection, as far as it is concerned with verb-noun collocations, suggests that only a limited range of verbs can occur with a particular noun, and reversely only a limited scope of nouns can go with a certain verb. In actual free written production, students’ major area of problems with verb-noun collocations is the inappropriate choice of verbs (Nesselhauf 2003) and learners have been found to overuse and misuse certain core verbs (e.g. take, get, give, have, make; Laufer and Waldman 2011). Calling them “lexical teddy bears”, Hasselgren (1994) explained the over-reliance on familiar lexical items in terms of their early presence in the L2 syllabus and high-frequency in English. A common feature of these core verbs lies in their “delexical” nature, i.e. they assume little perceivable lexical meaning by themselves apart from their prototypical meaning. Due to this delexical feature, each of these core verbs can take a large number of nouns to form collocations, serving to complete the meaning of the nouns. However, although being semantically less specific and being able to take a wide range of noun objects, these verbs are still subject to the general principle of arbitrary restriction on the co-occurring elements. Students’ excessive and sometimes erroneous use of collocations with these verbs either represents their unawareness of these verbs’ collocational
Students engaged in academic writing need to make constant lexical choices to form meaningful collocations. Without expert writing informants always being around, such a need is most likely to be satisfied when students are equipped with lexical resources to solve their immediate collocational problems. Useful resources to facilitate students’ collocation use may include learner’s dictionaries and online corpora. However, general learner’s dictionaries like the *Oxford Advanced Learner’s Dictionary* would not be the best place to be for detailed information on collocations since they are primarily intended for providing accurate word meaning. Although modern corpus-based learner’s dictionaries present words’ most representative collocation and colligation features, sometimes through enhanced typography (e.g. bold type and color) in examples, the information is still not adequate enough to provide comprehensive and reliable collocation guidance for academic writing. In contrast, collocation dictionaries for language learners are a much better resource for accessing collocational information. Take the *Oxford Collocations Dictionary* (Lea 2002) for example, collocates from each word class are provided for each sense of an headword and those from within a word class are separated according to their semantic relations with a sense of the headword for the user’s most informed reference. The other resource that students can benefit from once well trained is online corpora offering functionalities for collocation research. Providing valuable opportunities for data-driven learning and self-directed discovery learning as well as abundant information on the discourse features of particular collocations, online corpora can play an important role in facilitating students’ academic writing process.

In view of L2 students’ over-reliance on common verbs which may result in infelicitous verb-noun collocations and the dependable contribution to be made by lexicographical and corpus linguistics resources in facilitating students’ collocation production in academic writing, this paper aims to illustrate a detailed online corpus procedure for identifying verb collocates of nouns which have been found to co-occur with inappropriate verbs in student writing. Outcomes of this corpus procedure will be compared with consulting a collocation dictionary to identify areas where these two types of resource can complement each other. To provide a basis and empirical data for the lexicographical illustration, a corpus-based analysis of infelicitous verb-noun collocations as a result of misusing the delexical meaning of *make* is conducted first.

2. **The corpus-based analysis**

**2.1. Data**

The data used for the corpus-based analysis of infelicitous verb-noun collocations concerning the delexical meaning of *make* includes a corpus of 70 MA dissertations written by Chinese EFL students studying applied linguistics (hereinafter referred to as the MDC), with a size of around 1 million words. A contrastive corpus was compiled of 129 empirical research articles taken from 6 respected journals in the broad discipline of applied linguistics (hereinafter referred to as the PRC), with a total size of 971,726 words. The sample of the 70 dissertations is considered to come from a largely homogeneous population as all of them have been given a distinction grade. It needs to be noted that the PRC features authorship from a variety of L1 backgrounds, representing published writing of international authors rather than only limited to native speakers.

**2.2. Procedures**

Wordsmith Tools 5.0 was employed to automatically extract all concordance lines containing *make*, after which the infelicitous verb-noun collocations where *make* is a delexical verb were manually identified. The infelicity of a collocation consisting of *make* and a particular noun was first tentatively identified by the author, and then confirmed if the collocation had no
presence in the contrastive published corpus while that noun occurred with another verb to convey roughly the same meaning attempted by the “make” collocation. For example, make an explanation would be determined as infelicitous if this collocation does not occur in the PRC while provide an explanation is attested. The qualifier “infelicitous” was adopted rather than “erroneous” as in some studies of collocation (e.g. Laufer and Waldman 2011) because sometimes collocations unfavored in published writing are not really erroneous, e.g. make a thorough study, make an in-depth analysis; they are just very infrequently used, at least in some disciplines.

2.3. Results
2.3.1. Features of the nouns in infelicitous make-collocations
The nouns co-selected with the delexical meaning of make forming infelicitous and unidiomatic collocations are listed in Table 1 in decreasing order of frequency. The frequency includes instances of both singular and plural forms of the collocates, but the vast majority of the collocates are singular. There are a number of striking features discernable from the list of noun collocates. First, the collocates with the highest frequencies are more or less synonymous, belonging to the semantic category of academic study and analysis (the nouns in bold), e.g. analysis, study, research, and investigation; these nouns (tokens) account for 54% of all the instances (tokens) of the collocates of make. By using make with these nouns, the students introduced the entities analyzed and researched in their own studies and those of others. Another feature of the collocates is that few of the collocates are communication nouns, e.g. statement, claim, announcement etc, which represent the typical semantic domain of the noun collocates of make according to the Collins Cobuild English Grammar (1990). Where there are instances of communication nouns in the student corpus, such as explanation, instruction, question, and command, these nouns are rarely co-selected with make to construe verbal processes. Lastly, almost all of the collocates are instances of nominalization (except for idea, part, and statistics), either derived from a verb via nominalizing affixation (e.g. analysis from analyze) or through direct conversion from a verb without morphological transformation (e.g. study). In contrast, only about 62% of the noun collocates of delexical make used by English L1 students were nominalizations (Altenberg and Granger 2001), with quite a few of them being verbal communication nouns typically co-selected with make, e.g. argument, claim, statement, etc.

<table>
<thead>
<tr>
<th>Collocates</th>
<th>Freq. in MDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>analysis</td>
<td>39</td>
</tr>
<tr>
<td>study</td>
<td>20</td>
</tr>
<tr>
<td>research</td>
<td>17</td>
</tr>
<tr>
<td>classification</td>
<td>10</td>
</tr>
<tr>
<td>explanation</td>
<td>9</td>
</tr>
<tr>
<td>step</td>
<td>8</td>
</tr>
<tr>
<td>investigation</td>
<td>8</td>
</tr>
<tr>
<td>discussion</td>
<td>7</td>
</tr>
<tr>
<td>influence</td>
<td>3</td>
</tr>
<tr>
<td>experiment</td>
<td>3</td>
</tr>
<tr>
<td>instruction</td>
<td>3</td>
</tr>
<tr>
<td>idea</td>
<td>3</td>
</tr>
<tr>
<td>assignment</td>
<td>3</td>
</tr>
<tr>
<td>examination</td>
<td>2</td>
</tr>
<tr>
<td>definition</td>
<td>2</td>
</tr>
<tr>
<td>question</td>
<td>2</td>
</tr>
<tr>
<td>control</td>
<td>1</td>
</tr>
<tr>
<td>understanding</td>
<td>1</td>
</tr>
<tr>
<td>command</td>
<td>1</td>
</tr>
<tr>
<td>summarization</td>
<td>1</td>
</tr>
<tr>
<td>translation</td>
<td>1</td>
</tr>
<tr>
<td>representation</td>
<td>1</td>
</tr>
</tbody>
</table>
2.3.2. Possible factors contributing to the infelicity

Both intralingual and interlingual factors can be drawn upon to explain the Chinese students’ infelicitous use of collocations, including 1) the desire to achieve idiomaticity, 2) unskilled use of lexical verbs, 3) and negative L1 (Chinese) transfer. Regarding the desire for idiomaticity, research has shown that second language learners have a strong desire to be idiomatic (Liontas 2002) and idiomaticity can be achieved through the use of high-frequency, everyday words in their formulaic collocations (Martinez and Murphy 2011; Sinclair 1991). Thus, by constructing collocations with high-frequency verbs such as make students may feel as if they are writing idiomatically. If the hypothesis about the students’ aspiration towards idiomaticity is wrong, however, they must have adopted what Sinclair (1991) calls the open-choice principle, regarding the combinations of make and the nouns as largely compositional, i.e. freely co-selected based on the semantic compatibility between the component words. However, students may not have realized that since delexical verbs are semantically rather limited, they therefore have a limited range of co-selection possibilities, often semantically unmotivated (Allerton 1984).

The students’ somewhat unconstrained selection of nouns with delexical make points to their underdeveloped skills in retrieving semantically more appropriate lexical verbs, either to form collocations with the nouns or to substitute the whole collocations. In fact, in most of the infelicitous collocations it is the selection of make that results in the infelicity of the collocations, rather than the nouns. In other words, students’ selection of make is semantically motivated by the nouns which are the focus of the meaning being conveyed, not the other way around. Thus, unaware of the semantic restriction of make, students are shown to suffer from a lack of compatible verb collocates of the substantive nouns. Alternatively interpreted, concentrating on nominalizations may have blocked their view of the possibility to use the canonical verb form rather than nominalization, e.g. to classify students’ answers instead of to make a classification of students’ answers. As will be shown in the second part of the paper, classification is seldom preceded by a verb to form collocations for conveying the meaning of classify. Thus, using nominalizations with make makes the wording appear somewhat redundant and clumsy, apart from sounding unidiomatic. The Chinese students’ propensity for using nominalizations in this case might be supported by their belief that nominalization constitutes a distinctive feature of academic writing and frequent use of the device is a sign of being academic. In this sense, their understanding of nominalization and the discourse functions it serves would have to be adjusted.

The above interpretations have been made from an intralingual perspective, highlighting problematic areas of collocation that may apply to learner groups from different L1 backgrounds. The last factor, negative L1 transfer, carries on the discussion from an interlingual point of view. One possibility is that the Chinese verb for the “carrying out” of many of the nouns listed above can be zuo4 or zuo4 chu1, literally equivalent to do or make in English, e.g. do/make + study (zuo4 yan2 jiu2), do/make + investigation (zuo4 diao4 cha2). Being a versatile verb, zuo4 is able to enact a wide range of actions in Chinese. Probably conceiving do to be much too informal to appear in academic writing, the student writers have opted for make to complement the meaning of the nouns. But in fact, do is a more felicitous verb collocate for some of the nouns, especially the “study” and “analysis” nouns, e.g. study, analysis, research, and experiment. Of course there are other possible verb collocates for these nouns other than do that students can choose from, provided that they are equipped with useful lexicographical and corpus linguistics resources and tools.
The second part of the study aims to remedy students’ inadequate repertoire of verb collocates to complement the meaning of nouns by presenting details of a corpus-based method to search for appropriate verb collocates. Comparisons of this method are made with searching for collocates from a collocation dictionary to identify the relative strengths and weaknesses of each method.

3. Equipping students with lexicographical resources

It can be argued that students acquire productive knowledge of academic vocabulary mainly through reading academic texts (e.g. research articles, monographs, and theses/dissertations) whereby frequency effects play a major role in transforming input/intake into output. This transformative process does not always take place automatically and without barricade in spontaneous written production (cf. Larsen-Freeman 2002). Therefore, pedagogical treatment to facilitate students’ vocabulary choice in academic writing is in order. As students need to tackle various and individualized vocabulary challenges for their specific, local communicative purposes, it is impossible for teaching to cover all or even most of students’ lexical needs in their free written production. Thus, a more viable solution would be to provide students with lexicographical references and corpus linguistics tools for them to autonomously discover the specific lexis called for. Such an approach can be especially useful for dealing with collocation. The remainder of the paper illustrates how an online academic corpus can solve student writers’ collocation problems and compares its relative strengths and weaknesses with collocation dictionaries. Suggestions for vocabulary learning are also provided in due course.

The online corpus for demonstration is the Corpus of Contemporary American English (COCA; Davies 2008), the largest online corpus freely available. An advantage of COCA as a collocation reference tool is that it enables the user to search for either specific co-occurring words for a particular query word (e.g. to search for instances where make co-occurs with claim) or all the co-occurring words of that query word (e.g. all the words co-occurring with claim); the search engine also allows the user to specify the part of speech for either the query word or its co-occurrences (e.g. all the verbs co-occurring with claim as a noun). Besides, COCA is nicely sectioned into an academic subcorpus with four other genres (i.e. TV conversation, fiction, newspaper, and magazine), allowing the user to focus on language use in the academic section only or to compare usage across genres. These features make COCA a powerful tool for working with the collocations of specific words. The collocation dictionary referred to is the CD-ROM version of Oxford Collocations Dictionary for Students of English (Lea 2002), which is a widely-used corpus-based learner’s dictionary, deriving words’ collocates from their recurrent patterns attested in corpora.

3.1. Using COCA as a reference tool in writing

As introduced above, two types of processes are possible in gaining access to collocations in students’ written production: searching for a list of collocates and searching for specific collocates. Searching for a list of collocates is useful when students cannot come up with an appropriate collocate for a word (e.g. all the verb collocates for claim as a noun) while searching for specific collocates can help students test their hypotheses about the collocability of two words (e.g. if make can collocate with the noun influence). It is suggested that students go through both of these two processes, preferably with hypothesis testing preceding list generation for students to get more actively involved in their thinking and learning efforts. The following demonstration of how to use COCA for retrieving collocates begins with the list-generating activity and carries on to hypothesis testing. Examples from the nouns in infelicitous make-collocations listed in Table 1 will be used in the demonstration.

3.1.1. Generating a list of collocates

Although the previous corpus-based analysis of infelicitous collocations is derived from the delexical make, it is not the intention of the present demonstration to examine all the possible
noun collocates of delexical verbs and their semantic categories (cf. Altenberg and Granger 2001) but to present ways of searching for verb collocates of particular nouns. This is because in the actual production of verb-noun collocations students would normally come up with the noun first and then select from their lexicon an appropriate verb for collocation; it is the noun that activates the selection of its verb collocate. Therefore, to specifically address the collocation problems attested in the student corpus, the following illustration proceeds by showing how to obtain possible verb collocates of the nouns that the student writers wanted to express.

Figure 1 shows the initial interface of COCA, where we can see that there are four basic functions—LIST, CHART, KWIC (key words in context), and COMPARE. The LIST function needs to be selected for searching collocations. In the initial interface, the COLLOCATES and POS (part of speech) LIST buttons are inactive, so the user needs to activate the two functions by clicking on them, which changes the interface as shown in Figure 2. By now the user can start searching for collocations. In order to search for the left verb collocates of a particular noun, that noun should be put into the WORD(S) blank along with its POS tag, and the “verb.LEX” (meaning all lexical verbs) option should be selected from the POS LIST menu. Lastly, a span of 4 words to the left of the noun is allowed considering the possibility of premodifiers and determiners before the noun. The input and options for the search are shown in Figure 3, with the noun step as an example search word. The search syntax [explanation],[nn*] means that the query only involves step as a noun, in both its singular and plural forms, thus excluding all instances of the word’s other parts of speech. (Although explanation can only be a noun, there are homographs having more than one word class, e.g. study.)

Grounded in the academic subcorpus, the above search would return 1,751 lexical verbs (types) occurring to the left of explanation; the top 30 co-occurring verb types are listed in Table 2 with their raw frequencies. As can be seen, a number of verbs stand out as possible verb collocates of explanation, including provide, offer, require, give, find, suggest, and need, and accept. It needs to be cautioned that not all the verbs and not every instance of a verb may actually form a v + n collocation, e.g. suggest co-occurs with explanation both as verb-noun collocations and as suggest taking a complement clause in which explanation occurs. Therefore, students need to take this into careful consideration when working with corpus results.

Having access to these co-occurring verbs of explanation, students who have used make to collocate with the noun should be able to identify a better verb to realize the meaning they expect to make:

*The stereotype that teachers stand in the front to make explanation and students sit down there to listen and to take notes remains the dominant form in class.*

Thus, provide/give/offer would be more appropriate candidates for what the student had in mind to complement the meaning of explanation in this context. Moreover, information on the frequency distribution for each inflected form also allows students to have a general picture of the tense/aspect pattern of this specific collocation in academic writing, which they may draw upon in their own writing. Another bonus of using corpus to learn collocation knowledge is that concordance lines can be observed for other lexicogrammatical clues concerning the collocation in question. For example, students may notice that in the actual collocations of provide and explanation, explanation often appears in singular form with an indefinite article (provide an explanation) or in plural form without article (provide explanations), and that explanation is usually followed by an adjective premodifier (provide a satisfactory explanation). Back to the above example, the student would probably have used the plural form explanations had s/he had access to the concordance lines. Such value-added

---

3 For a detailed description of search syntax for COCA, see [http://corpus.byu.edu/coca/help/syntax_e.asp](http://corpus.byu.edu/coca/help/syntax_e.asp); for the POS tagset used by COCA, see [http://ucrel.lancs.ac.uk/claws/tags.html](http://ucrel.lancs.ac.uk/claws/tags.html).
information on the co-occurrence of two words would greatly facilitate students’ vocabulary choice in the writing process.

Figure 1. The basic search interface of COCA

Figure 2 COCA interface with activated COLLOCATES and POS LIST functions

Figure 3 Example search for left verb collocates of the noun explanation

Table 2 Top 30 lexical verb types occurring to the left of explanation in COCA-academic

<table>
<thead>
<tr>
<th>Verb</th>
<th>Count</th>
<th>Collocate 1</th>
<th>Count</th>
<th>Collocate 2</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provide</td>
<td>178</td>
<td>gave</td>
<td>30</td>
<td>giving</td>
<td>20</td>
</tr>
<tr>
<td>Provides</td>
<td>111</td>
<td>find</td>
<td>29</td>
<td>included</td>
<td>17</td>
</tr>
<tr>
<td>Offer</td>
<td>102</td>
<td>seek</td>
<td>28</td>
<td>needed</td>
<td>17</td>
</tr>
<tr>
<td>Offers</td>
<td>82</td>
<td>consider</td>
<td>27</td>
<td>needs</td>
<td>17</td>
</tr>
<tr>
<td>Offered</td>
<td>74</td>
<td>suggests</td>
<td>27</td>
<td>offering</td>
<td>17</td>
</tr>
<tr>
<td>Providing</td>
<td>43</td>
<td>given</td>
<td>26</td>
<td>suggested</td>
<td>17</td>
</tr>
<tr>
<td>Provided</td>
<td>41</td>
<td>found</td>
<td>24</td>
<td>believe</td>
<td>15</td>
</tr>
<tr>
<td>Requires</td>
<td>40</td>
<td>suggest</td>
<td>24</td>
<td>see</td>
<td>15</td>
</tr>
<tr>
<td>Require</td>
<td>38</td>
<td>used</td>
<td>22</td>
<td>think</td>
<td>15</td>
</tr>
<tr>
<td>Give</td>
<td>34</td>
<td>need</td>
<td>21</td>
<td>accept</td>
<td>14</td>
</tr>
</tbody>
</table>

3.1.2. Comparison with a collocation dictionary
Similarly, the Oxford Collocations Dictionary also provides a comprehensive list of verb collocates of explanation, as shown in Figure 4. However, an obvious convenience of this list is that the verb collocates are visually grouped (with words in bold type and an example in italics between two vertical bars) based on their meaning relationships with the noun—thus, give, offer, and provide are grouped together due to their similarity in meaning. This user-friendly layout design makes it convenient for students to identify the words compatible with
their desired meaning. This feature is an obvious advantage over the corpus collocate list from which students have to tease out useful information by themselves. However, what the collocation dictionary lacks is discourse correlates for the collocations, such as information on the tense/aspect patterns of each verb collocate and other lexicogrammatical features relating to the noun, such as its number, determiner, and premodification. As it would not be viable for dictionaries to include too much extra information, data from both the corpus and dictionary should be referred to for making more informed vocabulary choices.

![Figure 4 Verb collocates of explanation listed in the Oxford Collocations Dictionary](image)

Alongside being equipped with corpus linguistics and lexicographical means to access verb-noun collocations, students need to be made aware that some nominalizations seldom have to form verb-noun collocations to realize the verbal process (e.g. the Oxford Collocations Dictionary does not provide any verb collocates of classification meaning make or create), but rather their prototypical verbs are more frequently used to fill in the predicate position.

4. Conclusion

Having found that L2 research students, similar to undergraduates (e.g. Laufer and Waldman 2011; Nesselhauf 2003), also suffer from a lack of idiomatic verb-noun collocations and therefore end up resorting to very common verbs such as make to form felicitous collocations, this paper has detailed a procedure for identifying verb collocates of nouns based on a popular online corpus, COCA. It has been shown that the collocate functionality of COCA is a viable and effective tool for accessing verb collocates of a particular noun with much value-added lexicogrammatical information, including their frequencies when collocating with the noun, tense/aspect patterns, and within-collocation linguistic features. Compared with a well-designed collocation dictionary, the corpus tool has been shown to demand careful screening of real collocates from non-collocates on the part of the user although providing much more discourse-related add-ons. It is therefore suggested that both the two types of collocation resources referred to for their relative strengths.

References


Compilation of Traditional Building Construction Terminologies in Tamil

Saravananan Raja
Annamalai University, India
kiranrajaling@gmail.com

Abstract
The native vocabulary of different indigenous occupations is slowly being replaced by those taken from the English language, because of the influence of science and technology and the adoption of modern methods. This may lead to a permanent loss of rich occupational vocabulary in Tamil. This paper mainly focuses on different issues related to the collection of lexical items related to traditional building construction in Tamil. This paper explains the significance of a survey and compilation of a dialect dictionary. The significance of a dialect survey is to bring out the variant forms of regional dialects. By utilizing the structure of the dictionaries identified by Hausmann and Wiegand (1989) and Wiegand (1989) and the recent research in the field of metalexicography, the compilation of a dialect dictionary of traditional building construction in Tamil has been carried out. This paper explains the different structures of a dictionary ie. microstructure and macrostructure. The microstructure of the proposed dictionary has two important structures; the first one is the information structure, which is not a lexicographic definition, but the lexical description given by the construction workers. The second structure provides the variation found in the different regions. The microstructure consists of the lemma sign which is the spoken variety only. The results of the survey have been presented in the form of a dialect dictionary in Tamil.

Keywords: microstructure, macrostructure, information structure, lemma sign, lexicographic description.

Introduction
Lexicography is the discipline concerned with the principles and methods of compiling dictionaries. It is an applied field, whose theoretical background is partly utilized by linguistics. Almost all the branches of linguistics provide information in making or writing a dictionary. Architecture is one of the aspects of social life in which innovations through acculturation are more quickly and readily accepted and assimilated than in other areas like fishing, agriculture, weaving, carpentry etc. In this occupation, communication flows in both directions i.e. high to low (engineer, supervisor, maistris, coolies and vice-versa) along the social scale. Hence recording and documenting vocabularies of building construction in Tamil is essential for the future generations. The present paper envisages the design and production of dialect occupational dictionaries in general and in particular about dialect dictionary of building construction in Tamil.

The present paper deals with the relation and interaction between regional, social and occupational differences and their impact on language variation. The native vocabulary of different indigenous occupations is being slowly replaced by the English and international terms because of the influence of science and technology and adaptation of modern methods. This may lead to the permanent loss of rich occupational vocabulary of Tamil. The survey and the compilation of a dictionary will lead to the recording and preservation of that vocabulary. Furthermore, in the absence of a detailed dialect survey of Tamil Nadu and a fully-fledged dialect dictionary for construction terms, this study will provide a model for both these aspects and also for the future survey and recording of several other occupational vocabularies of Tamil. If the hypothesis that the social dialect differences are not reflected in...
the occupational vocabulary is proved, then this can serve as a basis towards solving the problem of communal differences and resultant communal clashes which are rampant among the working classes.

**Methodology**

The following procedures and methods have been adopted for conducting the survey of constructional vocabulary and the compilation of dialect dictionary.

**Survey**

*Selection of points:* Indigenous constructional vocabulary were collected from several points comprising of the five major dialect regions of Tamil Nadu viz., Eastern dialect region, Western dialect region, Southern dialect region, Northern dialect region and Central dialect region.

*Informants:* The informants were selected from densely populated areas. The informants, both male and female were of different age groups and caste groups. Even old persons in the age group of 50 – 70 who have no educational background were also selected for the elicitation of data.

*Areas or fields of investigation:* Lexical items relating to the following areas were collected.

Different types of constructions: traditional building construction, construction of huts, construction of wells, etc.

Different types of building materials: Brick, stones, lime and mortar, timber, grasses (used for roofing materials, fencing), tiles, etc.

Different categories of workers and their tools: Professional and casual labourers, different tools and their parts.

Different types of construction methods: Several stages and parts like:
2. Roof level:
   - Permanent and temporary roof, materials used, different shapes and uses, parts, construction methods
   - Opening on roofs: parapet walls, outlets and water courses, chimneys, etc. and their parts
   - Domes and upper stones, parts
3. Flooring types and methods, polishing, etc.
4. Types of doors, materials used and polishing
5. Types of windows, parts and uses
6. Niches in walls: types based on shapes and utility
7. Almirahs: types and parts
8. Types of walls
9. Pillars and columns: wood, stone, brick, etc.
10. Arches: different shapes
11. Rooms and verandahs
12. Stairs, ventilators
13. Balcony and its parts
14. Beams of different shapes and materials used
15. Measurements; partition and extension
16. Customs and faiths: social and religious customs, festivities and oblations, materials used.

*Points of observation:*

i. Evidence for regional variation including sub regions if any
ii. Whether any social differences are reflected in the regional differences  
iii. Age and sex differences and their reflections  
iv. Differences if any among cultural and linguistic groups settled in Tamil Nadu like Telugus, Kannadigas, Malayalees and Tribals.

Tools used in the collection of data: A Questionnaire was administered for collecting constructional data. The questionnaire consists of two parts: (i) background information of the informants like age, sex, education, caste or caste group, place, region, etc.,(ii) different questionnaires were prepared for collecting information about various systems of construction like construction of traditional buildings, construction of huts, construction of wells, etc.

Component parts and structure of a dictionary

There are three major structures of a dictionary viz., microstructure, macrostructure, and megastructure. Normally dictionaries have a double structure i.e. macrostructure and microstructure. According to Rey-Debove, ‘macrostructure is the arrangement of the list of entry words in each dictionary, the complete set of entries arranged in some orders, part of which is always used for vertical scanning when the user is looking for a practical piece of information’. The macrostructure is commonly called ‘nomenclature’ (Lehmann, 2006:202). The term macrostructure can be used as synonym for nomenclature, or word list, but it may be useful to distinguish the two: nomenclature refers to the list of lexical items that are treated in a dictionary and macrostructure refers to the same list once it has been organized to constitute the architecture of dictionary.

The term microstructure refers to the set of pieces of information as they are ordered in every article which are meant to be read horizontally after the entry word (Rey Debove, 1971:21). The microstructure of a dictionary is composed of a certain number of information items, or fields. Some metalexicographers have proposed the notion of address which refers to any element, headword, sub-head word, run-on, meaning etc., to which microstructural information - a definition, an explanation, an illustration, an example, etc. applies (Bejoint, 2010:14). Megastructure deals with the relationships and order between its main components. Hartmann and James (1998) have pointed out that ‘- structures have recently appeared’; megastructure for the organization of the different parts of dictionary from first to the last page including A to Z text; mediostructure for the system of cross references; iconostructure for the organization of pictorial illustration (Boulanger, 2003:13).

The macrostructure of the proposed dictionary has a single word list. This central word list will follow the Tamil alphabetical arrangement. The present proposed dictionary has monosemous lemmas in nature i.e. it represents one single part of speech and having a lexical description of that lemma. Regarding the microstructure of the proposed dictionary, entry is given in the following format. The lemma or head word is given in Tamil script with bold letters, followed by the phonemic script (transliteration). The grammatical information (abbreviated form) is also given. In addition to this, the English equivalent is provided. Normally semantic equivalence is given i.e both the translational equivalence and explanatory equivalence. Then, the lexicographic description is given in Tamil which is added for better understanding of the target language group (Tamil language group). Normally this lexicography description is encyclopedic in nature. After this, examples are also provided. In addition to this, visual illustrations are also provided. Finally, the most important feature of the present dictionary is the inclusion of labeled structure. Due to this salient feature/structure of the dictionary, this is referred to as a special dictionary.

Dictionary compilation

The results of the survey are presented in the form of a dialect dictionary showing the variations. While the starting point for the survey is the different dialect regions, the starting point for the dictionary is the different dialect terms or lexical units designating various
constructional concepts, processes and items. The presentation of lexical items is alphabetical. Among the various lexical units designating a single item of the study, the central regional dialect has been taken as the lemma sign or head word and other variants have been presented by indicating the region with abbreviated labels. Wherever necessary, encyclopedic information has been provided in the form of description and line drawings and photographs. English equivalents have been provided for wider use of the dictionary.

Conclusion

The main purpose of the present proposed dictionary is to record the native vocabulary of an indigenous occupation i.e. construction. Due to the influence of English language and due to the modern technological developments, many indigenous occupational vocabularies are becoming obsolete in nature. Hence it is necessary to record those indigenous vocabularies and preserve it in the form of dictionary for the benefit of the future generations.

References

Representation of the Pronunciation of Japanese Words in
Racuyoxu (1598) and Vocabulario Da Lingoa De Iapam (1603–04)

Kishimoto, Emi
Kyoto Prefectural University, Japan
kishimoto@kpu.ac.jp

Abstract
Approximately half a century after they came to Japan, the Jesuits printed two different Japanese
dictionaries: Racuyoxu (1598) and Vocabulario da lingoa de Iapam (1603–04). Racuyoxu is a Chinese-
character (kanji) dictionary showing how to read words written in Chinese characters by also
representing them in the Japanese kana syllabary. The Vocabulario da lingoa de Iapam is a Japanese
dictionary with translations and explanations in Portuguese, printed in Latin letters. These dictionaries
were complementary to the Jesuits’ learning of Japanese: Racuyoxu was compiled to aid the learning of
Chinese characters while Vocabulario the meanings of vocabulary items.

The names of the editors of both volumes are unknown. The editors of Vocabulario must have referred
to Racuyoxu, as Morita (1993) indicates by several clues. However, the two dictionaries include
different readings of several words with the same meaning. For example, we find the words baibainin
in Racuyoxu and baibaijin in Vocabulario, both of which were current in that period and mean
‘merchant’. In addition, some readings in Racuyoxu are regarded as less correct than those in
Vocabulario. For example, Racuyoxu includes the words cocumo ‘mother of the emperor’ and
cotozzute ‘message’, while Vocabulario indicates that cocumo and cotozzute are less correct
pronunciations than cocubo and cotoçute. The likelihood is that these differences were caused by the
volumes’ different sources and editors.

Keywords: Jesuit mission press in Japan (Kirishitan-ban), Chinese character (kanji) dictionary, Racuyoxu (Rakuyosu), Vocabulario da lingoa de Iapam (Nippo-jisho)

1. Introduction

When the first missionaries of the Society of Jesus, i.e., the Jesuits, most of whom were
Portuguese, came to Japan in 1549, they studied Japanese language and culture to propagate
the Catholic faith. It is believed that they wrote grammars and dictionaries for the study of the
Japanese language prior to the 1590s; unfortunately, these works are not extant today. The Jesuits began to print books for their missionary works in Japan, and these are called the
Jesuit mission press or Kirishitan-ban in Japan today. A list of the grammar books and
dictionaries from the Jesuit mission press in Japan that do survive today is provided below:

1594 De institutione grammatica (Latin grammar, originally by Manuel Alvarez)
1595 Dictionarium Latino Lusitanicum, ac Iaponicum (Latin-Portuguese-Japanese dictionary)
1598 Racuyoxu 落葉集 (Dictionary of Chinese characters)
1603–1604 Vocabulario da lingoa de Iapam (Japanese–Portuguese dictionary)
1604–1608 Arte da lingoa de Iapam (Japanese grammar by João Rodrigues)

In this paper, I would like to focus on the two different Japanese dictionaries: Racuyoxu 落葉集 (1598) and Vocabulario (1603–04). The names of the editors of both volumes are
unknown.

First, I would like to provide bibliographical information about the two dictionaries that show
the text. Racuyoxu is a Chinese-character (kanji) dictionary that demonstrates how to read
words written in Chinese characters by also representing them in the Japanese kana syllabary.
This work does not provide word meanings. Today, there are four complete copies of this text in the world: Archivum Romanum Societatis Iesu (ARSI), British library, Tenri library, and Earl of Crawford and Balcarres. I mainly use a reproduction of the ARSI copy, confirming using that in the Tenri library. The entire book comprises 109 folios, including approximately 2,400 distinct Chinese characters and approximately 500 words written in Chinese characters (Toyoshima, 2009). The dictionary consists of three parts: the main section (honpen), iroha-jishu 色葉字集, and sho-gyokuhen 小玉篇. In the main section (honpen), the words are arranged in the order of Sino-Japanese reading (on-reading), iroha-jishu in the order of Japanese reading (kun-reading), and Chinese characters are classified by Chinese radicals in sho-gyokuhen. It has been said that Racuyoxu was edited based on several kinds of dictionary that were widespread in Japan, such as Setsuyoshu 節用集 (a dictionary of Japanese words written in Chinese characters) and Wagocuhen 倭玉篇 (a dictionary of Chinese characters with on-reading and kun-reading). Below I provide an example of the main section, which I will examine in more detail later, added an example of iroha-jishu. In this paper, I romanized kana used in Racuyoxu as was the Jesuits’ method, while the English translations are in [ ].

Ex.1: words beginning with 専 in the main section (honpen) and 囲 in iroha-jishu in Racuyoxu

(right side: on-reading for each Chinese character)
専* -念 -要 -一 -輙
(xen nen you ichi teu [xen-nen, concentration, xen-you, most important, xen-ichi, first, xen-teu, arbitrary decision]
(right side and bottom: kun-reading)
cacomu [to surround] meguru [to circle]
(centre: Chinese character)囲
(left side: kun-reading)
(tayasuxi [easy]

The Vocabulario da lingoa de Iapam is a Japanese dictionary with translations and explanations in Portuguese, all printed in Latin letters. The main body was printed in 1603 and the supplement in 1604. The total body comprises 402 folios, including more than 32,000 headwords (Morita, 1993). There are two known complete copies, in the Bodleian library and Biblioteca Publica da Évora, and in this study I use a facsimile of the Évora copy. According to its preface, the Vocabulario was edited based on previous works by the Jesuits in Japan including several scripts and the Latin-Portuguese-Japanese dictionary (1595). Example 2 shows the same words as in Ex.1 in Racuyoxu as presented in the Vocabulario.
Ex.2: words *xennen* [concentration] and *meguru* [to circle] in *Vocabulario*

Xennen. Mopparani vomô. [to think deeply, dedicate] O trazer sempre algûa cousa no pensamento, & memoria como a saluaç ã o, &c. S.[always thinking about something, for example about salivation and so on. A literary word.]

Meguri, u, utta. [adverbial form, conclusive form, past form of the verb *meguru*] Rodear. [to circle]

Each dictionary has been used and studied quite extensively; however, there have been few comparative studies apart from Doi (1971) and Morita (1993). As Doi (1971) indicates, the two dictionaries were complementary for the Jesuits’ learning of Japanese: *Racuyoxu* was compiled to aid the learning of Chinese characters, and the *Vocabulario da língoa de Iapam* provided the meanings of vocabulary items. Furthermore, with regard to the relationship between these texts, there are several clues that the anonymous editors of the *Vocabulario* must have referred to the *Racuyoxu*, as Morita (1993) indicates. For example, in the main section (*honpen*) of the *Racuyoxu*, we find a word 起居動静 [daily life] with a reading qicodôyǒ, which should be qicodôjǒ. The same error is made in the supplement of the *Vocabulario*, which can be surmised to be based on the *Racuyoxu*.

However, the two dictionaries also include different readings of several words with the same meanings. In addition, as Doi (1971) indicates, some of the readings in the *Racuyoxu* are regarded as less correct than those in the *Vocabulario*. For example *Racuyoxu* has a word 善言 [good and useful words] with the reading *jengon*; however, the headword of *jengon* in the *Vocabulario* notes that the pronunciation *jenguen* is better than *jengon*. Indicating this example, Doi (1971) says this was possibly inconsistency because the pronunciation in contemporary spoken Japanese was important in the *Vocabulario*, while the editors of the *Racuyoxu* thought it important to show Chinese character forms and examples written in older materials.

I basically agree with Doi’s opinion; the important thing to consider is the difference in target reader. The target of the *Racuyoxu* was learners of a much higher level than of the *Vocabulario*, who already knew Chinese characters to some extent. The users of the *Racuyoxu* should know as many Chinese characters and their readings as possible, while the *Vocabulario* included amongst its users elementary learners who had never learned Chinese characters.

However, it should have been ideal for every Japanese learner that dictionaries showed one reading per word even if that word had several different readings. It remains a mystery why different readings occurred between two dictionaries printed by the Jesuits in Japan within only a few years of each other. In this paper, I would like to examine more examples from the two dictionaries and consider the causes of this inconsistency in the view of compilation.

2. Different readings between *Racuyoxu* and *Vocabulario*
Among the different readings of words between the Racuyoxu and Vocabulario, in this paper I will focus on Japanese words of Chinese origin (words of on-reading) consisting of more than two Chinese characters, not Japanese original words (words of kun-reading), because we can decide more easily in the case of the former whether the different readings of words had the same meanings. Moreover, I exclude words for which the pronunciation was in the process of change such as ji and gi, zu and zzu, because it is known that this evaluation of pronunciation was causing confusion. I also exclude simple differences in kana (kanazukai), which were rather common at that time in Japan and did not signify a difference in pronunciation, such as the written forms yuube/yufube [evening] and yafaraca/yauaraca [soft]. Yamada (1971) provides many such examples.

Let us now look at Table 1, which shows the numbers of the same and different reading between the Racuyoxu and Vocabulario in words beginning from A to Co, which includes approximately one-third of the main section. Today, we spell ‘A, I, U, E, O, Ka, Ki, Ku, Ke, Ko, Ga, Gi, Gu, Ge, Go’ based on English Romanization.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>I</th>
<th>V</th>
<th>Ye</th>
<th>Vo</th>
<th>Ca/Ga</th>
<th>Qi/Gui</th>
<th>Cu/Gu</th>
<th>Qe/Gue</th>
<th>Co/Go</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same</td>
<td>121</td>
<td>311</td>
<td>44</td>
<td>122</td>
<td>42</td>
<td>343</td>
<td>427</td>
<td>335</td>
<td>276</td>
<td>275</td>
<td>2296</td>
</tr>
<tr>
<td>Different</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>11</td>
<td>9</td>
<td>14</td>
<td>7</td>
<td>5</td>
<td>58</td>
</tr>
</tbody>
</table>

As is evident, most of the readings are the same between the two texts. First, let us examine some examples of the same reading. Ex.3 shows a common reading corai between the two. In Ex.4, Vocabulario has two different readings for 眼力, because the Chinese character 力 had two different pronunciations, riocu and riqi, in Japan and Racuyoxu has only the one reading ganriocu. When I find such cases, that is, when both dictionaries have several readings for one word but have one common reading, I include the example in the common reading group.

Ex.3: corai [from old times]
Racuyoxu: 古来 corai
Vocabulario: Corai. Inixiye, cono cata. [from old times to the present] Desdo tempo passado até gora. [from past time to now]

Ex.4: ganriocu/ganriqi [insight]
Racuyoxu: 眼力 ganriocu
Vocabulario: Ganriocu. Manacono chicara. [power of eye] Força, ou vigor da vista. [power or energy of the sight] …
Ganriqi. Idem. [the same]

The number of differences is quite small; however, they cannot be ignored because these learner’s dictionaries were edited by the same group. The examples of difference can almost be divided into two groups. One group is related to euphonic change called rendaku, in which unvoiced sounds are changed to voiced sounds for the first consonant of the no-initial mora, as in Ex.5. The Chinese character 制 has been given the on-reading xei in Japan and changes to jei when it comes after the character 禁. The other group is related to changes in the type of on-reading, as in Ex.6. The Chinese character 色 had two kinds of on-reading, xiqi and xociu, as a result of differences during the period when their pronunciation was introduced to Japan.

Ex.5: qinxei/qinjei [prohibition]
Racuyoxu:
禁制 qinxei
Vocabulario:
Qinjei. Prohibição. [prohibition] Vt. Qinjeiuo somuqu. [for example, to break prohibition]
Quebrantar a proibiçã o, ou ley. [to break prohibition or regulation]

Ex.6: giocuxiqi/giocuxocu [beautiful colour]
Racuyoxu:
玉色 giocuxiqi
Vocabulario:
Guiocuxocu. Tamano iro. [colour like jewels] Itçucuxij iro. [beautiful colour] Boa, ou fermo sa cor. [fine, or beautiful colour]

In fact, both types of different readings have often occurred due to changes or variations in pronunciation in the Japanese language. For example, today for rendaku, we can read 研究所 as kenkyusho or kenkyujo. The on-reading of the Chinese character 所 is sho but unvoiced and sequential sho is often changed to voiced jo, which is a euphonic change. As another modern example, the Japanese word 奥義 can be read as oogi or okugi, both of which denote the heart of certain studies or art. The Chinese character 奥 has two possible readings, oo and ok, in Japan because the Japanese people received various Chinese pronunciations at least twice during the ancient period. While oogi is an older reading than okugi, both pronunciations continue in use even today.

In addition, with regard to the Racuyoxu and Vocabulario, like the example of jenguen and jengon introduced at the beginning, it is remarkable that some of these differences are noted in the explanation of the Vocabulario. In Ex.7 and Ex.8, the Racuyoxu includes the words jengon [good and useful words] and cocumo [mother of the emperor], while the Vocabulario indicates that jenguen and cocubo are better pronunciations than jengon and cocumo without providing a detailed explanation. The Chinese character 言 has two possible pronunciations in Japanese: older gon and newer guen, and 母 likely has two pronunciation of older mo and newer bo.

Ex.7: jenguen/jengon [good and useful words]
Racuyoxu:
善言 jengon
Vocabulario:
Iengon. Yoqi cotoba. [good words] Melius, Ienguen. [jenguen is better than jengon] Boas palauras. [good words]

Ex.8: cocumo/cocubo [mother of the emperor]
Racuyoxu:
国母 cocumo
Vocabulario:

Among such pairs of on-readings in the Vocabulario, we can find approximately 40 words in the Racuyoxu that are regarded as better pronunciation in the Vocabulario. However, in contrast, at least 13 words are found in the Racuyoxu, including Ex.7 and 8 above: 国母 cocumo, 脱去 dacqio [escape, Vocabulario says ‘dacco is better’], 下筆 guefit [bad handwriting, ‘cafit is better’], 下官 guequan [officer of lower class, ‘caquan is better’], 下略 gueriacu [omission of under part of word or sentence, ‘cariacu is better’], 善言 jengon, 上略 jòriacu [omission of upper part of word or sentence, ‘xòriacu is better’], 勉賞 quanjô
[reward, ‘genjô is better’], 束香 sucô [bad incense, ‘socô is better’], 水泉 suixen [fountain, ‘sensui is better’]. 飲水 von sui [to drink water, ‘vonzui is better’], 清香 xeicô [refreshing incense, ‘xeiqiô is better’], 昇進 xôxin [to be promoted, ‘xôjin is better’]. Those words are classified into three groups: nine of which are related to the change of on-reading such as guefit/cafît, two are rendaku such as von sui/vonzui and xô xin/xô jin, and the other two are sucô/ socô and suixen/sensui.

Let us now consider what caused such differences. One main cause is probably the sources. As indicated previously, the main section (honpen) and iroha-jishu of the Racuyoxu are clearly and wholly based on Japanese dictionaries like Setsuyoshu. Racuyoxu is quite similar with Setsuyoshu not only on the point that they both show how to read words written in Chinese characters by also representing them in the kana syllabary, but that they do not provide word meanings. Today, it is known that approximately 50 types of Setsuyoshu were handwritten or a in a few cases printed before the Racuyoxu and it has not been decided which of these was the source; however, we can often find the same readings in the Racuyoxu and several kinds of Setsuyoshu, which were rather not found in the Vocabulario. In Ex.9, the Vocabulario has the headword gueqixin, but no giacuxin, which is found in the Racuyoxu. We can find both readings in the separated parts of Ekirin-bon Setsuyoshu, which was one of the Setsuyoshu printed in 1597 by Ekirin.

Ex.9: giacuxin/gueqixin [rebellious subject]
Racuyoxu:
逆臣 giacuxin
Ekirin-bon Setsuyoshu:
逆臣 giacuxin [in the headwords of Qi/Gui]
逆臣 gueqixin [in the headwords of Qe/Gue]
Vocabulario:
Gueqixin. Sacaximano xinca. [rebellious subject] Criado traidor, ou aleuantado. [rebellious subject, or traitor]

Ex.10 is an example of xôxin and xôjin, which demonstrates that we find xôxin in Racuyoxu, and that the Vocabulario says that xôjin is better than xôxin, and Ekirin-bon Setsuyoshu includes both readings in the separated sections inside. This seems to indicate that the word had two pronunciations in Japan and that the choices of their respective editors differed between the Racuyoxu and Vocabulario.

Ex.10: xôxin/xôjin [promotion, to be promoted]
Racuyoxu:
昇進 xôxin
Ekirin-bon Setsuyoshu:
昇進 xôxin [in the headwords of Xi/li (=Ji)]
昇進 xôjin [in the headwords of Xe/le (=Je)]
Vocabulario:
Xôxin. l, potiù s, Xô jin. [Xôxin. or, xôjin is better] Nobori susumu. [to climb and advance] O subir a dignidade, &c. [to rise to a higher rank, etc.] Vt, Curaini xôxin suru. [for example, to rise to a higher rank]

Furthermore, we should discuss the editors. Morita (1993) shows that the editors of the Vocabulario may have referred not only to the Racuyoxu but directly to the Setsuyoshu; however, it is clear that they did not contest the readings of several words. There are two possibilities with regard to the different readings: one is that the editors of the Vocabulario did not notice the different readings between their own knowledge and the Racuyoxu, and the other is that they noticed the difference but disagreed with the Racuyoxu. Thus, the evidence leads us to believe that that the two dictionaries had different editors.
It has also recently become clear that the *Vocabulario* and the previous Latin-Portuguese-Japanese dictionary the *Dictionarium* (1595) were edited by quite different Jesuit editors. It is true that the two dictionaries had basically different targets: The *Dictionarium* was for learners of both Latin and Japanese while the *Vocabulario* was for learners of Japanese. However, Kishimoto (2012, 2015) indicates the difference in usage of several Japanese words: the *Dictionarium* includes several words that the *Vocabulario* later notes as dialect or less proper. This is further evidence that each Jesuits’ dictionary had different sources and editors.

3. Conclusion

The inconsistency of readings between the *Racuyoxu* and *Vocabulario* was caused by their differing sources and editors. The editors of each dictionary should have used common materials including books and native speakers; however, each work presents the originality of its choice and usage of Japanese words. Moreover, from the viewpoint of Japanese language, the differences between these dictionaries reflect the fact that the Jesuits were faced with variety and change in the pronunciation of Japanese words, which may have been difficult for learners.

Acknowledgements

This work is supported by JSPS KAKENHI Grant Numbers 23320093, 24720209, 15K02573.

References

A. Primary sources

B. Secondary sources

A Study of The Prototypical Definition of the “walk” Headword
Based on the Characteristic of its Non Self-Sufficiency

Feng Haixia, Meng Wanjing
Ludong University, China
haixia872@163.com

Abstract
A result from current semantics research leads to a prototype semantic concept (prototypical definition): Firstly the definition can give the exact explanation of the prototypical sememe (center sememe). Secondly it also gives the explanation of the minor sememe, which adds or subtracts the most distinctive semantic features by offering the exact explanation of the sememe. As far as a semantic field is concerned, the prototypical sememe is a basic categorical term, it is a “basic level term”. A verb possesses characteristics of non-self-sufficiency; this can open a series of vacancies around the verb core requiring at least one to be filled; this completes the phrase that gives the verb an understandable meaning. With the support of semantic theories, we define the verb “walk” in a systematic and prototypical styled definition in the Language dictionary.

Keywords: language lexicography, prototypical definition, verb, non-Self-Sufficiency

1. The theoretical base

1.1 The verb and self-sufficiency
A pair of semantic concepts, self-sufficiency or non-sufficiency, was developed by Aristotle. In his book “Predicamenta” (Categories), he divided semantics into entity, amount, property, relation, place, time, posture, state, activity and sustainability. In the ten categories, “entity” refers to “human”, “horse” nouns. Human cognition itself is this category, which is included in the original. It is also the basis for the balance. The other nine are classified as “non-original” categories. They exist within the entity or they are dependent on it. According to Aristotle’s theory, the verb or the other words can be classified as noumenons. “If all the relative matter is correctly defined, it must connect and lead to that which it is related” or “the related exists at the same time” (Aristotle, 2003: 26-27). As it is shown, Aristotle’s main meaning is that noun belongs to the original category in the real world, therefore, a noun is “auto-sufficient” or “self-sufficient” while the verb or the other words must depend on “the related”—the original category, so it is “a non-original category” or “non-self-sufficient”.

A verb that is non-self-sufficient is always the opposite of a noun that is self-sufficient. It is a flawed opinion that Aristotle attributed all the nouns within the self-sufficiency category.

The verb sememe of non-self-sufficiency is one expression of the internal dependence of the semantic structure.

A semantics group from Moscow (Zhang Jiahua, 2003: 60) once said that “all semantic units can divide into two types, independent and dependent (variables)”. This group has pointed out that the unit of predicative semantics reflects the property of the surrounding participants and of their relationship as well as the related event. The scenic participants build up the unit of the predicative semantics which depicts the factor of the period. It not only refers to the individual but also includes the time, place, behavior and attributed behavior surrounding the variable. These participants can be divided into the two types: the necessary and unnecessary. One of the particular group must be a necessary condition of the particular situation. Every item varies from the change and breakup. These necessary participants are the explanation of
the related predicative semantics in the dictionary and are the correspondent with the semantic variables (abstract semantic parameter), named as the predicative semantic coordination valence. For example, the predicative verb “to nail” is explained in the dictionary as follows “put the nail or screw in a set position.” In the Modern Chinese Dictionary (MCD) “oder or screw” or “screwinen,” “the particular position” is equal to the semantic variables meanwhile are correspondent with the scenic method, object and location. If the explanation of the metalanguage in the dictionary is regarded as the perfect semantics in the verbal unit, the metalanguage explanation of the predicative semantic unit should be equal to the semantic structure of the predicative semantic unit. From that, we can draw an easy conclusion that the semantics group from Russia has inherited and developed traditionally the study of the verbs and its components. They thought that the verb sememe not only is the superficial structure, but also the meaning is in the deeper structure, which includes the completed and compulsory requirements of the necessary setting. It is the typical expression of the verbal semantic unit.

In the 1960s C.J. Fillmore had discovered that the most important aspect was the semantic roles of the verb. In the 70s or 80s, he pointed out that one word defined one scene, which stood out or emphasized one part in the scene, the various elements took the various semantic roles, meanwhile the verb was typically regarded as the related semantics within the internal components in different ways. These verbs in different ways “(evoke)” situation. For instance, a “commercial event” is the elements of the picture scene, which include the buyer who is interested in using money in exchange of goods, the seller who is interested in exchanging goods for money as well as goods and money. By making the full use of these elements, the verb “buy” emphasizes the buyer’s commercial manner and at the same time is the background of “seller” and “money”. However, the verb “sell” is the opposite. The verb “pay” emphasizes the relationships: the buyer/money/seller, with goods as the background. Therefore, Fillmore (2003: 382-412) clearly pointed out that “one only understands the verbs’ meaning by reliance on the situation (our understanding is obtained by the differing parts and transactions between buyers and sellers, so this scene is necessary background and incentive)...the frame structured the semantics, while the word semantics ‘evoke’ the frame”. It is shown in Fillmore frame theory that the meaning of the verb is the analysis of processing. That is to say, the process of the implied semantic roles of the verbal sememe is revealed.

Whether Андреян is the semantic valence which is the reprehensive of modern semantics “scene” in Russia or Fillmore is the top of reprehensive in the semantic frame. There is no exception. It proves that the verb is not self-sufficient. The lacking self-sufficiency of the verb opens a series of vacancies in the verbal kernel, which fills in at least one entity and constitutes the complete meaning in the verb coordinate valence.

1.2 The systematic perspective of Lexicon

Language scholars have generally regarded language as having a systematic nature since Saussure. But in the first half of the 20th century, the systematic nature of grammar and pronunciation was popular in the linguistic field, they believed that lexicon was disunited one. The theory of lexical systematics did not come up until the sixties when it was questioned severely. It was gradually accepted through the eighties.

Paraphrasing is a dependent variable of the lexical systematic nature. The compiling dictionary’s thought of the traditional puts a nuclear concept at the top while the contemporary put the whole as the domination. This shift results from the new trend of linguistic philosophy, linguistics, and psychology in the early 20th century. Modern middle-sized dictionaries were born with this new trend: “Concise Oxford English Dictionary (COED)” and “Le Petit Robert French Dictionary, Russian Dictionary and MCD”. However, “the related item” on compiling modern Chinese dictionary only represents the early shape of the whole, which is far away from its systematic structure, fullness, development and perfection.” (Zhang Zhiyi, 2012: 335)
Zhang Zhiyi and Zhang Qingyun (2012: 340) has developed the theoretical instruction of compiling Chinese dictionary from the whole view of semantic field by absorbing the essence of modern semantics. They clearly pointed out the theory of “the same field with the model”. That is to say, “the item in the same semantic field should be explained in the same model”. Xia Lixin (2007) also believed that the entry word in the same lexical set should be handled unified. He made a comparison between the Chinese Learning Dictionary and the Longman Contemporary English Dictionary and discovered some shortcomings. At the same time he suggested that the assumption and the measure of the paraphrasing model be based on glossary (semantic category). That is to say, the systematic quality of the sememe be broken up by its alphabetical order and built up the small unit of the polymerization of semantic category, in addition, realized the whole systematic nature of the dictionary and provided the supportive theory.

From the perspective of whole system theory, a sememe belongs to the whole language, therefore, one sememe (even an individual one can be correctly understood and explained. It is impossible to view the sememe in isolation but to observe the location of the meaning system, this regulates the nature of the individual. It is an emphasis that we view the relationship from the internal language, which includes the sememe valiancy. “The concept of “difference” gives us implications. When we paraphrase the sememe in a typical group, we go through the related theory of the systematic nature of language. That is to say, when we set up the meaning, there is a standard of the accepted or rejected choice among the relationship of the group in the sememe. A feature of semantic analysis is to depict selectively. It also is the tool whether this systematic characteristic of paraphrasing is strong or weak. An example is the listing for “bear” in the Modern Chinese Dictionary can be paraphrased into “bear” “black bear” “brown bear” “polar bear”

- **bear mammal**, it has a large head, short tail with four short and sturdy legs, can climb a tree and mainly eat animal food, sometimes eat fruits and nuts. There are several kinds of bears, for example, a brown bear or black bear. Some places called “hei xia zi”.

- **black bear** mammal, it has a huge body, short tail, big palms and sharp claws with the white stripes of the new moon shape in the chest, the rest part of the body is black, has a loneliness character, can swim and climb a tree, and even walk, named as gou xiu. Some places called “heixiazi”.

- **a brown bear** mammal, it has a huge body, shoulder uplift, brown color varies from the different areas, can swim and climb a tree, eat nuts, vegetables, worms, fish, birds, or animals sometimes can hurt people and livestock.

- **a polar bear** mammal, it has white thick fur with some yellow fur, has a black nose and black claws, lives in the north frigid zone and is good at swimming, named as a white bear.

It is shown that there are four kinds of the explanations in the MCD, but it is not possible to see that there is a relationship among them. The guess is the compiler was carefully describing every single entry.

1.3 The prototypical view of paraphrasing
One of the results in modern semantics leads to the prototypical semantic. Lyons (1995) believed that the prototypical view must pay attention to the application of the various words: natural words such as (dog, fish, tiger, lemon), cultural words (bachelor’s degree, tea cup, chair), verb, adjective, color.

The realization of the prototypical semantic view is that we select the center of the word (the prototype) in the meaning word, then develop the minor word, the common word and the marginal word. Finally, we can identify the semantic categories of several types and levels. Cognitive linguistics has made a new contribution to the study of semantics and the semantic field. The view of prototypical semantic brought out the prototypical definition. First is the correct explanation of the central sememe. The second is the minor sememe, which is equal to
the correct explanation of the prototypical sememe plus or minus of the feature of semantic property. In terms of the semantic field, the word lies in the basic category is the prototypical sememe, named as (Basic Level Terms) or the word of the basic concept, while the minor sememe of the word belongs to the up level of the meaning field or the down level of the meaning word.

“Bear” is a prototypical sememe of the semantic field. “black bear”, “brown bear” and “polar bear” belong to the down level entries, or are the non-central sememe of the semantic field, therefore, the later explanation is close to the up level word “bear” category, at the same time, choose the features of the semantic property. That is to say, the explanation of “bear” is based on the explanation of structure. The basic property of the explanation “bear” is that it has a big fat body, big palms, sharp claws and has a non-gregarious character, can climb treed while the non-different features are omitted. “A polar bear” has thick long fur, which feature is not differentiating, therefore, after sorting it out, we explain the followings:

black bear one kind of a bear has the white stripes of the new moon shape in the chest, the rest part of the body is black. It can walk, some places name it as hi xiaizi.
brown bear one kind of a bear has shoulder uplift and brown color fur, the fur’s color varies from the different places. It occasionally attacks human beings.
polar bear one kind of a bear has white fur with some yellow fur, has a nose and black claws, lives in the north frigid zone is a good swimmer, it can also be named as a white bear.

Therefore, the word entry in the same category has a systematic characteristic of the same model of the explanation and of the economy of the explanation of the dictionary, meanwhile it meets the cognitive regularity from easy to difficulty. Certainly, if we presume “a bear” correct explanation, the clarity of the explanation in the dictionary is close to the perfection in the integration of the explanation.

2. The inspection and optimization of the prototypical explanation of the “walk” entry based on it being non self-sufficient

In “the foot’s movement”, the “walk” shows the typical category. There are total 73 Chinese sememes and 95 English sememes in the “walk” entry.

The best model of the explanation is whether the sememe is located in the structure’s level, the explanation verbal select the prototypical sememe but as far as the “walk” integration category is concerned, the “walk” is selected. This is the best model to assess these dictionaries: MCD, COED, Longman Dictionary of Contemporary English (LDCE) and Collins COBUILD English Dictionary (CCED).

Table 1 the usage of the explanation of the verbal core meaning “walk” in the MCD

<table>
<thead>
<tr>
<th>Core verb</th>
<th>走</th>
<th>行</th>
<th>步</th>
<th>迈</th>
<th>踱</th>
<th>踉跄</th>
<th>踏</th>
<th>迈进</th>
<th>其他</th>
</tr>
</thead>
<tbody>
<tr>
<td>Times</td>
<td>22</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>17</td>
</tr>
</tbody>
</table>

In Table 1, there are 22 explanations of the sememe, which use the verbal core meaning, taking account 33.3%. Here it includes the variants of “walk” such as: 走走, 走过去, 走开. In
the table, the other refers to the synonym of the verbal core meaning or the up level verbs such as: 闲逛, 游逛, 散步, 趟, 徘徊.

Table 2 The usage of the explanation of the verbal core meaning “walk” in the COED

<table>
<thead>
<tr>
<th>Core verb</th>
<th>walk</th>
<th>move</th>
<th>others</th>
</tr>
</thead>
<tbody>
<tr>
<td>times</td>
<td>40</td>
<td>25</td>
<td>15</td>
</tr>
</tbody>
</table>

In table 2, there are 40 prototypes of the sememe in the “walk”, taking up 50%. In the up level words “walk”, “move” words take up 31.25%. The other words are selected such as synonym or up level words “go, come”.

Table 3 Usage of the explanation of the verbal core meaning “walk” in the LDCE

<table>
<thead>
<tr>
<th>Core verb</th>
<th>walk</th>
<th>move</th>
<th>others</th>
</tr>
</thead>
<tbody>
<tr>
<td>times</td>
<td>50</td>
<td>13</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 4 Usage of the explanation of the verbal core meaning “walk” in the CCED

<table>
<thead>
<tr>
<th>Core verb</th>
<th>walk</th>
<th>move</th>
<th>others</th>
</tr>
</thead>
<tbody>
<tr>
<td>times</td>
<td>40</td>
<td>17</td>
<td>14</td>
</tr>
</tbody>
</table>

From table 3, it is shown that 75.76% sememes in the “walk” category (LDCE) use the model as “differentiated factor + prototypical sememe (walk)” (CCED).

We have easily discovered that there are some problems among these digital number: 33.33% < 50% < 56.34% < 75.76%. In terms of a foreign language contrast to MCD, the “walk” category in the explanation needs optimizing. We can find out that each dictionary is not systematically controlled according to the range. There are several reasons for this:

First is a subjective reason, there are numerous sememes in the “walk” category. They do not only refer to the action and state of the person or animal, that is to say, this verb has another move except ‘walk’, for this part of the sememe, we must accept them within “walk” category. The lack of clarity and economy of a language enriches the sememe value, but it still cannot set up independent sememes, therefore, this word’s core meaning of the explanation cannot be selected just in the area “walk”, it needs the up level and a good generation of the verbal core. In one word, the complexity of the verbal sememe leads to the non-systematic nature of the core meaning, for instance:

迈进 大踏步前进。 (MCD)

slink move quietly with gliding steps, in a stealthy or sensuous manner. (COED)

trickle if people, vehicles, goods things trickle, they are moving slowly in small groups or amounts. (COED)

The second is the structure hierarchy and explaining expressiveness of a language. The “walk” category is based on the “walk category of the core and prototypical polymerization together. The “walk” is not located in the same level in MCD. In this typical category, these factors “speed factor”, “state factor”, “place factor”, “pathway factor” are built up and borrow from the smaller aggregation group. There are several examples: 徘徊, 徘徊, 盘桓, 盘旋, 低回踌躇, 彷徨, 蹑躚…”散步”it is the typical group of the bottom level, it includes: 信步, 溜达, 遛弯儿, 串游, 绕弯儿, 遛, 遛早儿, 闲逛, 逛… corresponding with English “walk” category, for example, according to the fast and slow speed, it can be divided into “speed” or “creep” the two of the aggregation group of the bottom level, and the action’s aimlessness can be categorized into the “wander” aggregation group. According to the different states, it can be categorized into “wander” aggregation group.

Facing this case, different Chinese dictionaries adopt different ways of the explanation, which mainly exemplifies the prototypical sememe of the typical group—or handles differently
“walk”. It belongs to the “up level word” or synonym. The ideal one is based on the explaining model of the category system (exception for the comprehensive way “synonym + labelled minor meaning”). Most non prototypical sememe are around the core meaning verb. The choice and prominence of the differentiated factor is clearly built on the comparison of the small typical group of sememe. For instance:

徘徊 ①在一个地方来回地走。
彷徨 ①在一个地方来回地走，不知往那个方向去;比喻犹豫不决。(这个释义是我们在照《现汉》调整的，参照了多本语文词典和同义词典，以及语料库)
遛 ①牵着牲畜或带着鸟慢慢走。
闲逛 闲暇时到外面随便走。
游逛 为消遣而闲走。
漫步 没有目的而悠闲地走。

waddle  walk with short steps and a clumsy swaying motion. (COED)
waddling walk with short steps, with your body moving from one side to another - used especially about people or birds (ducks) with fat bodies and short legs. (LDCE)
walk there with short, quick steps, swinging slightly from side to side. A person or animal that waddles usually has short legs and a fat body. (CCED)
hobble 1. walk awkwardly, typically because of pain. (COED)
wander with difficulty, especially because your legs or feet hurt. (LDCE)
if you hobble, you walk in an awkward way with small steps, for example because your foot is injured. (CCED)
limp walk with difficulty because of an injured leg or foot. (COED)
1. to walk slowly and with difficulty because one leg is hurt or injured. (LDCE)
if a person or animal limps, they walk with difficulty or in an uneven way because one of their legs or feet is hurt. (LCED)

In practice, the “walk” sememe of different levels is selected by the verbal core meaning. At most cases, we give up the ideal systematic model of the explanation in the MCD we select the up level verbal sememe or the core meaning of synonym. For instance:

迈 ①提脚向前走；跨。
跨 ①抬起一只脚向前或向左右迈(一大步)。
散步 随便走走。
遛早儿 早晨散步。
游玩 为消遣而闲走。
浪游 漫无目标地到处游逛。

To comparison with Chinese Dictionaries, there are less similar cases in the English dictionaries:
march 1. walk in a military manner with a regular measured tread. (COED)
defile (of troops) march in single file. (COED)
wander If you wander in a place, you walk around there in a casual way, often without intending to go in any particular direction. (CCED)
roam around it. (CCED)
if you roam an area or roam around it, you wander or travel around it without having a particular purpose. (CCED)

LCED or CCED. These two dictionaries must have correlated relation in the quantitative limit of the explaining metalanguage. 2000 explaining words are limited in the LDCE while 3500 or so in CCED. The quantitative control of the explaining metalanguage makes a lot of useful words not into the explaining words. To some extent, this limit blocks the possibility of coming into the explanation of the excursiveness, so that promotes the development of a high degree of unity of systematic explanation. For instance, we can see the choices of the verbal core meaning in all sememes of the explanation in “walk” category (LCED), non-meanings of “walk” and “move” take into account 95.49%. If we see the point from the view of discrimination, the explanation of the excursiveness damages the regularity and readability,
but it has some advantages: when the core meaning is close to the up level word or synonym, the word is easily and clearly be described. Taking the above “遛早儿” for an example, supposed that we must take the prototypical verbal meaning “walk” as an explanation, it is occurred these following’s explanation: “早晨随便走走”. Therefore, it is clear but redundant. If we try to choose closely the up level or synonym, the result is contrary to what we expect. If a reader cannot understand “游逛” the exact meaning, how can he or she even understand “浪游”? If we all use the meaning of “走”, we have to double the explanation. Therefore, the explaining metalanguage is a two-edged sword, whether we choose the categorical explanation with a high degree of uniformity or we choose the explanation of the excursiveness. We would better design the dictionary according to the character and size of the dictionary.

The last is due to codifiers’ view of the systematic explanation. Making an analytic comparison between COD and CCED, we have discovered that there are 24 verbal sememes in COD—no meanings of “walk”. In the differentiated factor, they do not change the verbal sememes, and the verbal sememe can completely replace the “walk”. In addition, we also have discovered the adjusting location of the core meaning of“行走, 步行, 走动”. We also believe that the codifiers pay less attention to the systematic explanation to a great extent, while paying less attention to the view of the prototypical explanation; therefore, they do not use the same model of the explaining sememe at the same level and in the same polymerization.

According to the above analysis, we can find out that the ideal explaining model of “walk” category should carefully describe “walk”, just like as a mirror description, while the another sememe is around the another category “walk”, it should feature the differentiate factors to some degree and with some aspects: such as aim factor, direction factor, state factor (speed, way), time factor, place factor as well as reason factor.

![Figure 1 Model of “walk” sememe explanation of category](image)

**Acknowledgments**

This work was supported by the National Social Science Foundation of China (No. 13FYY012)

**References**


A Comparative Study of War Metaphors in Business and General News Articles Using a Dictionary-based Database

Zhong Xian
Lingnan University
xianzhong@ln.hk / okpidia@hotmail.com

Abstract
“Metalude is an interactive database of conventional, lexicalized metaphors” (Goatly, 2002-2005). It has a fine categorization of lexis which has both literal and metaphorical meanings recorded in contemporary English dictionaries under different conceptual metaphors. The Metalude data supports the idea that activity in general, not just business activity, is pervasively understood as war, games, sailing. And that organizations and systems in general, not just the economy, is conceptualized as a machine, or a human body which is more or less healthy. These metaphors can be used in business discourse and general discourse to talk about business activity and activity in general. It thus becomes a question I investigated in this research, focusing on war metaphors, whether the war metaphors were used more in business discourse to talk about business activities, than they were in other texts to refer to activity in general. The study also attempted to compare the findings across English and Chinese.

Keywords: Metalude, war metaphor, business corpora, general corpora

1. Introduction

Metaphor is pervasive in various business-related materials and situations, including economics and financial news articles (Charteris-Black 2004; Koller 2004), business encounters (Handford and Koester, 2010), business research articles (Skorczynska and Deignan 2006) and books introducing organizational theories (Morgan, 1997; Weick 1979, 2005). Among the various types, war metaphor is found to be dominant in business discourse and it is used to conceptualize business activity as if with intense conflicts. Business life is conceptualized as a battlefield with intense conflicts. Corporate strategists prefer to use war as metaphor to translate tactics on the battlefield into policy for the boardroom (Peattie 1990). War metaphors constitute the third major metaphor cluster found in Charteris-Black and Musolff’s (2003) comparative study of metaphor for euro trading in British financial reporting. They also found that war metaphors in English could carry evaluative force, such as the adjective battered which personified the euro as a victim of physical combat. The successive use of war metaphor could achieve intensification and hyperbole. Many of the war metaphors were also found in Boers and Demecheleer (1997: 125-127) in their analysis of metaphors in economic discourse using corpora of English, French and Dutch as well as Charteris-Black and Ennis (2001: 256). Herrera and White (2000) summarize the steps for the communication of business acquisitions as challenge, open conflict and victory and subsume them under the conceptual metaphor BUSINESS IS WAR.

Based on a corpus of magazine and newspaper articles on marketing and sales, and another on mergers and acquisitions, Koller’s (2004) analysis shows the metaphor cluster of WAR/FIGHTING is the most prominent one in the discourses. It is highly masculinized as men play a constitutive role in the aggression and conflicts. The use of metaphors of aggression also tend to influence the harmony and the team spirits within corporations, as the individuals are motivated to engage in competitions and even conflicts within an environment conceptualized metaphorically as being highly aggressive. War metaphor is thus used to construct the business setting as a fighting scenario where the market forces are represented as
fighters. War metaphor also carries with it the value-judgment towards the entity or event referred to.

Though the conceptualization of business life as a battlefield with intense conflicts may be widespread across languages and cultures, subtle differences exist under this conceptual metaphor. The linguistic realizations of the war metaphor demonstrate that war itself is not a uniform domain, comprising a blend of both physical violence and military strategy (Koller 2004). In Alousque’s (2011) exploration of the ideological stance in metaphors used in the British and Spanish newspaper articles reporting the Endesa takeover, she finds that although the same detailed metaphorical mappings can be found in English and Spanish, there are more linguistic expressions as instantiations of the conceptual war metaphor in Spanish. Besides, the war metaphorical expressions in the Spanish news coverage emphasize the tough characters of Endesa’s CEO and highlight the strategy deployed by rival companies, while the war metaphorical terms in the English describe the process of takeover and the attitude of the competitors as hostile. Even though the conceptual metaphor of war is pervasive in economic and business discourse and it activates the same metaphorical mappings in different languages, the different instantiations and their uses in different languages can highlight selective aspects of the source of war and convey different ideological stance and implications to the readers.

While war metaphor is ubiquitous in business and economic discourse, the linguistic realizations of the war metaphor can be found in the database of Metalude, which is an interactive database of conventional, lexicalized metaphors” (Goatly 2002-2005). The database has a fine categorization of lexis which has both literal and metaphorical meanings recorded in contemporary English dictionaries under different conceptual metaphors. The Metalude data supports the idea that activity in general, not just business activity, is pervasively understood as war. The Metalude data shows that war metaphors can be used in business discourse and general discourse to talk about business activity and activity in general. So this paper aims to investigate whether the war metaphors are used more in business discourse to talk about business activities, than they are in other texts to refer to activity in general.

2. Data and method

The study adopts a corpus-driven approach. Two corpora of English and Chinese business news articles were compiled respectively. The English business corpus consists of business news articles taken from the American daily newspaper The New York Times published from 2008 to 2010. The Chinese business corpus is comprised of business news articles taken from Chinese daily newspaper Nanfang Daily published from the same period of time. The New York Times and Nanfang Daily were chosen as they share a lot of similarities. Both of them have a reputation for fair and accurate modern reporting. They are broadsheet daily papers originated from New York City and Guangzhou which have a significant influence on national commerce and finance. They have a clear Business section. Both newspapers have a similar daily circulation and according to the statistics presented above, they tend to attract readers who are of similar age, with higher educational background, business-oriented and with median to high income. The equivalence between the newspapers in terms of reputation, readership and style makes possible the compilation of two equivalent business corpora with business news reports from the two newspapers.

News reports from the New York Times and Nanfang Daily were extracted from the databases of Proquest and Wisenews respectively. The New York Times corpus has 2 million words, while the Nanfang Daily corpus has 3.2 million words. Both databases offer search engines for readers to search for specific articles. Articles in the business section were searched for and randomly selected every month during the time range between 2008 and 2010. The Wordsmith Tools were used to generate a word list of the English corpus and
lexical items denoting aggressive and hostile actions in war were identified from the list. The lexical items selected were manually examined in their concordance lines to find whether there was a contrast between the contextual meanings and their basic war-related senses. The software of Antconc was used to generate a word list of the Chinese corpus and words denoting aggressive actions in war were identified from the list. The war-related terms identified were manually examined in their concordance lines to see whether it was used metaphorically or not. The English and Chinese words which were confirmed to be war metaphors were noted down for further comparison with their occurrences in the reference corpora.

The general corpus in English is part of the Corpus of Contemporary American English (COCA) which was created by Mark Davies of Brigham Young University and contains more than 450 million words of text. The texts in the corpus are well classified according to the topics of spoken, fiction, popular magazines, newspapers and academic. Below the umbrella categorizations of the genres of spoken, fiction, popular magazines, newspapers and academic, there are specific divisions under each of the five general categories. There are, for instance, spoken texts from the broadcast ABC, NBC, CBS or CNN and articles in the magazine are divided among the topics of News/Opinion, Financial, Science/Technology, Society/Arts, Religion, Sports, Entertain, Home/Health, Women/Men, African-American, and Children. The study thus excluded texts from the sub-divisions of magazine and newspapers which had to do with money and finance and focused on war-manifested lexical items in texts from other discourse fields. The texts published between 2009 and 2010 were chosen and therefore covered a period similar to that by the business news articles in the self-compiled corpus.

The total number of words in the sub-genres of newspaper from 2009 to 2010 is around 7.37 million. Among the 7.37 million words, the concordance lines containing the war-related lexical items were generated and all manually analyzed. The metaphorical frequency analyzed from the concordance lines represented the metaphorical occurrences of elicited war-related lexical items in the sub-genres of newspaper. The metaphorical frequencies were used to compare with those in the English business corpus. Later, a statistical analysis was applied with the assistance of the SPSS program to investigate whether the frequency counts of war metaphors in the business corpus are more or less when compared with those of war metaphors in the general corpus. After the list of war metaphors found in the business corpus were searched for their uses in the general corpus and their frequency counts in the general corpus were noted down, the two sets of frequency counts, including the occurrences in the business corpus and the general corpus, were typed into the program SPSS and a paired samples test was conducted. If the significance level of the two-tailed test equals or is smaller than 0.05, then the significance level indicates that there is a statistically reliable difference between the two sets of frequency counts. Moreover, I’m interested in whether the use of war metaphors in the business corpus is more often or less often than that in the general corpus, so a one-sample t-test was also conducted by the SPSS program and a threshold test value was determined.

For the Chinese corpus, a section of the general reference corpus of CCL was used. CCL contains 477 million characters and in particular, the study made use of the newspaper reports within the corpus. The Xinhua News Agency reports published in 2004, amounting to 47 million characters, were taken from the CCL. After excluding the reports about business and economic activities from all the news reports, a reference corpus was compiled with the rest of the reports with 41.7 million characters. The collocations and compounds of war lexis found in the business corpus were searched in the reference corpus. Corresponding concordance lines were later generated for identification of the metaphorical uses of these collocations and war compounds. The metaphorical frequencies were used to compare with those in the Chinese business corpus. Later, a statistical analysis was applied with the assistance of the SPSS program to investigate whether the frequency counts of war metaphors
in the business corpus are more or less when compared with those of war metaphors in the
general corpus. After the list of war metaphors found in the business corpus were searched for
their uses in the general corpus and their frequency counts in the general corpus were noted
down, the two sets of frequency counts, including the occurrences in the business corpus and
the general corpus, were typed into the program SPSS and a paired samples test was
conducted. If the significance level of the two-tailed test equals or is smaller than 0.05, then
the significance level indicates that there is a statistically reliable difference between the two
sets of frequency counts.

3. Findings

3.1 Statistical comparison between the English business corpus and general corpus
Table 1 compares the frequencies per 1 million words of war and fighting lexical items used
metaphorically in the general and business corpora. The lexical items listed in the table
include word families, such as the occurrences of combat, include combats, combating, combated and combative as well.

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
<th>Adjective/adverb</th>
<th>Business tokens</th>
<th>General tokens</th>
<th>Ratio of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>retreat</td>
<td>to retreat</td>
<td>-</td>
<td>36</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>inroad</td>
<td>-</td>
<td>-</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>cohort</td>
<td>-</td>
<td>-</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>-</td>
<td>to pummel</td>
<td>-</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>-</td>
<td>to grapple</td>
<td>-</td>
<td>33</td>
<td>7</td>
<td>4.71</td>
</tr>
<tr>
<td>aggression</td>
<td>aggressive,</td>
<td>-</td>
<td>133</td>
<td>32</td>
<td>4.16</td>
</tr>
<tr>
<td></td>
<td>aggressively</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>resistance</td>
<td>to resist</td>
<td>-</td>
<td>53</td>
<td>13</td>
<td>4.08</td>
</tr>
<tr>
<td>assault</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>troop</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>invasion</td>
<td>to invade</td>
<td>-</td>
<td>7</td>
<td>2</td>
<td>3.5</td>
</tr>
<tr>
<td>blow</td>
<td>-</td>
<td>-</td>
<td>40</td>
<td>12</td>
<td>3.33</td>
</tr>
<tr>
<td>-</td>
<td>to batter</td>
<td>-</td>
<td>12</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>firepower</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>skirmish</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>hostility</td>
<td>hostile</td>
<td>-</td>
<td>11</td>
<td>4</td>
<td>2.75</td>
</tr>
<tr>
<td>gun</td>
<td>-</td>
<td>-</td>
<td>8</td>
<td>3</td>
<td>2.67</td>
</tr>
<tr>
<td>-</td>
<td>brutal</td>
<td>-</td>
<td>13</td>
<td>5</td>
<td>2.6</td>
</tr>
<tr>
<td>combat</td>
<td>to combat</td>
<td>combative</td>
<td>30</td>
<td>12</td>
<td>2.5</td>
</tr>
<tr>
<td>foray</td>
<td>-</td>
<td>-</td>
<td>7</td>
<td>3</td>
<td>2.33</td>
</tr>
<tr>
<td>under fire</td>
<td>-</td>
<td>-</td>
<td>7</td>
<td>3</td>
<td>2.33</td>
</tr>
<tr>
<td>-</td>
<td>beleaguered</td>
<td>-</td>
<td>7</td>
<td>3</td>
<td>2.33</td>
</tr>
<tr>
<td>bullet</td>
<td>-</td>
<td>-</td>
<td>9</td>
<td>4</td>
<td>2.25</td>
</tr>
<tr>
<td>hit</td>
<td>to hit</td>
<td>-</td>
<td>144</td>
<td>72</td>
<td>2</td>
</tr>
<tr>
<td>arms</td>
<td>to arm</td>
<td>armed</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>-</td>
<td>to safeguard</td>
<td>-</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>arsenal</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>siege</td>
<td>to besiege</td>
<td>-</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>-</td>
<td>to capitulate</td>
<td>-</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>tussle</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Action</td>
<td>Frequency</td>
<td>TPMI</td>
<td>CFRI</td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>----------</td>
<td>-----------</td>
<td>------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>to snipe</td>
<td>-</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>to torpedo</td>
<td>-</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>struggle</td>
<td>to struggle</td>
<td>227</td>
<td>127</td>
<td>1.79</td>
<td></td>
</tr>
<tr>
<td>ammunition</td>
<td>-</td>
<td>5</td>
<td>3</td>
<td>1.67</td>
<td></td>
</tr>
<tr>
<td>challenge, challenger</td>
<td>to challenge</td>
<td>236</td>
<td>145</td>
<td>1.63</td>
<td></td>
</tr>
<tr>
<td>struggle</td>
<td>-</td>
<td>50</td>
<td>32</td>
<td>1.56</td>
<td></td>
</tr>
<tr>
<td>casualty</td>
<td>-</td>
<td>3</td>
<td>2</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>fierce</td>
<td>18</td>
<td>13</td>
<td>1.38</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>to beat</td>
<td>67</td>
<td>50</td>
<td>1.34</td>
<td></td>
</tr>
<tr>
<td>camp</td>
<td>-</td>
<td>8</td>
<td>6</td>
<td>1.33</td>
<td></td>
</tr>
<tr>
<td>deployment</td>
<td>to deploy</td>
<td>9</td>
<td>7</td>
<td>1.26</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>embattled</td>
<td>5</td>
<td>4</td>
<td>1.25</td>
<td></td>
</tr>
<tr>
<td>tactics</td>
<td>-</td>
<td>15</td>
<td>12</td>
<td>1.25</td>
<td></td>
</tr>
<tr>
<td>fight, fighter</td>
<td>to fight</td>
<td>134</td>
<td>109</td>
<td>1.23</td>
<td></td>
</tr>
<tr>
<td>defense</td>
<td>to defend</td>
<td>50</td>
<td>43</td>
<td>1.16</td>
<td></td>
</tr>
<tr>
<td>punch</td>
<td>to punch</td>
<td>-</td>
<td>11</td>
<td>10</td>
<td>1.1</td>
</tr>
<tr>
<td>campaign, campaign</td>
<td>to campaign</td>
<td>-</td>
<td>274</td>
<td>255</td>
<td>1.07</td>
</tr>
<tr>
<td>strike</td>
<td>to strike</td>
<td>striking</td>
<td>46</td>
<td>44</td>
<td>1.05</td>
</tr>
<tr>
<td>weapon</td>
<td>-</td>
<td>10</td>
<td>10</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>enemy</td>
<td>-</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>bomb, bombshell</td>
<td>to bombard</td>
<td>-</td>
<td>6</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>-</td>
<td>to surrender</td>
<td>-</td>
<td>6</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>-</td>
<td>to spearhead</td>
<td>-</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>-</td>
<td>to lash</td>
<td>-</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>onslaught</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>brigade</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>to defuse</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>frontline</td>
<td>-</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>minefield</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>truce</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>to wield</td>
<td>-</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>-</td>
<td>to kill</td>
<td>18</td>
<td>18</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>battle,battlefield</td>
<td>to battle</td>
<td>78</td>
<td>79</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>attack, attacker</td>
<td>to attack</td>
<td>67</td>
<td>70</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>blitz</td>
<td>-</td>
<td>5</td>
<td>6</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>war, warfare, warrior</td>
<td>to war</td>
<td>warring</td>
<td>40</td>
<td>50</td>
<td>0.8</td>
</tr>
<tr>
<td>crusade</td>
<td>-</td>
<td>3</td>
<td>4</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>army</td>
<td>-</td>
<td>6</td>
<td>10</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>to enlist</td>
<td>-</td>
<td>4</td>
<td>7</td>
<td>0.57</td>
</tr>
<tr>
<td>shot</td>
<td>to shoot</td>
<td>33</td>
<td>63</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>barrage</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>stronghold</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>sword</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>
Table 1 shows that a majority of the war metaphorical lexical items are used slightly more frequently per million words in the business corpus than the general corpus. The number of occurrences for a metaphorical lexical item does not vary much between the business corpus and the general corpus and the biggest difference is 32 times per million words for the lexical item *retreat*. Another clear observation of Table 1 is that the use of descriptive adjectives in the business corpus is more frequent than in the general corpus. The adjectives include *brutal, fierce,* and *beleaguered*. Other lexical items among the majority are used slightly more often in the business corpus and their frequencies of occurrences in the business corpus and the general corpus are quite close. It also shows that war metaphors are well exploited in news reports.

The two sets of frequency counts were typed into the SPSS program to see whether they were significantly different with statistical evidence. The paired samples T-test performed by SPSS generated the mean difference of 4.99 (SD= 26.61) and the two-tailed p-value of 0.096, which is close to 1 and greater than the specified significance level of 0.05. So we can conclude that there is not a statistically significant difference between the two sets of frequency counts (t=1.69, df=80, p=0.096), and it means the use of war metaphors in the business corpus is not significantly more or less than that in the general corpus. Furthermore, the ratio between the frequency counts for each word was also calculated and a one-sample t-test was conducted using SPSS. The test result shows that there is a statistically significant difference between the frequency of each word in the business corpus and in the general corpus (t=4.80, df=80, p<0.05), which indicates that each word differs significantly in terms of their uses in the business context and the general context. The inconsistent results of the paired samples t-test and the one-sample t-test may be due to the significant difference between the frequencies of a single word in the two corpora rather than the whole set of war-related lexis.

**3.2 Detailed comparison between the English business corpus and general corpus**

There are altogether sixteen lexemes and the derivatives used with similar frequency in the business and general corpus, with the ratio around 1:1 in the business and general corpus. Three of the sixteen lexemes, *kill, onslaught* and *attack*, are sorted under the subtype Attack.
in the subfield Actions and four, *wield, defuse, bomb* and *weapon*, belong to the subfield of Weaponry. The other lexemes *enemy, brigade, spearhead; surrender* and *truce; lash* and *battle; frontline* and *minefield* are sorted under the subfields of Forces, Result, Types of action, and Topography, respectively.

Table 3.1 shows that there are ninety-one war metaphorical types found in the business English corpus and eighty-one war metaphorical types in the general English corpus. Fifty-seven types occur less frequently as metaphors in the general corpus, though their frequencies of occurrence do not vary greatly with those in the business corpus. Sixty-three percent of the types are used more frequently in the business English corpus than in the general English corpus. The lexeme *RETREAT* has the greatest frequency variance in the business and general corpus. Its ratio of frequency reaches 9:1 in the business and general corpus. The lexemes having relatively less frequency variance than *RETREAT* include *INROAD, PUMMEL* and *COHORT*. All of the three lexemes occur five times in the English business corpus as metaphors and one time in the general English corpus, which suggests that they are not frequently used as metaphors in both corpora. The lexeme *GRAPPLE* registers a frequency of 33 in the business English corpus and 7 in the general English corpus. So in the following, I will analyze some selected words as examples to demonstrate the differences.

The use of derivatives of *AGGRESS* is more frequent in the business corpus than in the general corpus, reaching the ratio of 4.1:1. The disparity in terms of frequency of use suggests that derivatives of *AGGRESS* are much more frequently used as metaphors in the business corpus than in the general corpus. Typically, the adjective form registered frequent occurrences in the business corpus to describe the competitive business and financial activities or confident and assertive entities involved in business dealings. The first type of description presents the financial practices as “angry and threatening”, as shown in the following concordance lines:

1. Aid. Output dropped at a 1.1 percent rate in the previous period. *Aggressive* cost-cutting has pushed productivity
2. Pan, Britain enjoyed more than a decade of booming growth, fed by *aggressive* bank lending and real estate investments.
3. Anaging the risk of asset price bubbles as a result of last year’s *aggressive* expansion of credit. ”Stephen Green,

While the second type describes the individuals or companies involved in financial practices as “decisive and determined”, which are shown in the following concordance lines.

1. Rajaratnam since his days at Needham, said Mr. Rajaratnam was an *aggressive* buyer and seller of stocks. "His was
2. Forward to a safe and enjoyable shopping experience at Wal-Mart. "*Aggressive* shoppers are common the day after
3. Lenders are desperate to avoid the fate of Corus Bank shares, an *aggressive* condominium lender that federal regulators

The comparative and superlative forms of the adjective can be observed often in the business corpus to describe the more and most assertive entities in business dealings or the levels of competition that financial activities hold. However, the adjectival form is used far less often as an expression of fighting metaphor in the general corpus when compared to its counterpart in the business corpus. The pattern of using comparatives and superlative forms of the adjective can not be found in the general corpus. The metaphorical uses of *aggressive* in the general corpus can be applied to realms such as politics and sports.

The most frequent war metaphors found in the business English corpus, including *campaign, challenge* and *struggle*, also occur frequently in the general English corpus. The frequency
ratios of these three metaphors are around 1 and the highest one is 1.79 of struggle. The small frequency variances suggest that campaign, challenge and struggle are conventional metaphors often used in both corpora and their uses do not vary much. For instance, the metaphorical use of struggle can be applied in different discourse fields as the following concordance lines from the general English corpus show.

1. Mainstay of familiarity and assistance in a new country. #” Gay Latinos constantly struggle with multiple identities,
2. With China, we must combine partnership, cooperation, deterrence, and power struggle. Achieving that requires cooperation
3. Badly, and I didn’t want to. “ # Most mothers say they struggle to make peace with hurtful comments. # Swanson says

The first concordance line concerns the gay Latinos who are trying hard to manage their identities and be accepted by the society. The second concordance line concerns politics with the mention of partnership, cooperation, deterrence and power struggle in the context. The third concordance line concerns mothers who want to raise their kids according to what the pediatricians suggest, whose method is not understood by other mothers. The third concordance line is thus about daily life struggles. The three concordance lines involve the struggles in different areas of life for different purposes, which suggest that wide and pervasive metaphorical use of struggle.

### 3.3 Statistical comparison between the Chinese business corpus and general corpus

Table 2 compares the frequencies per 1.6 million words of war lexical items used metaphorically in the Chinese general and business corpora, after checking all the concordance lines.

Table 2 War terms in the Chinese business and general corpus with descending ratio

<table>
<thead>
<tr>
<th>Chinese word</th>
<th>Business</th>
<th>General</th>
<th>Ratio of comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>弹药</td>
<td>7</td>
<td>0.03</td>
<td>233.33</td>
</tr>
<tr>
<td>失守</td>
<td>95</td>
<td>0.52</td>
<td>182.69</td>
</tr>
<tr>
<td>撤退</td>
<td>18</td>
<td>0.1</td>
<td>180.00</td>
</tr>
<tr>
<td>收复</td>
<td>78</td>
<td>0.45</td>
<td>173.33</td>
</tr>
<tr>
<td>进驻</td>
<td>50</td>
<td>0.45</td>
<td>111.11</td>
</tr>
<tr>
<td>地雷</td>
<td>3</td>
<td>0.03</td>
<td>100.00</td>
</tr>
<tr>
<td>关口</td>
<td>119</td>
<td>1.31</td>
<td>90.84</td>
</tr>
<tr>
<td>抗战</td>
<td>2</td>
<td>0.03</td>
<td>66.67</td>
</tr>
<tr>
<td>正规军</td>
<td>6</td>
<td>0.1</td>
<td>60.00</td>
</tr>
<tr>
<td>司令</td>
<td>5</td>
<td>0.1</td>
<td>50.00</td>
</tr>
<tr>
<td>子弹</td>
<td>5</td>
<td>0.14</td>
<td>35.71</td>
</tr>
<tr>
<td>火线</td>
<td>7</td>
<td>0.21</td>
<td>33.33</td>
</tr>
<tr>
<td>追兵</td>
<td>1</td>
<td>0.03</td>
<td>33.33</td>
</tr>
<tr>
<td>按兵不动</td>
<td>2</td>
<td>0.07</td>
<td>28.57</td>
</tr>
<tr>
<td>根据地</td>
<td>2</td>
<td>0.1</td>
<td>20.00</td>
</tr>
<tr>
<td>突围</td>
<td>8</td>
<td>0.45</td>
<td>17.78</td>
</tr>
<tr>
<td>集团军</td>
<td>2</td>
<td>0.14</td>
<td>14.29</td>
</tr>
<tr>
<td>战略性</td>
<td>109</td>
<td>7.77</td>
<td>14.03</td>
</tr>
<tr>
<td>冲锋</td>
<td>4</td>
<td>0.31</td>
<td>12.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-----</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>战争</td>
<td>10</td>
<td>0.81</td>
<td>12.35</td>
</tr>
<tr>
<td>转战</td>
<td>14</td>
<td>1.14</td>
<td>12.28</td>
</tr>
<tr>
<td>排头兵</td>
<td>15</td>
<td>1.24</td>
<td>12.10</td>
</tr>
<tr>
<td>先锋</td>
<td>42</td>
<td>3.54</td>
<td>11.86</td>
</tr>
<tr>
<td>重兵</td>
<td>2</td>
<td>0.17</td>
<td>11.76</td>
</tr>
<tr>
<td>反攻</td>
<td>27</td>
<td>2.34</td>
<td>11.54</td>
</tr>
<tr>
<td>前锋</td>
<td>11</td>
<td>1.02</td>
<td>10.78</td>
</tr>
<tr>
<td>突袭</td>
<td>8</td>
<td>0.76</td>
<td>10.53</td>
</tr>
<tr>
<td>防卫</td>
<td>1</td>
<td>0.1</td>
<td>10.00</td>
</tr>
<tr>
<td>刀光剑影</td>
<td>1</td>
<td>0.1</td>
<td>10.00</td>
</tr>
<tr>
<td>兵强马壮</td>
<td>1</td>
<td>0.1</td>
<td>10.00</td>
</tr>
<tr>
<td>炮轰</td>
<td>4</td>
<td>0.41</td>
<td>9.76</td>
</tr>
<tr>
<td>生力军</td>
<td>15</td>
<td>1.65</td>
<td>9.09</td>
</tr>
<tr>
<td>搏杀</td>
<td>10</td>
<td>1.1</td>
<td>9.09</td>
</tr>
<tr>
<td>防御</td>
<td>42</td>
<td>4.75</td>
<td>8.84</td>
</tr>
<tr>
<td>重创</td>
<td>33</td>
<td>3.96</td>
<td>8.33</td>
</tr>
<tr>
<td>短兵相接</td>
<td>2</td>
<td>0.24</td>
<td>8.33</td>
</tr>
<tr>
<td>主战场</td>
<td>6</td>
<td>0.79</td>
<td>7.59</td>
</tr>
<tr>
<td>攻坚战</td>
<td>4</td>
<td>0.55</td>
<td>7.27</td>
</tr>
<tr>
<td>战场</td>
<td>13</td>
<td>1.93</td>
<td>6.74</td>
</tr>
<tr>
<td>退守</td>
<td>2</td>
<td>0.31</td>
<td>6.45</td>
</tr>
<tr>
<td>利剑</td>
<td>3</td>
<td>0.48</td>
<td>6.25</td>
</tr>
<tr>
<td>杀</td>
<td>10</td>
<td>1.65</td>
<td>6.06</td>
</tr>
<tr>
<td>抵抗</td>
<td>15</td>
<td>2.68</td>
<td>5.60</td>
</tr>
<tr>
<td>标兵</td>
<td>4</td>
<td>0.72</td>
<td>5.56</td>
</tr>
<tr>
<td>锋芒</td>
<td>4</td>
<td>0.76</td>
<td>5.26</td>
</tr>
<tr>
<td>一线</td>
<td>31</td>
<td>5.95</td>
<td>5.21</td>
</tr>
<tr>
<td>坚守</td>
<td>17</td>
<td>3.41</td>
<td>4.99</td>
</tr>
<tr>
<td>精兵强将</td>
<td>2</td>
<td>0.41</td>
<td>4.88</td>
</tr>
<tr>
<td>肉搏</td>
<td>1</td>
<td>0.21</td>
<td>4.76</td>
</tr>
<tr>
<td>战火</td>
<td>3</td>
<td>0.65</td>
<td>4.62</td>
</tr>
<tr>
<td>武器</td>
<td>5</td>
<td>1.1</td>
<td>4.55</td>
</tr>
<tr>
<td>拉锯战</td>
<td>4</td>
<td>0.89</td>
<td>4.49</td>
</tr>
<tr>
<td>主攻</td>
<td>11</td>
<td>2.54</td>
<td>4.33</td>
</tr>
<tr>
<td>中坚</td>
<td>10</td>
<td>2.31</td>
<td>4.33</td>
</tr>
<tr>
<td>剑拔弩张</td>
<td>4</td>
<td>0.93</td>
<td>4.30</td>
</tr>
<tr>
<td>主力军</td>
<td>16</td>
<td>3.75</td>
<td>4.27</td>
</tr>
<tr>
<td>大刀阔斧</td>
<td>6</td>
<td>1.41</td>
<td>4.26</td>
</tr>
<tr>
<td>战备</td>
<td>1</td>
<td>0.24</td>
<td>4.17</td>
</tr>
<tr>
<td>滑铁卢</td>
<td>4</td>
<td>1</td>
<td>4.00</td>
</tr>
<tr>
<td>异军突起</td>
<td>10</td>
<td>2.51</td>
<td>3.98</td>
</tr>
<tr>
<td>攻克</td>
<td>10</td>
<td>2.79</td>
<td>3.58</td>
</tr>
<tr>
<td>连珠炮</td>
<td>1</td>
<td>0.28</td>
<td>3.57</td>
</tr>
<tr>
<td>交战</td>
<td>3</td>
<td>0.86</td>
<td>3.49</td>
</tr>
<tr>
<td>军心</td>
<td>2</td>
<td>0.62</td>
<td>3.23</td>
</tr>
<tr>
<td>词语</td>
<td>数量</td>
<td>分数</td>
<td>类型</td>
</tr>
<tr>
<td>-----</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>遭遇战</td>
<td>1</td>
<td>0.31</td>
<td>3.23</td>
</tr>
<tr>
<td>抵御</td>
<td>24</td>
<td>7.5</td>
<td>3.20</td>
</tr>
<tr>
<td>胜出</td>
<td>16</td>
<td>0.31</td>
<td>51.61</td>
</tr>
<tr>
<td>新军</td>
<td>3</td>
<td>1</td>
<td>3.00</td>
</tr>
<tr>
<td>收兵</td>
<td>1</td>
<td>0.34</td>
<td>2.94</td>
</tr>
<tr>
<td>炮弹</td>
<td>2</td>
<td>0.76</td>
<td>2.63</td>
</tr>
<tr>
<td>初战</td>
<td>1</td>
<td>0.38</td>
<td>2.63</td>
</tr>
<tr>
<td>战役</td>
<td>9</td>
<td>3.44</td>
<td>2.62</td>
</tr>
<tr>
<td>守</td>
<td>36</td>
<td>0.87</td>
<td>41.38</td>
</tr>
<tr>
<td>抗衡</td>
<td>9</td>
<td>3.54</td>
<td>2.54</td>
</tr>
<tr>
<td>壁垒</td>
<td>3</td>
<td>1.2</td>
<td>2.50</td>
</tr>
<tr>
<td>攻守</td>
<td>4</td>
<td>1.62</td>
<td>2.47</td>
</tr>
<tr>
<td>大军</td>
<td>7</td>
<td>2.86</td>
<td>2.45</td>
</tr>
<tr>
<td>警戒线</td>
<td>6</td>
<td>2.62</td>
<td>2.29</td>
</tr>
<tr>
<td>进军</td>
<td>48</td>
<td>21.62</td>
<td>2.22</td>
</tr>
<tr>
<td>实战</td>
<td>8</td>
<td>3.72</td>
<td>2.15</td>
</tr>
<tr>
<td>大战</td>
<td>18</td>
<td>8.69</td>
<td>2.07</td>
</tr>
<tr>
<td>追击</td>
<td>2</td>
<td>1.03</td>
<td>1.94</td>
</tr>
<tr>
<td>告捷</td>
<td>3</td>
<td>1.55</td>
<td>1.94</td>
</tr>
<tr>
<td>持久战</td>
<td>1</td>
<td>0.52</td>
<td>1.92</td>
</tr>
<tr>
<td>攻势</td>
<td>7</td>
<td>3.68</td>
<td>1.90</td>
</tr>
<tr>
<td>袭击</td>
<td>8</td>
<td>0.28</td>
<td>28.57</td>
</tr>
<tr>
<td>攻破</td>
<td>8</td>
<td>4.65</td>
<td>1.72</td>
</tr>
<tr>
<td>战斗</td>
<td>12</td>
<td>7.09</td>
<td>1.69</td>
</tr>
<tr>
<td>攻击</td>
<td>31</td>
<td>18.6</td>
<td>1.67</td>
</tr>
<tr>
<td>战略</td>
<td>330</td>
<td>202.52</td>
<td>1.63</td>
</tr>
<tr>
<td>双刃剑</td>
<td>2</td>
<td>1.24</td>
<td>1.61</td>
</tr>
<tr>
<td>战士</td>
<td>1</td>
<td>0.62</td>
<td>1.61</td>
</tr>
<tr>
<td>拼杀</td>
<td>1</td>
<td>0.62</td>
<td>1.61</td>
</tr>
<tr>
<td>制胜</td>
<td>6</td>
<td>3.85</td>
<td>1.56</td>
</tr>
<tr>
<td>激战</td>
<td>3</td>
<td>2</td>
<td>1.50</td>
</tr>
<tr>
<td>争夺战</td>
<td>7</td>
<td>4.68</td>
<td>1.50</td>
</tr>
<tr>
<td>阵营</td>
<td>17</td>
<td>11.46</td>
<td>1.48</td>
</tr>
<tr>
<td>侵占</td>
<td>2</td>
<td>1.38</td>
<td>1.45</td>
</tr>
<tr>
<td>据点</td>
<td>1</td>
<td>0.72</td>
<td>1.39</td>
</tr>
<tr>
<td>参战</td>
<td>1</td>
<td>0.72</td>
<td>1.39</td>
</tr>
<tr>
<td>攻防</td>
<td>5</td>
<td>3.65</td>
<td>1.37</td>
</tr>
<tr>
<td>对抗</td>
<td>12</td>
<td>9.56</td>
<td>1.26</td>
</tr>
<tr>
<td>强攻</td>
<td>4</td>
<td>3.37</td>
<td>1.19</td>
</tr>
<tr>
<td>军团</td>
<td>23</td>
<td>19.61</td>
<td>1.17</td>
</tr>
<tr>
<td>抗击</td>
<td>15</td>
<td>12.94</td>
<td>1.16</td>
</tr>
<tr>
<td>遭战</td>
<td>1</td>
<td>0.96</td>
<td>1.04</td>
</tr>
<tr>
<td>打击</td>
<td>145</td>
<td>154.34</td>
<td>0.94</td>
</tr>
<tr>
<td>瞄准</td>
<td>1</td>
<td>1</td>
<td>0.91</td>
</tr>
<tr>
<td>防线</td>
<td>6</td>
<td>6.92</td>
<td>0.87</td>
</tr>
<tr>
<td>狙击</td>
<td>5</td>
<td>0.38</td>
<td>13.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>攻关</td>
<td>11</td>
<td>13.73</td>
<td>0.80</td>
</tr>
<tr>
<td>混战</td>
<td>2</td>
<td>2.51</td>
<td>0.80</td>
</tr>
<tr>
<td>背水一战</td>
<td>1</td>
<td>1.31</td>
<td>0.76</td>
</tr>
<tr>
<td>挑战</td>
<td>92</td>
<td>125.09</td>
<td>0.74</td>
</tr>
<tr>
<td>打败</td>
<td>2</td>
<td>2.93</td>
<td>0.68</td>
</tr>
<tr>
<td>交锋</td>
<td>7</td>
<td>10.39</td>
<td>0.67</td>
</tr>
<tr>
<td>备战</td>
<td>18</td>
<td>26.77</td>
<td>0.67</td>
</tr>
<tr>
<td>战术</td>
<td>14</td>
<td>21.77</td>
<td>0.64</td>
</tr>
<tr>
<td>战线</td>
<td>3</td>
<td>4.71</td>
<td>0.64</td>
</tr>
<tr>
<td>失败</td>
<td>25</td>
<td>40.75</td>
<td>0.61</td>
</tr>
<tr>
<td>攻坚</td>
<td>3</td>
<td>5.13</td>
<td>0.58</td>
</tr>
<tr>
<td>决战</td>
<td>3</td>
<td>5.2</td>
<td>0.58</td>
</tr>
<tr>
<td>侵袭</td>
<td>3</td>
<td>6.57</td>
<td>0.46</td>
</tr>
<tr>
<td>冲突</td>
<td>6</td>
<td>14.63</td>
<td>0.41</td>
</tr>
<tr>
<td>进攻</td>
<td>16</td>
<td>44.91</td>
<td>0.36</td>
</tr>
<tr>
<td>惨败</td>
<td>2</td>
<td>5.85</td>
<td>0.34</td>
</tr>
<tr>
<td>作战</td>
<td>7</td>
<td>21.42</td>
<td>0.33</td>
</tr>
<tr>
<td>防守</td>
<td>11</td>
<td>40.82</td>
<td>0.27</td>
</tr>
<tr>
<td>出击</td>
<td>18</td>
<td>4.65</td>
<td>3.87</td>
</tr>
<tr>
<td>重击</td>
<td>1</td>
<td>0.31</td>
<td>3.23</td>
</tr>
<tr>
<td>不堪一击</td>
<td>1</td>
<td>0.34</td>
<td>2.94</td>
</tr>
<tr>
<td>袭</td>
<td>45</td>
<td>17.12</td>
<td>2.63</td>
</tr>
<tr>
<td>胜利</td>
<td>12</td>
<td>79.97</td>
<td>0.15</td>
</tr>
<tr>
<td>战胜</td>
<td>14</td>
<td>104.65</td>
<td>0.13</td>
</tr>
<tr>
<td>反击</td>
<td>37</td>
<td>15.37</td>
<td>2.41</td>
</tr>
<tr>
<td>回击</td>
<td>2</td>
<td>0.96</td>
<td>2.08</td>
</tr>
<tr>
<td>失利</td>
<td>3</td>
<td>24.26</td>
<td>0.12</td>
</tr>
<tr>
<td>后卫</td>
<td>4</td>
<td>35.27</td>
<td>0.11</td>
</tr>
<tr>
<td>出战</td>
<td>1</td>
<td>9.08</td>
<td>0.11</td>
</tr>
<tr>
<td>突击</td>
<td>10</td>
<td>8.02</td>
<td>1.25</td>
</tr>
<tr>
<td>助攻</td>
<td>1</td>
<td>14.32</td>
<td>0.07</td>
</tr>
<tr>
<td>攻</td>
<td>84</td>
<td>74.78</td>
<td>1.12</td>
</tr>
<tr>
<td>兵</td>
<td>15</td>
<td>13.93</td>
<td>1.08</td>
</tr>
<tr>
<td>夹击</td>
<td>1</td>
<td>1.15</td>
<td>0.87</td>
</tr>
<tr>
<td>胜负</td>
<td>6</td>
<td>8.78</td>
<td>0.68</td>
</tr>
<tr>
<td>阻击</td>
<td>1</td>
<td>1.92</td>
<td>0.52</td>
</tr>
<tr>
<td>击败</td>
<td>2</td>
<td>75.36</td>
<td>0.03</td>
</tr>
<tr>
<td>领军</td>
<td>27</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
The two sets of war term frequency counts in Table 2 were typed into the SPSS program to see whether they were significantly different. The paired samples t-test performed by SPSS generated a mean difference of 6.18 (SD=25.31), and a significance level of 0.003, which was far below the specified significance level of 0.05 which suggests that there is a statistically significant difference between the two sets of frequency counts (t=3.041, df=154, p=0.003). The result suggests the use of war metaphors in the business corpus is rather different from that in the general corpus in terms of frequency. Furthermore, the ratio between the frequency counts for each word was also calculated and a one-sample t-test was conducted using SPSS. The result show there is a statistically significant difference between the frequency of each word in the Chinese business corpus and in the general Chinese corpus (t=4.440, df=147, p<0.05). The war lexis is used significantly more in the Chinese business corpus than in the general corpus. The test result also indicates that each word differs significantly in terms of its uses in the business context and the general context and confirms the result of the pair-sample t-test.

3.4 Detailed comparison between the Chinese business corpus and general corpus

According to Table 3.2, the most obvious difference between the war metaphors used in the Chinese business corpus and the general corpus is that some expressions related to weapons in warfare are hardly used metaphorically in the general corpus. While danxiao (弹药, ammunition) is used metaphorically seven times in the business corpus referring to money or capital, the compound word is used 0.03 times metaphorically, which suggests it is basically used literally in the general corpus. The same applies to dilei (地雷, landmine) with its 3 occurrences all metaphorical in the business corpus and nearly no metaphorical occurrence in the general corpus and zidan (子弹, bullet) with 5 occurrences in the business corpus and almost no metaphorical occurrence in the general one. Words such as zhanqu (战区, war zone), zhantuan (战团, war clan) and lingjun (领军, lead-army) are not used as metaphors in the general corpus. Expressions referring to troops withdrawing, such as chetui (撤退, withdraw) 18 times all metaphorically, cheli (撤离, withdraw) 11 times all metaphorically all metaphorically in the business corpus, are almost never used metaphorically in the general corpus. Certain collocations can only be found in the business corpus, such as the collocation of the character gong (攻, attack) and prepositions. Formed with the bound morpheme gong (攻, attack) and the preposition as well as the orientation metaphor shang (上, up), the collocation shanggong (上攻, up-attack) can not be found in the general corpus. The collocation is a metaphorical expression referring to the rise in stock market index. Collocations like this include shadie (杀跌, kill-fall).

Some war terms are used less often in the general corpus than in the business corpus, such as kangji (抗击, beat back), juntuan (军团, army group) and qianggong (强攻, storm). The word gongji (攻击, attack) is also less often used in the general corpus than in the business one with a ratio of 0.6 : 1. Further investigation of the word gongji (攻击, attack) shows that it can occur alongside terms which manifest the liquid metaphor, e.g. gongji-lang (攻击浪, attack-wave=a series of rising stocks). The compound gongji-lang (攻击浪, attack-wave) is a mixed metaphor with the war/fighting metaphor combined with the liquid metaphor. In the Chinese stock market, labels are assigned to different shares or stocks according to their propensity to rising or falling prices. If a stock usually has a shorter cycle of its price rising and falling which can be induced in a short period of time, then it is called the “offensive or aggressive” type of stock. On the contrary, the other type of stock is called the “defensive” type of stock. It is also a way to comment on the market and its performance as being “aggressive” or “defensive”. As the market intends to reach or exceed a certain amount of trading volume, constant and steady input of large amounts of investment to influence the market prices via the process of trading in and out using multiple accounts can raise or devalue the stock prices. The Ground of this war metaphor is that the stock prices constantly hover over a specific price
or amount. The specific level or price is seen as a benchmark for the market stocks. If they want to have their prices rise further, they need to reach the benchmark level. Thus the manipulators are making use of their power to influence the prices in order to reach and exceed the level. The Ground for the liquid metaphor is that when efforts were made by manipulators to increase or decrease the trading volumes, a wavy line in resemblance to tidal waves can be identified when drawing connections between the trading volume points. It is because the input of intensive capital in the market left the line chart looking like a wave. The shape of the line probably forms the grounds of the liquid metaphor which is in turn caused by the war metaphor itself. Thus the mixed metaphor is a good illustration of the war metaphor and liquid image metaphor. In terms of the word gongji-lang (攻击浪, attack-wave) which shows interaction between the war metaphor and liquid metaphor, if we are to reactivate the image which can be evoked by the S-terms, we can get the image of humans performing aggressive and attacking action and the image of tidal waves. They may seem incongruous at first sight, but we can still understand them as attempts and actions of real attack taking place continuously. The word gongji-lang (攻击浪, attack-wave) is rarely used in the general corpus. Its only occurrence takes place in the field of sports, referring to successive attempts made by Chinese footballers to score points against the Japanese football team.

4. Conclusion

The review of metaphor studies shows that war metaphor is dominant and deeply entrenched in business discourse. War metaphor is more often exploited to talk about business activity and conceptualize business life as a battlefield full of conflicts and hostility. However, the comparison of frequency counts of war lexis in the English business and general corpus shows no statistically significant difference, suggesting the war metaphors are no more frequent for talking about business activity than they are for talking about activity in general. War metaphor is not dominant for the conceptualization of business activity, as the English data shows. The emphasis on the dominance of war in English business discourse may lead to the misbelief that business activity is more often conceived as competition and war and the overstatement of competition and war in the business field. On the other hand, the war and fighting lexis is used significantly more in the Chinese business corpus than in the general corpus, which confirms the researchers’ argument that business activity is conceived more in terms of competition and war than activity in general and it is dominant and deeply entrenched in Chinese business discourse.

References


A Diachronic Comparison of the Strategies Adopted by The New York Times in Translating Expressions with Chinese Characteristics

Meiyan Liu, Bo Li
Northwestern Polytechnical University, China
myliu@nwpu.edu.cn

Abstract
With the deepening of globalization and growing international influence, news about China appears in the main media like The New York Times more frequently than ever before. Thus the words and expressions with Chinese characteristics catch the eyes of the scholars in recent translation studies. Based on cultural translation theory, the author did a diachronic study about the translation strategies adopted by The New York Times from 2007 and 2013 in translating expressions with Chinese characteristics to find out the changes and development trend. The research results reveal that The New York Times mainly adopts five translation strategies. The diachronic comparative result indicates that the proportion of transliteration and literal translation are gradually rising, while the proportion of transliteration plus explanation and literal translation plus explanation drops slightly from 2007 to 2013. Moreover, there is an obvious decrease of liberal translation. Evidence suggests that The New York Times is showing an increasing tendency to adopt foreignization strategy in translating expressions with Chinese characteristics.

Keywords: translation strategy, expressions with Chinese characteristics, diachronic comparison, foreignization, domestication

1. Introduction
China’s rapid development and increasing international status arouses much attention from news media all over the world. They begin to report China more frequently and widely. For instance, Newsweek published a special report named China’s Century on May 9th, 2009, giving a twenty-one-page-long coverage of China from politics, economics, education, culture, etc. Under such circumstances, an increasing number of scholars shift their attention to the translation of words and expressions with Chinese characteristics.

However, as Hu (2008) holds, the studies are paying too much attention to general description from macroscopic aspects instead of detailed analysis and specific case study. Wang (2009) also argues that the existing studies are mainly based on top-down approach instead of bottom-up. What’s more, little diachronic study has been done systematically with statistics from international media. Therefore, the paper intends to figure out the strategies adopted by The New York Times from 2007 and 2013 in translating expressions with Chinese characteristics and make a diachronic study with abundant first-hand data.

2. Literature review
2.1 Susan Bassnett’s cultural translation theory and two translation strategies suggested
2.1.1 Susan Bassnett’s cultural translation theory
The late 1980s witnessed a growing popularity of translation studies from the perspective of cultural aspect. Susan Bassnett, the most outstanding and influential scholar points out that translation is not a pure language behavior and it is deeply rooted in the culture of the language. Thus translation is a kind of communication between inner culture and inter culture. In her book “Translation Studies”, she highlighted the close relationship between language, culture and translation: “language is the heart of culture, and the surgeon, operating the heart,
cannot neglect the body that surrounds it, so the translator treats the text in isolation from the culture at his peril” (Bassnet, 1988: 14).

In 1990, Susan Bassnett and Andre Lefevere published their co-edited book *Translation, History and Culture*, in which they officially put forward the “cultural turn” in translation studies and explored cultural translation theory. First, translation should take culture as the translation unit, not always linger on the discourse level. Second, translation is not a simple process of decoding and recording, but an activity of communication. Third, translation should not be restricted as the description of source text but viewed as the functional equivalence in the target language culture. Fourth, translation will have different principles and norms in different times in order to meet different needs. Translation, in a word, is to satisfy the needs of different cultures and the needs of different groups in a certain culture (Bassnett, 1990: 4).

From Bassnett’s idea, we can conclude that translation is not a pure linguistic activity. It includes not only the language transformation but also the cultural values transfer.

2.1.2 Foreignization and domestication
As translation studies attach more importance to culture, whether to adopt the strategy of foreignization or domestication has become a hot topic. According to Venuti (1995), foreignization means taking the reader over to the foreign culture and making him or her see the differences; while domestication means bringing the foreign culture close to the reader in the target culture, making the text recognizable and familiar. Which of the two strategies adopted determines the specific translation methods such as literal translation or liberal translation to be selected during the translation activity.

2.2 Present studies on translation strategies of expressions with Chinese characteristics
Since China English was first proposed in the early 1980s, many linguists and translators like Wang (1991), Li (1993), Xie (1995), etc. have all expressed their understanding of China English. Although they define China English in different ways, the basic characteristics of China English they convey are the same. First, China English is an objective existence of linguistic phenomenon in culture exchange between Chinese and English-speaking people. Second, it takes normative English as its core to express things peculiar to Chinese society and culture. Third, China English, as an English variety, can be accepted by native English speakers and it has found its way into English language.

It is widely believed that the features of China English are most obvious and distinctive at the lexical level. According to Kirkpatrick (2007: 147), localized lexis of China English has been created through transliteration, direct translation and semantic shift. Jia (2008) analyses China-specific words in Newsweek and concludes that foreignization approach is subject to ideology to a great extent, so translators are required to combine foreignization and domestication in translation. Fan (2012), after making a survey of the translation strategies adopted by some international media about Chinese culture-loaded words, summarizes that the techniques of foreignization is mainly used in translating expressions with Chinese characteristic.

Generally speaking, the studies based on data from international media are still limited. Furthermore, the studies are mainly based on top-down approach instead of bottom-up. And most of the studies focus on synchronic study rather than diachronic study.

Therefore, in order to compensate for the weakness in the study of China English translation, the author did a systematic diachronic study of the strategies and methods and the change of the strategies by The New York Times from 2007 to 2013 to reveal how native- speakers translate expressions with Chinese characteristics in the hope to shed light to the related domestic study.
3. Research methodology

3.1 Research questions
(1) What strategies are adopted by The New York Times in translating expressions with Chinese Characteristics?
(2) Are there any changes about the translation strategies adopted by The New York Times from 2007 and 2013? If any, what is the development trend?

3.2 Research subject
The New York Times is selected as the subject of the research, because it is regarded as the most authoritative and influential English newspaper not only in its own country, but also in the world and it covers the widest and most comprehensive reports about China among American media (Fan, 2012). Thus, China English used in The New York Times can, to a greater extent, gain their entry to international society. Furthermore, almost all the reports are written by native speakers, which make a great contribution to the accuracy and intelligibility of China English in The New York Times. Thus, the samples collected from The New York Times could make the findings and results of the research more valid, reliable and representative.

3.3 Data collection
The author first downloads all the articles about China from the official website of the New York Times with a period of six months from January to June in the year of 2007 and 2013 respectively. Altogether 760 articles have been collected with about 1.25 million words. All the collected articles are given clear indication of their publication dates and pages in its print version rather than those which are only circulated on the internet.

Then each article is read carefully, and 567 expressions with Chinese characteristics are selected out, together with their publication dates and pages, and the corresponding Chinese meaning is annotated by the author as well. Then the data is carefully classified according to the translation strategies adopted, followed by a diachronic comparison between 2007 and 2013.

3.4 Research findings
With careful data collection and analysis, the author concludes the translation strategies about expressions with Chinese characteristics in 2007 and 2013 as follows.

<table>
<thead>
<tr>
<th>Table 1 Classification of Translation Strategies about Expressions with Chinese Characteristics in 2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Number</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>Percentage</td>
</tr>
</tbody>
</table>

427
Table 2 Classification of Translation Strategies about Expressions with Chinese Characteristics in 2013

<table>
<thead>
<tr>
<th></th>
<th>Transliteration</th>
<th>Transliteration plus explanation</th>
<th>Literal translation</th>
<th>Literal translation plus explanation</th>
<th>Liberal translation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Number</td>
<td>78</td>
<td>42</td>
<td>148</td>
<td>33</td>
<td>68</td>
<td>369</td>
</tr>
<tr>
<td>Percentage</td>
<td>21.1%</td>
<td>11.4%</td>
<td>40.1%</td>
<td>8.9%</td>
<td>18.4%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 3 Comparison of Translation Strategies about Expressions with Chinese Characteristics between 2007 and 2013

<table>
<thead>
<tr>
<th></th>
<th>Transliteration</th>
<th>Transliteration plus explanation</th>
<th>Literal translation</th>
<th>Literal translation plus explanation</th>
<th>Liberal translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>15.2%</td>
<td>12.6%</td>
<td>34.3%</td>
<td>12.6%</td>
<td>25.3%</td>
</tr>
<tr>
<td>2013</td>
<td>21.1%</td>
<td>11.4%</td>
<td>40.1%</td>
<td>8.9%</td>
<td>18.4%</td>
</tr>
<tr>
<td>Trends</td>
<td>5.9% ↑</td>
<td>1.2% ↓</td>
<td>5.8% ↑</td>
<td>3.7% ↓</td>
<td>6.9% ↓</td>
</tr>
</tbody>
</table>


According to the data shown in Table 1 and Table 2 above, in translating expressions with Chinese characteristics, *The New York Times* mainly adopts five translation strategies: transliteration, transliteration plus explanation, literal translation, literal translation plus explanation and liberal translation. The detailed classification of the collected 567 expressions with Chinese characteristics are as follows: 108 samples are transliterated; 67 ones are transliteration plus explanation; 216 ones belong to the strategy of literal translation; 58 ones are literal translation plus explanation; and 118 go to the category of liberal translation. Obviously, the techniques of literal translation is the most commonly used one, with 34.3% in 2007 and 40.1% in 2013, followed by Transliteration, with 15.2% in 2007 and 21.1% in 2013 and liberal translation, with 25.3% in 2007 and 18.4% in 2013. The following is a detailed analysis of the strategies used by *The New York Times*.

4.1 Transliteration

Transliteration is one of the strategies adopted by *The New York Times* in dealing with some proper names or objects, things and phenomenon which are peculiar to Chinese culture with no correspondences in English. Altogether 108 samples (30 in 2007 and 78 in 2013) are found, which account for 15.2% in 2007 and 21.1% in 2013. For example,

(1) But since David wasn’t around when I checked in, I put my bags away, walked out through the lane — where grandmothers played mah-jongg outside pink stucco homes and cicadas chirred in the trees — and grabbed a quick snack of jian bing, a crepe stuffed with egg, chili sauce and a piece of fried dough, from a street vendor (1 yuan). (January 21, 2007, page TR12)

(2) At the well-tended Tianshouyuan Cemetery in Beijing, an idyllic spot on the outskirts of the city that is famous for its feng shui, grave plots for ashes sell for as much as $46,000 per square meter. (April 5, 2013, page A5)

The above “Mahjong” and “feng shui” are transliterated from Chinese “风水”, and “麻将” respectively. Blessed with strong Chinese characteristics, they suffer semantic vacancy in
English. Thus, the techniques of transliteration is adopted by The New York Time to maintain the “Chinese flavor”.

The following are more examples:

<table>
<thead>
<tr>
<th>Chinese Items</th>
<th>English Translation</th>
<th>Published Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>太极</td>
<td>tai chi</td>
<td>April 5, 2007, page G1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>March 17, 2013, page TR6</td>
</tr>
<tr>
<td>捞面</td>
<td>lo mein</td>
<td>March 8, 2007, page A1</td>
</tr>
<tr>
<td>阴阳</td>
<td>yin and yang</td>
<td>April 13, 2007, page E29</td>
</tr>
<tr>
<td>户口</td>
<td>Hukou</td>
<td>February 6, 2013, page A13</td>
</tr>
<tr>
<td>小笼包</td>
<td>xiao long bao</td>
<td>March 17, 2013, page TR7</td>
</tr>
<tr>
<td>人民币</td>
<td>renminbi</td>
<td>March 17, 2013, page TR6</td>
</tr>
<tr>
<td>磕头</td>
<td>Kowtow</td>
<td>May 11, 2013, page A1</td>
</tr>
</tbody>
</table>

4.2 Transliteration plus Explanation
Building strong exotic appeal and rightly filling up the cultural gap with additional explanations, the technique of transliteration plus explanation wins a wide support from translators both home and abroad, and has become a very popular and effective skill in translating words with Chinese characteristics adopted by The New York Times. A total of 67 examples are collected, of which 25 ones (12.6%) are from 2007 and 42 (11.4%) are from 2013. Evidence shows that many concepts about Chinese culture such as music, religion, food, architecture, new concepts, hot issues and specific phenomenon in current Chinese society etc. are dealt with “transliteration plus explanation” approach. For example,

(1) To ensure her good character and fortune, he insisted that her wuguan — *a feng shui-like reading of the sense organs on the face* — show perfect harmony. (March 10, 2013, page BU1)

(2) Seated cross-legged in her farmhouse on the *kang, a brick sleeping platform warmed by a fire below*, Meng Shujing lifted her chin and sang a lullaby in Manchu, softly but clearly. (March 18, 2007, page A8)

Here Wuguan(五官) and (Kang)炕, with local culture connotation, will definitely be less intelligible to most native English readers, so annotation (wuguan) —*a feng shui-like reading of the sense organs on the face* and (kang), “a brick sleeping platform warmed by a fire below” are given to eliminate communication barriers. Here are some more examples:

<table>
<thead>
<tr>
<th>Chinese Items</th>
<th>English Translation</th>
<th>Published Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>新华社</td>
<td>Xinhua, the official state news agency</td>
<td>January 12, 2013, A7</td>
</tr>
<tr>
<td>石库门</td>
<td>shikumen, or stone gatehouses that are built around internal lanes</td>
<td>April 7, 2007, page A4</td>
</tr>
<tr>
<td>富二代</td>
<td>fuerdai, or “second-generation-rich,”</td>
<td>March 10, 2013, page BU1</td>
</tr>
<tr>
<td>胡同</td>
<td>hutongs — alleys formed by the outer wall of courtyard homes</td>
<td>March 3, 2013, page</td>
</tr>
<tr>
<td>萧</td>
<td>xiao, an ancient, flutelike instrument</td>
<td>April 10, 2013, page C1</td>
</tr>
<tr>
<td>气功</td>
<td>Qigong, which entails coordinating slow movements with breathing to cultivate the flow of energy, or qi, in a sort of graceful, fluid dance</td>
<td>April 5, 2007, page G1</td>
</tr>
</tbody>
</table>
4.3 Literal translation
The data from *The New York Times* reveals that the strategy of literal translation is used most frequently with a total of 216 examples collected, which make up 34.3% in 2007 and 40.1% in 2013 in the total samples. And multiple of items concerning economy, politics, society, culture, etc fall into this category. For example,

(1) Since his appointment as party chief, Mr. Xi has championed the slogans of "the Chinese dream" and a "great revival of the Chinese nation" to appeal to ordinary citizens who are often angry over official corruption and wealth disparities. (中国梦; 中华民族的伟大复兴) (March 24, 2013, page A8)

(2) The Chinese police said on Tuesday that they had detained Gong Ai’ai, popularly known as *House Sister*, in a corruption case that has gripped China for weeks after it was revealed that Ms. Gong and her family had accumulated dozens of houses and other properties. (房姐) (February 6, 2013, page A13)

As is shown in the above examples, these literally translated items, vividly expressed, can be easily understood with contextual information. Furthermore, literal translation, by retaining the foreignness of the source text as much as possible, refreshes target readers with original Chinese culture.

Here are more examples from the collected data:

<table>
<thead>
<tr>
<th>Chinese Items</th>
<th>English Translation</th>
<th>Published Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>和平崛起</td>
<td>peaceful rise</td>
<td>January 20, 2007, A7</td>
</tr>
<tr>
<td>第一桶金</td>
<td>first bucket of gold</td>
<td>February 5, 2007, page A2</td>
</tr>
<tr>
<td>人民当家做主</td>
<td>let the people be the masters of their own home</td>
<td>March 17, 2007, A3</td>
</tr>
<tr>
<td>经济特区</td>
<td>special economic zone</td>
<td>January 27, 2013, page SR7</td>
</tr>
<tr>
<td>摸着石头过河</td>
<td>crossing the river by groping stones</td>
<td>February 15, 2013, page A1</td>
</tr>
<tr>
<td>农村医疗保险</td>
<td>rural health insurance</td>
<td>March 6, 2013, page A12</td>
</tr>
<tr>
<td>学雷锋纪念日</td>
<td>Learn From Lei Feng Day</td>
<td>March 12, 2013, page A5</td>
</tr>
<tr>
<td>一个中国政策</td>
<td>“one-China policy”</td>
<td>April 14, 2013, A6</td>
</tr>
</tbody>
</table>

4.4 Literal translation plus explanation
Based on data from *The New York Times*, a total of 58 samples (12.6% in 2007 and 8.9% in 2013) relating to politics, social phenomenon, history, etc. are literally translated with explanation. For example,

(1) I have also shot the so-called "ant tribe," which comprises graduates from provincial and unprestigious universities who flock to the capital in search of their Beijing Dream, only to find themselves working low-paying jobs and living in dorm-style hostels on the outskirts of the city. (February 10, 2013, page A1)

*Ant tribe* (蚁族) is a new term popular in current Chinese society, which is used to describe the poor living conditions of the young graduate in big cities. Here, literal translation conveys the vivid rhetorical effect of Chinese language, and explanation details the specific meaning and background information.

(2) The show featured talent from China, Tan Weiwei and Li Yuchun, two winners of China’s *Super Girl contest* — a competition that attracts a far larger audience than “American Idol.” (April 19, 2007, C9)
As a Chinese singing contest organized by Hunan TV station, Super Girl (超级女声) is one of the most popular entertainment shows. Despite of its big hit in China, it appears to be unfamiliar to American people. The reporter here compares it to “American Idol”, which is similar to Super Girl in many aspects and enjoys unparalleled success in American broadcasting history. With this annotation, the target readers will have a better understanding of China’s Super Girl contest.

The following are more examples:

<table>
<thead>
<tr>
<th>Chinese Items</th>
<th>English Translation</th>
<th>Published Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>钉子户</td>
<td>“nail house,” as many here have called it because of the homeowner’s tenacity, like a nail that cannot be pulled out</td>
<td>March 27, 2007, A4</td>
</tr>
<tr>
<td>小皇帝</td>
<td>“little emperors,” the singletons mandated by the one-child policy of the past quarter of a century</td>
<td>April 1, 2007, MAGZINE</td>
</tr>
<tr>
<td>老虎苍蝇一起打</td>
<td>take on both “flies” and “tigers” — wayward local officials and corrupt senior leaders</td>
<td>January 24, 2013, A13</td>
</tr>
<tr>
<td>微博</td>
<td>Sina Weibo, the Chinese equivalent of Twitter</td>
<td>March 2, 2013, A4</td>
</tr>
<tr>
<td>地沟油</td>
<td>“ditch” cooking oil that has been used and dumped in drains and processing it to be resold for cooking.</td>
<td>May 4, 2013, A4</td>
</tr>
</tbody>
</table>

4.5 Liberal translation
When literal translation fails to convey the cultural connotation accurately, liberal translation is often adopted. The data from The New York Times reveals that a total of 118 samples (50/25.3% in 2007 and 68/18.4% in 2013) adopt the strategy of literal translation. And many items concerning politics, economy, culture, language, geography, etc. fall into this category. For example,

(1) He praised the real estate industry as an essential part of the national economy, but he called on developers, who have built expensive projects all over China, to also focus on building affordable housing and not to threaten “primary farmland.” (March 5, 2007, A8)

In this example, “经济适用房” is difficult to translate word by word. Based on their meaning, the writer liberally translates it into “affordable housing”

<table>
<thead>
<tr>
<th>Chinese Items</th>
<th>English Translation</th>
<th>Published Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>小康社会</td>
<td>moderately prosperous society</td>
<td>January 27, 2013, A5</td>
</tr>
<tr>
<td>春运</td>
<td>traditional New Year trips</td>
<td>February 17, 2013, page A1</td>
</tr>
<tr>
<td>彩礼</td>
<td>bride price — a sort of dowry used in rural China</td>
<td>March 10, 2013, page BU1</td>
</tr>
<tr>
<td>银耳</td>
<td>snow fungus</td>
<td>May 13, 2013, page A1</td>
</tr>
<tr>
<td>公款吃喝</td>
<td>taxpayer-financed banquets</td>
<td>April 5, 2013, page A7</td>
</tr>
</tbody>
</table>
Comparing the figures in Table 3, differences can be found in translation strategies adopted by *The New York Times* in 2007 and 2013. First, the proportion of transliteration and literal translation increased by 5.9% and 5.8% respectively in 2013 compared with that of 2007, which indicate that these two methods enjoy an increasing popularity in international media in dealing with expressions with Chinese characteristics.

Second, the proportion of transliteration plus explanation and literal translation plus explanation decreased to some extent. In the process of data analysis, the author notices that many expressions, which were transliterated plus explanation or literally translated plus explanation in 2007, begins to be transliterated or literally translated without annotations in 2013. This phenomenon can be best illustrated by the examples in the following table.

<table>
<thead>
<tr>
<th>Transliterated plus explanation / literally translated plus explanation in 2007</th>
<th>Transliterated or literally translated in 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>the Chinese currency, the yuan or renminbi (人民币)January 26, 2007, C3</td>
<td>renminbi (January 9, 2013, page A10) Yuan (May 1, 2013, page A1)</td>
</tr>
<tr>
<td>Qigong(气功), which entails coordinating slow movements with breathing to cultivate the flow of energy, or qi, in a sort of graceful, fluid dance ( April 5, 2007, G1)</td>
<td>Qigong (June 7, 2013, page C8)</td>
</tr>
<tr>
<td>Confucianism(儒家思想), the political and social philosophy of ancient China, which treats harmony as a core virtue (April 20, 2007, A1)</td>
<td>Confucianism (March 1, 2013, page A23)</td>
</tr>
<tr>
<td>feng shui(风水), the ancient Chinese practice of positioning objects to create a harmonious environment ( May 2, 2007, A4)</td>
<td>feng shui (April 5, 2013, page A5)</td>
</tr>
<tr>
<td>Great Hall of the People in Beijing, the country’s most important political meeting place(人民大会堂) (March 27, 2007, C3)</td>
<td>Great Hall of the People (March 5, 2013, page A4)</td>
</tr>
<tr>
<td>The Yangtze(长江), the third longest river in the world after the Nile and the Amazon (April 17, 2007, A10)</td>
<td>Yangtze River (March 17, 2013, page TR6)</td>
</tr>
<tr>
<td>Three Gorges Dam( 三峡大坝 ), the world’s biggest hydropower project (April 17, 2007, A10)</td>
<td>Three Gorges Dam (March 17, 2013, page TR6)</td>
</tr>
</tbody>
</table>

Third, as is shown by the figure, there is an obvious reduce of the proportion of liberal translation by 6.9 %, indicating that the approach of liberal translation is less adopted in *The New York Times*.

To sum up, of the five translation strategies adopted by *The New York Times*, transliteration, literal translation, transliteration plus explanation and literal translation plus explanation keep the original Chinese language features and cultural uniqueness to a great extent and thus can be classified into foreignization approach. Liberal translation, adopting the expressions of the target culture, is a kind of domestication. The data implies that *The New York Times* tends to adopt the strategy of foreignization more in 2013 compared with that of 2007 in dealing with
expressions with Chinese characteristics. Besides, the drop of transliteration plus explanation and literal translation plus explanation also add weight to the conclusion.

6. Conclusion and implications

Based on the studies discussed above, the major findings can be drawn as follows: First, Five translation methods are mainly adopted by The New York Times in translating expressions with Chinese characteristics. They are the approach of transliteration, transliteration plus explanation, literal translation, literal translation plus explanation and liberal translation, of which the most frequently used are literal translation, liberal translation and Transliteration. Second, as indicated by the diachronic comparison, the proportion of transliteration and literal translation rise, while liberal translation, the typical domestication approach is falling with time pass-by, showing an increasing tendency of foreignization approach in the translation of expressions with Chinese characteristics by The New York Times.

Many reasons can account for the change. First, due to frequent cross-cultural communication among nations, no culture can refuse to accept others’ culture, thus different cultures may infiltrate into each other. Second, people all over the world are realizing the necessity and importance of showing respect to cultural diversity and the widely adopt of foreignization approach in The New York Times reveals their respect for foreign cultures. Third, to grab the readers’ attention is the highest priority in news report writing, so an effective way is to adopt foreignization translation method to attract readers’ interest and retain the original cultural identity with exotic flavor. Fourth, the rhetorical style of news report in American media also accounts for their preference to foreignization since foreignization contributes to the objectiveness and authenticity of news. By adopting transliteration and literal translation, the reporter also reveals his professionalism and proficiency in reporting China. Furthermore, loan words are said to be preferred in news report in America. Besides, the advocating of novelty and ingenuity in news report in western society also add weight to the adoption of foreignization. Above all, due to China’s booming economy and increasing international status, Chinese language and culture has been sweeping the world. In the context of a global “Chinese fever”, many countries have seen a growing number of Confucius Institute. Therefore, more and more Chinese loanwords are gaining their entry to English vocabulary. As is said by Paul Payack, the chairman of Global Language Monitor, “with tremendous increase of China’s economy, China English exerts far more effect on international English than English-speaking countries” (Li, 2009: 66). However, liberal translation is still an important method adopted in translating Chinese culture-loaded words where gaps occur between the two cultures.

The study could shed light on translation in the following aspects. First, expressions with Chinese characteristics should be dealt flexibly in translation. As the westerners, eager to know what are happening in the outside world, have been fully prepared to accept new knowledge and new cultures to broaden their horizon, transliteration and literal translation would be the best choice. So Chinese translators can confidently adopt foreignization in translation to convey unique Chinese language and culture and attract foreign readers with exotic flavor. Second, despite the merits of foreignization, translators should be cautious of over-foreignization. Be it foreignization or domestication, intelligibility and readability of the target text are the key points to keep in mind. Furthermore, liberal translation should be adopted when cultural gap is difficult to fill so as to ensure intelligibility and readability of the translated text.

References


The Effects of Task Types on Incidental Vocabulary Acquisition through L2 Listening

Ximing YUAN
Northwestern Polytechnical University, China
yuanxm@nwpu.edu.cn

Abstract
As Nagy, Herman and Anderson proposed, incidental learning occurs when learners acquire new aspects of their L2 without being focused on doing so. Researchers have since been long interested in enlarging new vocabulary through incidental learning. Various empirical studies have been carried out but mostly in the area of reading. This study intends to investigate: 1) whether learners can acquire vocabulary through listening; 2) the effects of different task types on incidental vocabulary acquisition based on the Involvement Load Hypothesis, especially the effects on instant and delayed vocabulary recall and retention. 74 non-English major freshmen in Northwestern Polytechnical University have been involved in the study. The study reveals very meaningful findings. In the end, the study discusses its implications for vocabulary teaching in college English teaching.

Keywords: incidental vocabulary acquisition, Involvement Load Hypothesis, task type

1. Introduction

Incidental learning is the process of learning something without the intention of doing so. It is also learning one thing while intending to learn another (Richards and Schmidt 2002). Language educators and researchers have long focused on the acquisition of new vocabulary knowledge through incidental learning (Day, Omura, and Hiramatsu 1991; Jenkins, Stein, and Wysocki 1984; Nagy, Herman, and Anderson 1985; Saragi, Nation, and Meister 1978). From the vast number of study, incidental vocabulary acquisition has been proved effective in the uptake of new vocabulary, especially from the context of reading.

Craik and Lockhart (1972) presented the levels of processing framework as an alternative to the theories of memory that postulated separate stages for sensory, working and long-term memory. According to the framework, stimulus information is processed at different levels simultaneously depending on its characteristics. While in incidental vocabulary acquisition, they suggested that the “deeper” the processing, the more would be remembered. In light of this theory, Laufer (2005) proposes that form focused instruction is indispensable for L2 lexical acquisition. Form-focused vocabulary tasks prove superior to reading alone because they push learners to engage in more elaborate processing of previously unknown words. Laufer and Hulstijn (2001) proposed the Involvement Load Hypothesis, which leaves much more room for future researchers in incidental vocabulary learning.

2. Involvement Load Hypothesis

Involvement load hypothesis claim that word learning and retention depend on the amount of effort and attention allocated to processing these words. Task-induced involvement is considered as a motivational-cognitive construct entailing three factors: need, search, and evaluation. Need is motivational, non-cognitive element of involvement when learners are required to use new words to complete a given task. Need is moderate when it is imposed by the task, such as answering reading comprehension questions; while strong when learners are intrinsically motivated. Search and evaluation are the cognitive elements of involvement because they entail information processing and require focusing on word forms and meaning. Search refers to the attempt learners make out the meaning of unknown words in a task.
Search is present when learners try hard to complete the task through different ways like looking up in the dictionary and consulting teachers; while it is present when no such effort is required, e.g. reading comprehension tasks accompanied by marginal glosses. Evaluation is a decision making process, which entails “a comparison of a given word with other words, a specific meaning of a word with its other meaning, or comparing the word with other words in order to assess whether a word (i.e., a form-meaning pair) does or does not fit its context” (Laufer and Hulstijn 2001: 14). Evaluation is moderate when learners are required to identify the differences between words provided in a given context (e.g. deciding which meaning of a target word best fits the context in which it is encountered) and strong when the task requires making decisions about new words and combining them with known words in original contexts, such as sentence making and writing.

3. The study

A large number of previous research studies have been conducted to prove the gains of incidental vocabulary learning. Eelly (1989: 174-487) found that reading stories aloud to children could promote word learning; Vidal compared gains from listening and reading (2011), and found out that students acquired more from reading than from listening. Vidal also explored incidental vocabulary acquisition gains from L2 listening (2003) by asking students to complete fill-in-blank passages and answer true & false questions after watching 3 academic video lectures. As to the effectiveness of task-induced involvement load on incidental vocabulary learning, Beal (2007) found that participants in the sentence production group performed better than the multiple-choice glosses and textual glosses. Keating (2008) assigned three different tasks in his vocabulary learning study, reading only, fill-in-blank task, and sentence construction. With the comparison, he found subjects in the sentence writing task group surpassed the other two. However most studies are carried out in reading and writing instead of L2 listening. Findings are limited as well.

This study aims to provide a more complete picture of incidental vocabulary knowledge learning from listening. With the basis of Involvement Load Hypothesis, the researcher tries to test and compare the effect of word acquisition and retention under different task type with different involvement load. The analysis was designed to answer the following research questions:

1. Whether language learners can acquire productive word knowledge in L2 listening?
2. Whether language learners can acquire receptive word knowledge in L2 listening?
3. Does task type strongly affect foreign language incidental vocabulary acquisition and retention in L2 listening?
4. What type of task will evolve better effect in word acquisition in listening?

4. Research methodology

4.1 Participants
The participants were 74 undergraduates at Northwestern Polytechnical University in China. They were non-English major freshmen from two classes, belonging to College English Band I level. To ensure they have the similar English listening and vocabulary level, a couple of weeks prior to the treatment, 4 classes were randomly chosen to complete a listening test. Two of them showing no significant differences in the test were eventually chosen to attend the study.

4.2 Research materials and target words
Material selection will decisively affect the test result. Based on the findings that 95% lexical coverage is sufficient for adequate listening comprehension of narrative text (van Zeeland and Schmitt 2012), it was ensured that this coverage level was reached by all participants. The listening passage used was about the biggest snack maker in America and its chips. The whole
passage contains 338 words with a familiar topic relating to daily life. 10 students who were not included in the study were invited to have a pretest and circled out 12 new words. With further confirmation, the researcher finally chose 10 target words.

### 4.3 Research procedure

The study intended to investigate three tasks with different involvement loads. The involvement factors need and search were controlled constant, while the three tasks only differed in the evaluation element they induced. After a careful study of Hulstijn and Laufer’s task design, more details are as below:

**Task A: Listening + Reading comprehension + Word list:**
Participants assigned to Task A were offered with seven listening comprehension questions after listening to the passage. The meanings of the ten target words were explained both in L1 (Chinese) and L2 (English) on the word list given to the participants. The directions of the listening comprehension encouraged students to listen for meaning and to make use of the word list for comprehension. In task A, need was moderate. But the available access to the meanings of the target words made search and evaluation absent. Therefore, the involvement load index for this task was 1 (1+0+0).

**Task B: Listening + Fill-in blanks + Word list:**
In task B, a summary of the listening passage was offered with 10 words deleted. Participants were instructed to complete the blanks with 10 target words plus another additional 4 newly given words related to the topic. Still, they could use word list for help. Task B induced moderate need because participants needed to know the meaning of the target words to fill in the blanks; the task induced moderate evaluation when the participants carefully study and distinguish the target words; the task induced no need with the word list provided. Then the involvement load index was 2 (1+0+1).

**Task C: Listening + Composition writing + Word list:**
After listening, participants assigned to task C were required to write a composition relating to the topic of the passage with the ten target words. Need in task C was moderate because it was imposed by the task and search was still absent. Evaluation was strong because participants had to decide how to use these words together with his previous knowledge to write appropriate sentences. The involvement load index for this task was 3 (1+0+2).

The testing procedure was simple. Immediately after listening to the material, participants were randomly appointed to three groups to fulfill different tasks. And they would then instantly take the posttest. To see the effect of the retention, one week later, all participants would take the delayed posttest without knowing it in advance. The test type and items were the same as the instant posttest only with reversed order.

<table>
<thead>
<tr>
<th>Knowledge measured</th>
<th>Test type</th>
<th>Scoring standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Receptive knowledge (the Chinese equivalent)</td>
<td>Meaning translation</td>
<td>a. Not hearing it 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b. Heard it but don’t know 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>c. The meaning is… 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total: 20 points</td>
</tr>
<tr>
<td>2. Productive knowledge</td>
<td>Sentence completion</td>
<td>a. Using the wrong word or with mistakes above 2 letters 0</td>
</tr>
<tr>
<td></td>
<td>with the correct form of the given word</td>
<td>b. Using the wrong form of the word or with 1 letter spelling mistake 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>c. Using correct form and spelling 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total: 20 points</td>
</tr>
</tbody>
</table>
5. Data analysis

Having collected all data from the instant and delayed post-tests, the researcher employed SPSS (21.0) for Windows Product for data analysis.

Table 2 T test analysis comparing receptive knowledge acquisition in both instant and delayed posttest $p<.05$

<table>
<thead>
<tr>
<th>Task</th>
<th>Instant posttest</th>
<th>Delayed posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>A</td>
<td>9.40</td>
<td>6.12</td>
</tr>
<tr>
<td>B</td>
<td>12.3</td>
<td>10.9</td>
</tr>
<tr>
<td>C</td>
<td>12.9</td>
<td>11.13</td>
</tr>
</tbody>
</table>

Table 3 T test analysis comparing receptive and productive knowledge acquisition in both instant and delayed posttest $p<.05$

<table>
<thead>
<tr>
<th>Task</th>
<th>Instant posttest</th>
<th>Instant posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>A</td>
<td>9.40</td>
<td>7.88</td>
</tr>
<tr>
<td>B</td>
<td>9.83</td>
<td>7.70</td>
</tr>
<tr>
<td>C</td>
<td>10.5</td>
<td>7.69</td>
</tr>
</tbody>
</table>

Table 1 shows the test type and scoring standard of the posttest, which intends to see the participants’ receptive and productive vocabulary knowledge acquisition. The total score for the 4 posttests is 20 points each.

Receptive knowledge and productive knowledge of a word have commonly been used to describe the degree of learners’ word knowledge (Nation 2001; Read 2000). Receptive knowledge in vocabulary involves recognizing the form of a word while reading or listening and retrieving its meaning from mental lexicon; productive vocabulary knowledge involves expressing a message through speaking and writing and producing the appropriate word form appropriately in context. The figures in Table 2 clearly answer the first two research questions. From the mean scores of each group, it is obvious that learners have incidentally acquired vocabulary knowledge in L2 listening, with an improved learning both in receptive
and productive knowledge. Furthermore, they demonstrate better knowledge of the target words in instant posttest than in delayed posttest, which obeys the common rules of word acquisition and retention.

In Table 3, for both instant and delayed posttests, the mean scores in all groups are much higher in the gaining of receptive knowledge than in productive knowledge, and the *t* values also show significant differences. The result indicates that productive knowledge seems harder to acquire. Compared with Task A and B, Task C group has done better in all the tests.

The results in Table 3 already fully support the Involvement Load Hypothesis in that Task B and Task C were more effective than Task A both in receptive and productive instant posttests. One week after treatment, however, Task B was superior to Task A, but Task C was not more effective than Task B. It seems students in Task B group have done better in the delayed posttest. That means with the time difference, different task type may involve different learning effects.

The figures clearly point out that students fulfilling the task of fill-in-blank and the task of writing all did well in learning words, and they are much better than students in Task A group, which refer to the fact that tasks with more involvement load with lead to more effective vocabulary acquisition in L2 listening. Epically students in Task C group, they all felt they could learn new words much easier and better than before.

In delayed posttest, thought students didn’t do better than in the instant posttest, but just as Hulstijn (2001) notes, one expects a decline in knowledge over time in the absence of rehearsal or additional exposure to the target words. The participants in this study were not exposed to the target words between testing intervals. Thus, it is not surprising that there would be a decline in word acquisition.

6. Conclusion

From the analysis above, we can conclude that learners can incidentally acquire vocabulary in L2 listening in both receptive knowledge and productive knowledge learning, but can do much better in receptive knowledge gaining. In the meantime, Students with the task of fill-in-blank and students with task of writing showed obvious gaining of the target words and they did much better than the students with listening comprehension task. Therefore, the study confirms that the type of task with greater involvement load will induce better learning effect. The results of the delayed posttest indicate that students in all the three groups have gained vocabulary retention to some degree, especially the group with writing the task. Based on these results it can be suggested that teachers should design tasks with a higher involvement load to enhance students’ vocabulary retention. Future research on incidental vocabulary learning could include designing more effective task type for students to perform more output and get better word retention.

Acknowledgments

The deepest gratitude goes to Northwestern Polytechnical University. This research study is specially sponsored by Humanities, Social Sciences and Management Research Fund, Western Polytechnical University. The project number is [3102014RW0021].

References


Improving an Electronic Dictionary for Morphological Analysis of Japanese: Use of historical period information

Tomoaki Kouno, Toshinobu Ogiso
National Institute for Japanese Language and Linguistics, Japan
tkouno@ninjal.ac.jp, togiso@ninjal.ac.jp

Abstract
In this paper, we present our improvements to the electronic dictionary UniDic made during the construction of an annotated diachronic corpora of Japanese. In our work, we use UniDic to attach morphological annotations. Each entry of UniDic has a hierarchical structure and consists of three levels: the lemma, form, and orthographic levels. UniDic contains over 240,000 entries and covers not only contemporary but also archaic words; however, using all of these entries is inefficient. In fact, UniDic with all entries caused an increase in machine learning time. In order to cope with this difficulty, we controlled UniDic entries by using historical period information: information on the time span a word was used over. We added historical period information to each entry level while considering historical changes. As a result, UniDic discriminates among the vocabularies used in each period. Applying this improved UniDic. We built two sub-dictionaries specialized for specific periods: one dictionary for Early Middle Japanese and one for Kyōgen (Middle and Early Modern Japanese). Compared to UniDic which uses all entries, UniDic for Early Middle Japanese is smaller (by half), and the time needed for machine learning is remarkably diminished (by 75%). Moreover, machine analysis maintains a high level of accuracy (97%).

Keywords: electronic dictionary, historical period information, diachronic corpus of Japanese, morphological analysis

1 Background

The National Institute for Japanese Language and Linguistics (NINJAL) has been developing annotated corpora of the Japanese language. First, we constructed and released the Balanced Corpus of Contemporary Written Japanese (BCCWJ) containing 100 million words (Maekawa et al. 2014). We then developed the following diachronic corpora: the Corpus of the Heian Period (Early Middle Japanese; 738,000 words), Kokumin no Tomo Corpus (Modern Japanese; 1 million words), and Kyōgen Corpus (Middle and Early Modern Japanese; 235,000 words). All of these corpora are richly annotated, and the morphological information (e.g. word segmentation, parts of speech, reading, and conjugation patterns) is especially useful for linguistic research.

We have used an electronic dictionary called UniDic to attach morphological annotations to our corpora. UniDic contains over 240,000 entries and covers not only contemporary but also archaic words. The text of the corpus is automatically tagged with morphological information by the state-of-the-art morphological analyzer MeCab trained with Unidic (Kudo et al. 2004). MeCab analyzes Contemporary Japanese with great accuracy (98%).

In the process of constructing diachronic corpora, we have added many words appearing in historical texts to the UniDic database (Ogiso et al. 2012). While some words are obsolete now and inessential to the analysis of contemporary Japanese texts, others are newly-coined and unnecessary for old texts; therefore, using all entries for all corpora is inefficient. In fact, UniDic with all entries imposed heavy costs on machine learning and caused an increase in learning time. Moreover, it can negatively affect the accuracy of machine analysis.
This paper aims to improve UniDic by using historical period information (HPI) and handle the problem. In Section 2 we give an outline of our research method and discuss the periodization of Japanese. In Sections 3 and 4, we explain and give examples of HPI. In Section 5, we show the effects of HPI with numerical values, and in Section 6, we present our conclusions.

2 Method

2.1 Outline

In order to cope with the difficulty, we employed HPI: information on the time span a word was used over. To start, we examined the historical period when a word was used, referring to general dictionaries and our developed corpora. This information was attached to each word entry in UniDic as HPI. With HPI, the use of each word entry was limited to the proper period. We went on to build sub-dictionaries specialized for specific periods (e.g. UniDic for Early Middle Japanese, UniDic for Kyōgen) based on the HPI. As a result, the morphological analyzer uses dedicated entries for texts of particular time periods (for example, it uses the Early Middle Japanese entries for Early Middle Japanese texts) and can analyze historical texts more efficiently.

2.2 Periodization of the Japanese language

Recorded Japanese documents date back to the eighth century. From the eighth century to the twenty-first century, Japanese has undergone both lexical and grammatical changes at some stages. To categorize the historical periods of Japanese, we adopt a sixfold classification: Old, Early Middle (E-Mid.), Middle, Early Modern (E-Mod.), Modern, and Present (Pres.) periods. This linguistic periodization roughly corresponds to the political one (see Figure 1).

![Figure 1 Periodization of the Japanese Language](image)

3 Historical period information

3.1 The hierarchy of UniDic entries

Each word entry of UniDic has a hierarchical structure consisting three levels: the lemma, form and orthographic levels. The lemma is like the headword of a general dictionary and is the highest level of the hierarchy. The form level distinguishes different forms and conjugation types while the orthographic (orth) level distinguishes variant spellings. This hierarchical representation is useful for linguistic researchers who study variations in orthography and word form.
Figure 2. Hierarchical Structure of UniDic

Figure 3. Structure of the Entry *izuko* 何処

Figure 3 shows the internal structure of the entry *izuko* 何処 (‘where’). The word *izuko* has three form variations: *izuko*, *izuku*, and *idoko*. The form *izuko* is spelled *いずこ* or *いづこ*, and so on.

### 3.2 Attachment of HPI

The HPI of a word entry varies depending on the lemma, form, and orthographic levels: therefore, we should attach HPI to the entry at each respective level.
Figure 4 shows the result of the HPI attachment to the entry いずこ. HPI is displayed as time bars to the right of each entry. The lemma いずこ is used in all historical periods, so the HPI is ‘FROM Old TO Pres.’ The form いずこ was first used in the Early Middle period and is still used today. Therefore the HPI is ‘FROM E-Mid. TO Pres.’ いずこ is the old form of いずこ. The いずこ form has been used in all periods, so the HPI of this form is ‘FROM Old TO Pres.’ and is the same as that of the lemma いずこ. どこ is a rare form and was used only in the Early Middle period. In the present day, いずこ and どこ are spelled いずこ and いずこ respectively. Both いずこ and いずこ are historical kana usage and are no longer in use.

3.3 Consistency

HPI is attached to an entry at each level in the hierarchical structure of UniDic, but the HPI value is not freely selected. HPI should satisfy the following consistency condition: HPI on one level must be equal to or less than the HPI of the higher level.

Figure 5 shows the consistency condition. The HPI of each level inherits the HPI of the higher level as the default value. It is possible to modify the HPI as long as it does not violate the above-mentioned condition.
4 Examples

4.1 Lemma Level
We first see an example of entries with the same meaning but different HPI at the lemma level. Haha 母 (‘mother’) is one of the basic kinship terms in Japanese. In Old Japanese, the word *omo* was also used, but by the Early Middle period, it was obsolete.

<table>
<thead>
<tr>
<th>Lemma</th>
<th>HPA</th>
<th>Old</th>
<th></th>
<th></th>
<th></th>
<th>Pres.</th>
</tr>
</thead>
<tbody>
<tr>
<td>haha</td>
<td>母</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>omo</td>
<td>母</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6. HPI of *haha 母* and *omo 母*

The word *omo* appears only in Old Japanese literature; therefore, we limit this word entry in UniDic to the Old period by assigning the HPI ‘FROM Old TO Old’ (see Figure 6). As a result, the morphological analyzer can use *omo* only when analyzing an Old Japanese text.

The second example concerns entries with the same lemma reading but different meanings and HPI. The word *morokoshi 唐土* means ‘China’, and is used in all historical periods. *Morokoshi 蜀黍* means ‘sorghum’ (a kind of grain) and first appeared in the Early Modern dialect dictionary *Butsurui shōko* (1775). The HPI restricts the word entry *morokoshi 唐土* in UniDic to the Early Modern, Modern, and Present periods (see Figure 7); this entry is used only for analyzing the texts of these periods.

<table>
<thead>
<tr>
<th>Lemma</th>
<th>HPA</th>
<th>Old</th>
<th></th>
<th></th>
<th></th>
<th>E-Mod.</th>
<th>Pres.</th>
</tr>
</thead>
<tbody>
<tr>
<td>morokoshi 唐土</td>
<td>母</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>morokoshi 蜀黍</td>
<td>母</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 7. HPI of *morokoshi 唐土* and *morokoshi 蜀黍*

Thirdly, the following examples are words which have the same meaning (‘lunch’) and are often spelled in the same way (昼飯), but differ in word origin.

<table>
<thead>
<tr>
<th>Lemma</th>
<th>HPA</th>
<th>Mid.</th>
<th></th>
<th></th>
<th></th>
<th>Pres.</th>
</tr>
</thead>
<tbody>
<tr>
<td>chūhan 昼飯</td>
<td>母</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hirumeshi 昼飯</td>
<td>母</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hiruhan 昼飯</td>
<td>母</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 8. HPI of *chūhan 昼飯*, *hirumeshi 昼飯*, and *hiruhan 昼飯*

*Chūhan* is a Sino-Japanese word (*chū* 昼 ‘noon’+*han* 飯 ‘food’) used since the Middle period. *Hirumeshi* is a native Japanese word (*hiru* 昼 ‘noon’+*meshi* 飯 ‘food’) and is newer than *chūhan*. *Hiruhan* is a hybrid word coined in the Meiji period, which quickly became outdated. The use of these word entries for morphological analysis is limited to the historical periods as shown in Figure 8.
4.2 Form level
The first example concerning the form level is the word *dare* 誰, which has multiple forms.

*Figure 9. HPI of dare 誰*

*Dare* 誰 had only the form *tare* until the Middle period, but the form *dare* has been used since the Modern period. This form variation is attributable to phonological change in the Early Modern period when *tare* became voiced. It is possible that the analyzer using UniDic with all entries may choose the incorrect form *dare* for Old Japanese texts. With HPI, the form of 誰 in UniDic is determined uniquely as *tare* before the Early Modern period, and the morphological analyzer can correctly analyze 誰 as *tare* for texts in these periods.

Secondly, we see an example in which the same form (*chikashi*) appears in the two lemmas (*chikai 近い* and *chikashii 近しい*).

*Figure 10. HPI of chikai 近い and chikashii 近しい*

These two lemmas are similar in meaning: *chikai 近い* means ‘near; close; familiar’ and *chikashii 近しい* means ‘intimate; familiar’. These entries have the same form *chikashi*, and the morphological analyzer has difficulties in distinguishing between the lemmas. We attached HPI ‘FROM E-Mod. TO Mod.’ to the form *chikashi* of the lemma *chikashii 近しい*
(see Figure 10) so that the analyzer would use it only for Early Modern and Modern Japanese texts.

Thirdly, we mention the distinction of conjugation types. Conjugation types of verbs and adjectives have changed throughout the history of the Japanese language, with one major change being the transformation of *bungo* conjugation types to *kōgo* conjugation types. This change was progressing in the Early Modern and Modern periods. UniDic differentiates between conjugation types at the form level. The HPI restricts *kōgo* conjugation types so that they are used only after the Early Modern period, and *bungo* conjugation types so that they are not used for the Present period (see Figure 11).

![Figure 11. HPI of *kōgo* and *bungo* Conjugation Types](image)

The conjugation types of some words have undergone minor changes. There are different *bungo* conjugation types (upper monograde/upper bigrade) for the verb *hanahiru* 噛ひる (‘sneeze’). The upper monograde was used only in Old Japanese and was thereafter replaced by the upper bigrade (see Figure 12).

![Figure 12. HPI of *hanahiru* 噛ひる](image)

Using UniDic with HPI, the analyzer can correctly select the upper monograde *hanahu* for Old Japanese texts and the upper bigrade *hanahiru* for the texts of other periods.

### 4.3 Orthographic level

An important example of HPI at the orthographic level concerns *kana* usage. There are two kinds of *kana* usage in the Japanese writing system: historical *kana* usage and present-day *kana* usage. For example, the lemma *kyō* 今日は (‘today’) is written けふ with historical *kana* usage and also written きょう with present-day *kana* usage. Historical *kana* usage was abolished in 1946 by the Japanese government, which adopted present-day *kana* usage instead. Therefore, the HPI of orths with historical *kana* usage is ‘TO Mod.’ and with present-day *kana* usage ‘FROM Pres. TO Pres.’

Some peculiar orthographies carry special meaning. The greeting *yoroshiku* 宜しく (‘nice to meet you’) is sometimes spelled 夜露死苦 by today’s motorcycle gangs. The Chinese characters 死 and 苦 mean ‘death’ and ‘agony’ respectively. Motorcycle gangs select this
orthography to show ironic humor. The orthography 夜露死苦 was produced during the Present period, so the HPI is ‘FROM Pres. TO Pres.’

Effects of HPI limitations

With HPI, UniDic discriminates among the vocabularies used in each historical period. Table 1 shows how the entries in each period are limited. The amount of conjugated orthographies diminished drastically in the Old, Early Middle, Middle, and Present Japanese entries because in UniDic, kōgo conjugation types are used only after the Early Modern period, and bungo conjugation types cannot be used for the Present period.

Table 1 Number of Entries in Each Period

<table>
<thead>
<tr>
<th>Period</th>
<th>Lemmas</th>
<th>Forms</th>
<th>Orths</th>
<th>Conjugated Orths</th>
</tr>
</thead>
<tbody>
<tr>
<td>All periods</td>
<td>240459</td>
<td>270356</td>
<td>448532</td>
<td>1479585</td>
</tr>
<tr>
<td>Old</td>
<td>180428</td>
<td>187871</td>
<td>280421</td>
<td>692061</td>
</tr>
<tr>
<td>Early Middle</td>
<td>183161</td>
<td>191391</td>
<td>286617</td>
<td>715320</td>
</tr>
<tr>
<td>Middle</td>
<td>186357</td>
<td>195317</td>
<td>296935</td>
<td>748205</td>
</tr>
<tr>
<td>Early Modern</td>
<td>188666</td>
<td>212063</td>
<td>339505</td>
<td>1373684</td>
</tr>
<tr>
<td>Modern</td>
<td>202492</td>
<td>229672</td>
<td>349964</td>
<td>1401782</td>
</tr>
<tr>
<td>Present</td>
<td>221322</td>
<td>237365</td>
<td>327670</td>
<td>909864</td>
</tr>
</tbody>
</table>

Using this improved UniDic, we built two sub-dictionaries specialized for specific periods: one dictionary for Early Middle Japanese and one for Kyōgen. The former is based on Early Middle entries and the latter on Middle and Early Modern entries. As shown in Table 2, these dictionaries decrease in size (by half, in the case of UniDic for Early Middle Japanese).

Table 2 Comparison of Dictionary Sizes

<table>
<thead>
<tr>
<th></th>
<th>Source file (Lex.csv)</th>
<th>matrix.def</th>
<th>Binary dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniDic with all entries</td>
<td>315MB</td>
<td>3694MB</td>
<td>800MB</td>
</tr>
<tr>
<td>UniDic for Early Middle Japanese</td>
<td>145MB</td>
<td>832MB</td>
<td>268MB</td>
</tr>
<tr>
<td>UniDic for Kyōgen</td>
<td>294MB</td>
<td>3497MB</td>
<td>747MB</td>
</tr>
</tbody>
</table>

The time for machine learning was shortened by 75% from 577 minutes (using UniDic with all entries) to 142 minutes (using UniDic for Early Middle Japanese).

Tables 3 and 4 show the analysis accuracy of UniDic for Early Middle Japanese and UniDic for Kyōgen compared to UniDic with all entries. Level 1 is the accuracy of word segmentation. Level 2 is about whether part-of-speech tagging is correct in addition to segmentation. Level 3 is the accuracy of lemmatization for the correct entries at Levels 1 and 2. Lastly, Level 4 is the accuracy of discerning allomorphs. The accuracy of the two new dictionaries is equal to or higher than that of UniDic with all entries, despite their compactness.

Table 3 Comparisons of Accuracy (Analyzing Early Middle Japanese Text)

<table>
<thead>
<tr>
<th>Level 1: Segmentation</th>
<th>Level 2: POS tagging</th>
<th>Level 3: Lemmatize</th>
<th>Level 4: Allomorph</th>
</tr>
</thead>
</table>

448
Table 4 Comparisons of Accuracy (Analyzing Kyōgen Text)

<table>
<thead>
<tr>
<th></th>
<th>Level 1: Segmentation</th>
<th>Level 2: POS tagging</th>
<th>Level 3: Lemmatize</th>
<th>Level 4: Allomorph</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniDic with all entries</td>
<td>0.9885</td>
<td>0.9698</td>
<td>0.9608</td>
<td>0.9590</td>
</tr>
<tr>
<td>UniDic for Kyōgen</td>
<td>0.9887</td>
<td>0.9700</td>
<td>0.9613</td>
<td>0.9595</td>
</tr>
</tbody>
</table>

6. Conclusions

We added HPI to each entry level of UniDic. UniDic restricts the word entry so that it is used only in its appropriate periods. The morphological analyzer uses these restricted entries and can then efficiently analyze the texts of each period. This improvement to UniDic brings about three beneficial effects: (a) more compact sub-dictionaries; (b) a reduction in time for machine learning; (c) an equal to or higher analysis accuracy than that of UniDic with all entries. We plan to attach more precise HPI, which will promote accuracy when analyzing new texts.

References


Portable Electronic Dictionary (PED) Use in Spanish as Second Language Students in Japan

Lidia Sala Caja
Aichi Prefectural University, Japan
sala@for.aichi-pu.ac.jp

Abstract

It is well known that Japanese students of English are massive users of pocket electronic dictionaries (PED). The same can be said of Japanese students of other foreign languages, despite the fact that they begin their second language acquisition later in their education at university, and that fewer dictionaries are available to them (for the Spanish this is limited to three bilingual dictionaries and one monolingual). It is understandable, therefore, that their user profile or habits have been largely ignored. Recently however, some teachers have felt the need to adjust their teaching of lexicographical skills to the PED, while others remain doubtful. To shed some light to this question, I prepared a questionnaire for Aichi prefectural University students of Spanish. This paper presents some of its results, focusing on the user profile of students of Spanish as first foreign language.

Keywords: portable electronic dictionaries, dictionary use, Spanish, Japan

1. Introduction

The portable electronic dictionary (PED) is a common learning tool in Japan. With already 36 years of history (Tono 2009), it has reached many specialization fields (medical, economical, etc.) and potential users (secondary, tertiary students and professionals). Since September 2004, Japanese learners of Spanish also have their pick, although the number and type of dictionaries available to them are fewer than those for English learners ones. The offer for 2015 was limited to two bilingual Spanish-Japanese dictionaries, one bilingual Japanese-Spanish and one bilingual bidirectional Spanish-English. As for monolinguals dictionaries, a few years ago Casio launched a CD with the Gran Diccionario de la Lengua Española Larousse (2003) that could be installed to some models, and Sii includes a Spanish bilingualized dictionary, the Spanish Kernermar 西西和辞典.

In the Japanese education system, there is virtually no Spanish instruction in high schools. Learning begins at the university as a first or second foreign language. The student’s level when they start is therefore absolute beginner and eventually those in charge of such students find themselves in the situation of having to teach them the basic mechanics of dictionary lookup. Traditionally, paper dictionaries were used to give such guidance but, as PED became more popular in the Spanish classroom, doubts arose among teachers as to which dictionary support was more convenient or what should be taught in class and how.

In order to find a way out of that impasse, following earlier studies (Kobayashi, 2008; Ronald and Ozawa, 2009; Chen 2010; Boonmoh, 2010) I designed a questionnaire with the purpose of obtaining an initial user profile of students and teachers to provide the latter with useful information so they could fine tune their recommendations and guidance regarding dictionaries. Note that it is an exploratory approach, it was not intended to evaluate their actual PED use or its impact on learning, which is a following endeavour.

1 The first one was the Casio XS-HA01.
3 Oxford Spanish Dictionary
4 Same concerns expressed Bower and Mc Millan (2007) and Loucky (2010).
In this paper, after a brief description of the questionnaire and its results, I will focus on some differences found between first and fourth year student answers regarding experience and sense of improvement.

2. The questionnaire
The questionnaire was addressed to three different respondent types: teachers, first foreign language students (i.e. from the Spanish Studies Section) and second foreign language students (i.e. from other sections, departments and faculties). Items targeted ownership, frequency of use, experience, training, attitudes and believes. It was divided in three sections: (1) questions as owners and users of dictionaries, (2) as owners and users of a PED and (3) as Spanish teachers/learners. Answers could be closed, open-ended, partially categorized or a 6-degree Likert scale. Most of the items in parts (1) and (2) were identical in all questionnaire types while part (3) items are somewhat different, I tried to keep them symmetric in order to be able to compare their results afterwards⁵. To minimize misunderstandings, students handled a Japanese version and teachers, a Spanish one.

It was conducted at Aichi Prefectural University (Japan) on July 2014. It was answered by 19 teachers, all of them in charge of Spanish language courses (conversation, reading, composition, grammar), and by 385 students: 211 from the Spanish Studies Section and 174 from the rest of the University. After, I maintained informal conversations with 10 teachers. In this occasion, I will only report mainly on data obtained from the second ones.

3. Results
3.1. Spanish first language student profile
Data obtained were similar to recent surveys on PED use in Japan for English learners. Like Kobayashi (2006: 227), we found that “the majority of students owned an ED, and those who owned it tended to use it exclusively”: ownership started at 96% in first year and reach 100% before graduation⁶. Frequency of use only decreased in 4th year, being “every day” the most common answer (77%) followed by “several times a week” (20%). Also, like Kobayashi (2008) and Nesi (2002), we found that they were satisfied with their dictionaries (73% satisfaction overall), but not completely a-critical. Participants no longer considered their PED a waste of money nor a symbol of status, but a necessity to follow language studies⁷, which anyway did not prevent them from making it an unreflecting purchase⁸.

Compared to previous surveys, perceived PED pros with respect to paper dictionaries have increased. Respondents considered them more advantageous not only for reading (79.2%) but also for writing, with a 20.8% average for “strongly agree”, 27.5% for “agree” and 23.7% of “somewhat agree”⁹. This trend can be explained if it is taken into account that recent PEDs contain complete editions of renewed bilingual dictionaries that improved design and content to better fulfil reception and production tasks. This, allied with the proverbial fast access to information offered by PED, makes their preference logical. Interestingly, though, its wonders did not reach vocabulary learning, where disagreement outnumbered agreement by 7%¹⁰.

---

⁵ For example, the teacher version asked “Do you give any dictionary guidance in class?” and the student version, “Have you received any dictionary guidance in class?”
⁶ Bower and McMilland (2007), for example, collected similar percentages in foreign language learning faculties.
⁷ The statement “Buying a PED is a waste of money” only received 6% of some sort of agreement. The item “Students buy a PED because it is a symbol of status” presented a 43% for “strongly disagree”; 13% for “disagree” and 25% for “somewhat disagree”. Teachers’ questionnaires exhibited similar averages.
⁸ 47% checked the option “Because it was sold in the University shop” as the first reason to purchase his actual PED and another 24% “Because it included a dictionary that I wanted” (i.e. the Spanish ones). Only 10 students (out of 201) mentioned teacher advice as reason to purchase showing lack of guidance on this point.
⁹ Weschler and Pitts (2000) found that students seldom used PED for production, like Chen (2010) and Jian’s advanced students (2009). Conversely, Ronald and Ozawa (2009) participants reported to use often their PED for both writing and reading.
¹⁰ The teachers, strikingly, expressed a little more faith in PEDs as vocabulary learning tools, 58% thought it could be helpful.
As for other dictionary ownership, 43.6% purchased a paper dictionary (PD) and 32.7% access other electronic or online dictionaries but, as table 1 shows, abandonment proliferates and happens fast with PD while frequency of use resists within the other digital options: around 50% declared to consult them at least several times a week. Results on this point differ from Kobayashi (2008) and Chen (2010), where PD ownership and frequency of use were higher.

Eventually, being in possession of multiple lexicographic resources is becoming the norm. Looking at the figures, the most common combination is PD + PED (29.3%), followed by PED + other ED (13.7%) but it must be born in mind that many of that PD are actually out of use, unlike their other electronic counterparts. Thus, it appears that we are in a transition phase, where the more widespread combination\(^\text{11}\) of PED + PD is being replaced by PED + other electronic type dictionaries.

As for the monolingual choice, online dictionaries and applications helped raise its numbers\(^\text{12}\), being the most remarkable fact that even a few first year students declared to have one. Its low frequency of use, however, ends up depicting the average Japanese foreign language student, as being highly dependent on the bilingual dictionary.

3.1. Experienced users

Experience has been considered an element that may propel or inhibit efficient ED usage. For example, Kobayashi (2006) mentioned habit transference and lack of experience to justify that no significant difference was found between data from the PED and PD users. On the other hand, Ronald and Ozawa (2009) listed familiarity with PED as a factor that might affect user behaviour and Loucky (2010:169) concluded that computer and engineering major students more efficient use might “depend on the individual student’s level of English proficiency and experience using computers”. In fact, the role that digital literacy plays in electronic dictionary use is getting more attention from researchers and scholars recently\(^\text{13}\).

Also, it has been long observed that there is a correlation between experience, language knowledge and paper dictionary usage: the more you handle them, the more successful you are in your lookup; the more you know about a language, the more skillful you will be interpreting the information contained. In Jian’s 2009 survey low proficiency students preferred PED while high ones rely more on PD, which made him conclude that language level affected the learner needs towards PED. And finally Okuyama and Igarashi (2007) and Loucky (2010) hypothesized such correlation too for EFL Japanese students using PED\(^\text{14}\).

In the questionnaire participants were asked whether they had more than two years experience using PED: only 15% (N=211) gave a negative answer. Also, they were asked about previous PED familiarity and ownership to find out how many years of experience using PED they had. For 112 undergraduates (54%), the device on use was not the first one. Not surprisingly, as this tendency has already been observed and described, if only first year students were considered, average raised to 61%. Many of them purchased their first PED when they started senior high school, so by the time they join the Spanish classroom, they have already accumulated two or three years of experience.

\(^{11}\) Kobayashi (2008): “ED owners tended to use an ED almost exclusively although they also owned a PD”.

\(^{12}\) 50 students declared to have one, but only 6 of them in printed form.

\(^{13}\) See for example Aguilar Escobar (2006), or Lew (2014), who explicitly mentions digital literacy and information literacy as a factor that influences performance.

\(^{14}\) “What is more likely is that less proficient language learners are less experienced and less proficient users of EDs” (Loucky 2010:158).
These data carry two implications, the first one, already pointed out by Yukio Tono, is that students have short experience in handling paper dictionaries, and therefore limited influence from that should be expected. The second one is that, unlike in English foreign language learning, for many of them there will be no proportional correlation between language (in that case Spanish) level and PED handling: our beginner learners come to class with rooted usage habits, which might affect their pattern of use. Their senior’s, on the other hand, will offer a better insight of how constant PED use and progress in Spanish influenced the acquisition of new reference habits as well as the bettering of skills and, consequently, their perception of progress as users.

3.2. Progress and experience

According to the questionnaire, more students (63.3%) than less believed they had improved their dictionary skills, but, intriguingly, the gap between positive and negative answers was smaller by far among first years (6 point difference against 33, 20 and 55 points in the following school years). Moreover, in previous research it has been found that students perceived themselves as competent PED users, and that they feel they did not require any further training. The total agreement obtained in such statement (65%) would take us to the same conclusion. But when analysed by entrance year, significant differences were found between junior and senior students (vid. table 2): 70% of the former were much more prone to skip training while only 48% of the latter thought likewise.

Explanation of such data among first year students cannot be Spanish related, but must come from their previous experience with PED, or tangentially from familiarity with other digital environments. An instance of which is the ability to transfer PED usage habits from their secondary education years.

Evidence of such transfer was found in the students’ belief that PED is a useful tool to learn how to pronounce Spanish. Teachers, specifically native ones, disregarded it and consequently gave no relevance to the sound option (no one checked it). Certainly, phonetic transcription has not been a traditional feature in Spanish paper lexicography, the reason being that grapho-phonemic correspondence is pretty straightforward in this language. Furthermore, Japanese and Spanish phonological systems are very much alike and it does not entail a big obstacle for Japanese learners. Nevertheless, students believe PEDs play an important role in supporting pronunciation (69.9% agreement average, 74% in first year) and exhibit a consistent wide choice of the sound function. However, discrepancy with the instructor group can be explained if usage transfer is considered: influenced by their previous PED usage habits of studying English, they disregarded the fact that the sound function may

15 For Japan, Kobayashi (2008: 103) has even stated “some younger learners have never even used a printed dictionary” to learn English.
16 Learning needs towards PED increased significantly in the last two University years: specialty seminar classes, where extensive and dense reading takes place, start in third year; and in their last year, students take a compulsory class on Academic Writing.
17 In Ronald and Ozawa’s (2009) survey 70% of the participants said no to guidance.
18 The difference in proportions is significant, $x^2(1, N=199) = 9.5414, p<0.05$.
19 Phonetic transcription is absent even in some Spanish learners paper dictionaries like Clave or Diccionario Salamanca de la lengua española, and online dictionaries, like IULA’s Diccionario de aprendizaje del español como lengua extranjera (http://www.iula.upf.edu/rec/duele/). For Humberto Hernández its entrance in the Spanish dictionaries “se explica por una razón puramente mimética hacia otras lexicografías de mayor prestigio” (Hernández, 2000: 101). It must be said, though, that phonetic transcription is present in the learner’s pioneer Diccionario para la enseñanza de la lengua española and, naturally, in the bilingual Spanish-Japanese dictionaries.
20 In EFL literature, Bower and McMilland (2007) complained about little student awareness and use of sound function, but admitted their popularity, Kobayashi (2010) and Chen (2010) later studies on the other hand found that it was a used function, perceived as one of the PED greater utilities. In our questionnaire, sound function takes the lead as the most employed one among freshmen (38.7%) and it was checked often alone and not very far came the hyper-textual jump (36.6%).
be superfluous when studying Spanish but nevertheless remember its usefulness and resort to it.

Another advantage of prior PED experience is that it allows them to easily acquire lexicographical habits that fit their needs with the new language. This is the case of the inflected form lookup, an available option in recent PED models to satisfy the users need to cope with the Spanish rich verbal morphology21. Data gathered in the questionnaire showed that this type of search was favoured by first year students even ahead of the word meaning lookup (vid. table 2), drawing more a cliff-shaped feature acquisition process than an s-shaped one22. Although its importance decreases the following years, as one would expect, it still occupies a remarkable position in their lookup preferences in the following years, overtaking search by combination of words or by example some years.

The means first year students declared to use in order to improve their PED usage hold another clue as to why beginners can be so sure about their dictionary skills. On the top of the list is “by themselves” (57%) and then “thanks to my classmates” (23%), teachers receiving only a 9% (that is, 6 respondents). Later on, the perception of being on their own is unchallenged (65% average; 58% in forth year) but classmate influence leaves pass to an increasing teachers presence in third and forth year (15% and 25% respectively). This was to be foreseen because of the growing importance of reading and composition in that period, so instructor hints about dictionary use became both more frequent and more relevant for them.

In short, the possibility of recycling old habits plus the fast, autonomous and effortless acquisition of new ones that allowed previous acquaintance with PED, might explain no only why PD frequency of use plummeted23 despite teachers active efforts to encourage its use, but also the student’s overconfidence and limited prospects of progress.

Unlike them, as much as 74% of the last year students were convinced that the way they handle PED had changed. Growing figures about PED being used alongside other lexical tools as well as the number of dictionary functions used seems to back up such perception24.

Another sign of user evolution was found in the nature of PED criticism. Kobayashi (2008:109) also noticed this point. She reckoned that students’ perception of PED advantages “progressed from the obvious (i.e. speed) to more subtle aspects (i.e. jump options)”. Indeed first year student comments concentrate on the device (speed, information on screen), but from the second year onwards, respondents’ remarks concentrated on PED content (like the lack of a built-in monolingual or a thesaurus in Spanish) and in the last two years fine reflections about the lexicographic works (misleading examples, translation accuracy) appeared. And nevertheless, they feel they need training.

Function types remain unchanged along the four years: the first two positions are occupied by jump search and sound function, all of them with relatively small figures: an average of 56.2% students chose the former among the three most resorted functions and 49%, the latter. Final year results witness a noteworthy raise of the jump function (62%) as well as a drop in the frequency of use of the sound function (40%). However, the still important presence of this function, a non-crucial Spanish learner feature, could be interpreted negatively at this point, as an habit inertia. Furthermore, bigger numbers in other functions like history search,

21 This function has raised some controversy. Lew (2014) foresees that it will make redundant the skill of “deciding on the appropriate form of the look-up item” and sees in this type of search the advantage of reducing look-up failure, but some interviewed teachers loath it because they believed it hinders verb conjugation acquisition.
22 Lew (2015:7-8) proposed a s-shaped learning curve to illustrate the process of learning how to use a dictionary.
23 Kobayashi (2008) reported similar findings in her survey.
24 62% of last year students reported that they conducted more combined searches than two years ago and 74% of them declared using more functions.
only mentioned by two students, or flashcards are missed. This data resembles considerably that of other surveys which concluded that students do not get the optimum use out of their dictionaries.

Concerning lookup type, the slow progress towards the second place of word combination search that table 2 describes suggests decoding-related skill improvement but the bottom place occupied by example search points at a stagnation of encoding-related skills. Without either a collocation dictionary or an easily accessible learner monolingual built in the PED, it is mandatory to look at the examples provided in the bilingual dictionaries to know the usage of a word25, but not all the entries provide one and they might be displayed on a different screen. Search by example is essential to encoding because through the accumulated sentences students access relevant information (collocation prepositions choice, transitivity related issues) about a word. Moreover, as the example comes with a Japanese translation, the sentence comprehension burden disappears.

With the exception of first years only, example lookup appears in last place of the students’ preferences, which further illustrates the difference between knowing and using a dictionary feature. Impossible to miss in every bilingual dictionary first screen, PED experienced user students are well aware of it, but skip it anyway. It is likely that they are unable to interpret the information that it contains, or fail to apply it to the writing task. It is only when faced with the reality that their digital experience is not sufficient to fulfil their language requirements, that they realised both their own limitation as users and the need of proper dictionary training.

**Conclusion**

In this article, I have reported on the results of a questionnaire administered to learners of Spanish as a second foreign language in Japan. Although the knowledge obtained, based on self-reporting, can be only illustrative to a certain degree, it has been useful to notice some general trends than might benefit future teaching. Data gathered showed that Japanese learners of Spanish might have arrived late at the world of PEDs but are catching up fast despite their limited range of choice. Furthermore, high PED ownership and steady if limited frequency of use of other electronic dictionaries stood out against a shrinking PD use average. It would therefore seem prudent for teachers to shift their first dictionary guidance towards the electronic support accordingly.

Our learners also exhibit weaknesses found in their ESL mates like the reckless consultation glimpsed through the excessive and long-lasting importance given to the inflected form lookup. Also, under-exploitation of the PED’s capabilities, common to other educational settings, could be confirmed through the scarce use of functions like history search or the relegated position of search by example.

However, thanks to previous familiarity with PED and other digital environments, an initial good knowledge of the device is guaranteed from the beginning of their Spanish studies. It allows them to incorporate useful new features and lays the foundations for their usage pattern in that language. Teachers can also take advantage of that situation by focussing less on the mechanics of the lookup, which are already known, and more on understanding the dictionary conventions or the component parts of the entry that might be new for them (like the noun gender abbreviations f. and m.).

Nonetheless, as we have seen, digital experience does not assure a steady progress as users nor does the incorporation lexicographic habits automatically equate to an improvement of the skills required for successful consultation. The students’ evolution as users is challenged by the fact that their Spanish learning experience is not accompanied by training. Given the

---

25 In monolingual consultation, “the most important source of information seems to be the examples rather than the explicit grammatical information” (Chan 2012: 90)
choices at hand, teachers should centre their efforts in skills such as deriving and applying information from examples. Research on electronic dictionaries from the user perspective is still incipient in Spanish, but very much needed. In order to gain knowledge about actual dictionary use questionnaire results like the one I have reported must be rounded up by further research using other data collection means. Meanwhile, the teaching of reference skills should not be forgotten, and specific training that will adjust to the Spanish learner profile should be developed.

Table 1 Paper dictionaries frequency of use

<table>
<thead>
<tr>
<th>DICTIONARY TYPE</th>
<th>FREQUENCY</th>
<th>ENTRANCE YEAR</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2014</td>
<td>2013</td>
</tr>
<tr>
<td>Bilingual Spanish-Japanese</td>
<td>Once a day</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2, 3 times a week</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3, 4 times a month</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>From time to time</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Never in 6 months</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>19</td>
<td>13</td>
</tr>
<tr>
<td>Bilingual Japanese-Spanish</td>
<td>Once a day</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2, 3 times a week</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3, 4 times a month</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>From time to time</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Never in 6 months</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>15</td>
<td>10</td>
</tr>
</tbody>
</table>

Last year students comprise 2011 and 2010 to include students that spent one year abroad. Data is not presented in percentages because raw frequencies were too small.
Table 3  Lookup type⁶

<table>
<thead>
<tr>
<th>LOOKUP TYPE</th>
<th>Meaning</th>
<th>Inflected form</th>
<th>Word combination</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st option</td>
<td>2nd option</td>
<td>1st option</td>
<td>2nd option</td>
</tr>
<tr>
<td>ENTRANCE YEAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>18</td>
<td>22</td>
<td>23</td>
<td>12</td>
</tr>
<tr>
<td>2013</td>
<td>23</td>
<td>7</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>2012</td>
<td>29</td>
<td>0</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>2011</td>
<td>18</td>
<td>3</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>2010</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>TOTAL 1st opt.</td>
<td>99</td>
<td></td>
<td>42</td>
<td>3</td>
</tr>
<tr>
<td>TOTAL 2nd opt.</td>
<td>32</td>
<td></td>
<td>51</td>
<td>38</td>
</tr>
</tbody>
</table>

Most common 1st option

Most common 2nd option

<table>
<thead>
<tr>
<th>LOOKUP TYPE</th>
<th>Meaning</th>
<th>Inflected form</th>
<th>Word combination</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3rd option</td>
<td>4th option</td>
<td>3rd option</td>
<td>4th option</td>
</tr>
<tr>
<td>ENTRANCE YEAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2013</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2012</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>2011</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>2010</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>TOTAL 3rd opt.</td>
<td>5</td>
<td></td>
<td>13</td>
<td>56</td>
</tr>
<tr>
<td>TOTAL 4th opt.</td>
<td>4</td>
<td></td>
<td>29</td>
<td>34</td>
</tr>
</tbody>
</table>

Most common 3rd option

Most common 4th option

⁶Some respondents (43 out of 202) failed to mention the preference order of each lookup type. The table only includes the ranked answers.
References


The Design of an App for Learning Chinese Idioms

Shan Wang*, Yujia Wu
The Hong Kong Institute of Education
swang@ied.edu.hk, s1114428@s.ied.edu.hk

Abstract
The effective use of Chinese idioms allows for effective communication and adds vigor to the language. However, due to the richness of idioms, foreign learners encounter many difficulties in grasping them. Unfortunately, existing resources provide limited help. First, existing printed and electronic dictionaries have a number of drawbacks in showing the information of idioms: (i) they tend to list senses randomly; (ii) contain little or no information concerning the grammatical function; (iii) list synonyms or antonyms in isolation. To resolve this problem, the proposed App provides rich information concerning idiomatic language, including: (i) dynamic sense arrangements, (ii) the frequency of grammatical functions, and (iii) the lexical semantic network. Second, the App provides multimedia functions to enrich students' learning experience. Third, most Apps are for reference only and very few are designed for learning idioms. Our app is designed to have an interactive learning environment, including (i) the self-test system, (ii) peer assessment, and (iii) online games. In sum, the newly designed idiom-learning App can help students efficiently grasp idioms.

Keywords: App, sense, grammatical function, lexical semantic network, learning environment

1. Introduction
With the development of technology, language learners have been transforming their way of learning (de Schryver 2003, Chen 2010, Dziemianko 2010), such as using Apps on smartphones and tablet computing devices. Though there are some Chinese-English dictionary Apps in App stores, their function is mainly for reference, not for learning. Chinese idioms have deep cultural connotations and are highly generalized and expressive. Therefore effective use of them brings convenience to communication and adds vigor to the language. However, due to the richness of idioms, foreign learners encounter many difficulties in grasping them (Liu 2013b, An & Wang 2015). This paper introduces the design of a user-friendly App for learning Chinese idioms. Compared to existing resources, the design has three distinctive features: (i) rich and dynamic information of idioms, (ii) multimedia functions, and (iii) interactive learning environment.

2. Related research
This section reviews related research from three respects: semantic and syntactic functions of idioms, printed and electronic idiom dictionaries, and lexical semantic network.

2.1 Semantic and Syntactic Functions of Idioms
It is common that learners make errors in using Chinese idioms, such as semantic, syntactic and pragmatic errors (Zhang 1999, Yang 2011, Shi 2014). Liu (2013a) collected 77 idiom errors from the in-class and after-class homework of Year 3 Chinese learners at Beijing Language and Culture University. She found 16 semantic errors (20%); and 24 syntactic errors taking up (31%). Moreover, she also examined 64 idiom errors in “HSK Dynamic Composition Corpus”1, out of which 17% are semantic errors and 34% are syntactic errors. The two results indicate that semantic and syntactic errors account for over half of idiom errors. An in-depth study to the semantics and syntactic functions of idioms is essential to help learners reduce errors and improve the accuracy in using Chinese.

2.2 Printed and Electronic Idiom Dictionaries

There has been an increasing number of learners who take advantage of electronic dictionaries (EDs). Zarei and Gujjar (2012) summarized 10 disadvantages of printed dictionaries (PDs) and 12 advantages of ED. In addition, many studies have reported that students find ED beneficial and facilitative for their learning (Nesi & Haill 2002, Golonka et al. 2014).

2.3 Lexical Semantic Network

Everyone who acquired a language has a mental lexicon. Meara (1982) and Meara (1984) found that there might be significant differences between the mental lexicon of a second language learner and that of a native speaker. The connections between words in learners’ mental lexicon are less stable than in native speakers; the semantic links between words in learners’ mental lexicon differ systematically from those of native speakers.

These findings reveal that second language learners have a much weaker mental lexicon compared to native speakers. Hence, constructing a lexical semantic network through systematic word relations will evidently benefit second language learners to achieve the proficiency of native speakers. However, none of the existing idiom Apps has such a function.

3. The reasons of designing the App

Chinese Idioms have their historical and cultural connotations, and thus many learners are highly interested in this area. However, many students find them difficult to learn because a large number of them have two layers of meaning and unique cultural characteristics. In addition, EDs are mainly for native Chinese checking senses. There is very little software designed for student to learn idioms. To solve these problems, this paper aims to design an efficient idiom App to help students learn them.

We conducted a survey on existing PDs and EDs for idiom learning before designing the new idiom learning App.

(1) Xu (2006):
【按图索骥】àn tú suǒ jì
索：寻求。骥：好马。按照图像寻求好马。比喻办事死板，拘泥成规。也比喻按照已有的线索去寻找事物。
语本《汉书·福传》“察伯乐之图求骐骥于市”。郭沫若《苏联纪行·七月二十二日》：“贵族写给他宠姬的信里泄露了这项秘密……～，全部都被收获。”△
“骥”不读yì。

(2) Huang (2008):
按图索骥 àn tú suǒ jì
【典出】明·杨慎《艺林伐山》卷七载：春秋时，秦国有个相马高人伯乐，此人根据自己多年的相马经验写了《相马经》一书并配有插图。伯乐有个愚笨的儿子，一次他决定根据书中的插图去寻找良马。刚一出门，他就看到田埂上有只大蛤蟆，额头隆起，两眼突出，酷似《相马经》中好马的样子，于是将其带回家中，并对伯乐说：“我找到了一匹好马，其头和眼如你书里的画一样，只是蹄子不同。”
伯乐见后，哭笑不得，叹道：“此马好跳，不堪御也。”
【典义】索：寻找。骥：骏马。按照画像去寻求好马。比喻墨守成规办事。也比喻按照线索去寻求。也作“按图索骏”。
【例句】我们按图索骥，终于发现了藏宝之处。

(3) Tan (1993):
按图索骥 àn tú suǒ jì
指按照线索去寻找好马。也比喻办事机械，死板。to look for a steed with the aid of its picture // to locate sth. by a plan/chart // to try to locate sth. by following up a clue //
to search for sth. based on the hints given

例：每见一班~者，多失于骊黄牝牡。（《葬书问对》）Things are often like this: Those who look for a steed with the aid of its picture will probably fail to recognize a good one, even differentiate a mare from a stallion. / 他们→, 提审了严家忠, 攻下了曹约翰, 然后才杀回马枪找到田玉堂。（方·内）F ollowing up one lead after another, they interrogated Yan Jiazhong, launched an attack against John Cao, and ultimately moved in for the kill on Tian Yutang.

Regarding PDs, we surveyed three types of dictionaries: student idiom dictionaries, such as Xu (2006) in (1); idiom allusion dictionaries, such as Huang (2008) in (2); bilingual dictionaries, such as Tan (1993) in (3). Take one widely used idiom 按图索骥 àntúsuǒjì as an example, (1) simply lists three senses followed by its sources and one example sentence. The drawbacks are: (a) it is hard for learners to know which sense is used by the example sentence; (b) the source is shown by only one short sentence written in classical Chinese, which is difficult for students to understand. (2) introduces the story first, followed by its senses and one example sentences. Though it has a complete story of the source, it has the drawback of (1)(a) as well. (3) has bilingual interpretation, but there is no source of this idiom and its allusion. Moreover, the English does not match the Chinese very well. For example, the English sense “to locate sth. by a plan/chart” is not in the Chinese interpretation. Vice versa. The Chinese metaphorical meaning “也比喻办事机械，死板” is not explained in English.

Furthermore, PDs are not easy to follow the changes and get updated. They have a mono-way of explaining the senses through listing them directly, lacking relations between them. Therefore it is difficult for students to grasp the main meaning and understand idioms in context.

Regarding idiom Apps, there are already some, but most are small game Apps for native Chinese to play with idioms. There are very few Apps aimed at learning idioms and the existing ones are in Chinese with no other language translation. Below let’s see an idiom dictionary on Apple shown in (4): Multi-function Idiom Dictionary. Due to the large capacity of EDs, its explanation of senses and functions are much more abundant, including the English translation, specific usages and sample sentences.

But there are still some shortcomings. (i) There are no rules when the App lists the senses and thus it is not conducive for students to understand and remember them. (ii) The "usage" says that this idiom is negative, but the sense that it lists is not negative. (iii)Though it provides two sample sentences, it does not indicate the grammatical function of this idiom and its frequency of use. Hence it is not contributing for students to learn how to apply the idiom. (iv) It lists the synonyms and antonyms, but these isolated idioms are difficult for students to remember one by one.
Therefore, designing an idiom learning App that is more scientific and efficient in helping students learn idioms is of urgent demand.

4. Ideas of Designing an App for Learning Chinese Idioms

This section describes how to design a Chinese idiom learning App to help learners understand and apply idioms accurately. Compared to existing resources, the design is comprised of three key features: (i) rich and dynamic information of idioms, (ii) multimedia functions, and (iii) interactive learning environment.

4.1 Rich Information of Idioms

The App improves idiom information in three aspects: dynamic sense arrangements, frequency of grammatical functions, and lexical semantic network.

4.1.1 Dynamic Sense Arrangements

Many idioms not only have a literal meaning, but also always have an extended meaning. The two are linked implicitly. Due to such features, there is a need to help students build relations among each meaning and enable them to remember. This was our main concern when arranging the senses. In achieving this, this App contains a dynamic sense arrangement system for the polysemous idioms. Multiple senses can be ordered by their historical development, logical relation and frequency.

(i) Sense Ordering by Historical Development

Polysemous idioms have a certain sequence in forming the senses, so if they are listed by chronological order, learners can clearly understand the origin and the development. Take 按图索骥 àntúsuǒjì as an example, the design is like the following: First, list its original meaning, which is interpreted literally: to look for a steed according to a picture. This is the basic layer of meaning. The story of it can also be followed by this sense. Second, the extended meaning 1: based on the original meaning, a metaphorical meaning arises, which is analogous to mechanical work and rigid dogma. The source of this sense can be listed below it with translation. Third, the extended meaning 2: this idiom has come out with another sense, indicating find things according to the clues. The historical development of the senses can be shown below:

1. look for a steed with the aid of its picture
2. do things rigidly
3. try to locate something by following up a clue

For each sense, we can give examples to help students understand. The following example shows the second extended meaning:

旅客按图索骥,尝尽香港的经典美食。(CCL)
Lǚkè àntúsuǒjì, cháng jǐn xiānggǎng de jīngdiǎn měishí.
“Visitors follow the clues and taste all the classic Hong Kong cuisine.”

(ii) Sense Ordering by Logical Relations

Existing dictionaries do not show the logical connection between senses. However, idioms have holistic semantics, and thus if you can sort out the logical relation, it will help students remember the idiom better. For example, 按图索骥 àntúsuǒjì has this logical relation:

1. look for a steed with the aid of its picture
2. do things rigidly
3. try to locate sth. by following up a clue

First, the original meaning “look for a steed with the aid of its picture” has the implication that such behavior is inflexible, rigid, so its ironic metaphor is that doing things rigidly.
Second, in the original meaning, the picture is the clue of looking for the steed, so the meaning is extended to “try to locate sth. by following up a clue”, which is not negative.

(iii) Sense Ordering by the Frequency

Due to the complexity of doing sense annotation, there is no Chinese idiom dictionary that is compiled according to the sense frequency. Among the consulted dictionaries and Apps, when ordering the senses of 按图索骥 àntúsuǒjì, most of them just introduced two senses. That is, some even do not have the third sense. These dictionaries show that the major sense is the negative one: "doing things rigidly." But according to the corpus of "Center for Chinese Linguistics PKU" (CCL), we get an unexpected results. Among a collection of 130 sentences with 按图索骥 àntúsuǒjì, we annotated all their senses and the result is shown in Table 1. It shows that the most commonly used sense nowadays is the third one, which developed later than the other two senses and it is not negative at all.

<table>
<thead>
<tr>
<th>Senses</th>
<th>No.</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>117</td>
<td>90.00%</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>9.23%</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.77%</td>
</tr>
<tr>
<td>Total</td>
<td>130</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Therefore, when learning idioms, if students can see the permutations of usage frequency, it will be beneficial for them to grasp and use the idiom quickly. Furthermore, senses frequency statistics can also help lexicographers effectively improve dictionary’s quality. In the case of 按图索骥 àntúsuǒjì, they won’t say that this idiom is often negative and they do will list the third sense.

With the three kinds of sense order available, learners can click a button to get a way that they want. At the very beginning, to understand the idiom, the historical order and logical order can be more helpful. After they know its meanings, the most important will be which sense is used more often, and therefore the frequency order will be more useful. No matter which way is used, showing the connection among senses is beneficial for students. Besides, we need to mention that stories, pictures and audios are provided under each sense as well for learners to have a better learning experience.

4.1.2 Provide the Frequency of Grammatical Functions

For the learners who learn Chinese as a second language, the main purpose of learning idioms are not only to read and remember but also to use. Since Chinese is not their native language, they cannot use idioms and make sentences by their language sense, but they need to use their learned sentence structure and grammar to apply them.

At present, although there are some idiom dictionaries that have already listed the grammatical functions, most of them are casual. That is, only list them separately in each selected example sentence. It has obvious disadvantages. First, whether the selected example is representative. Second, certain grammatical functions are common but the dictionary just did not choose such a sentence, so it is impossible to mark the function. Third, it is difficult for students to determine which function is often used.

To avoid such problems, the idea of our design is to annotate the grammatical functions to get the frequency information. When providing examples, select representative ones and give more examples for the most frequent function, listing them from high to low frequency. We annotated the grammatical functions of 130 sentences of 按图索骥 àntúsuǒjì from CCL. The result is shown in Table 2.

---

Table 2 The Frequency of Syntactic Functions of 按图索骥 àntúsuǒjì

<table>
<thead>
<tr>
<th>Syntactic functions</th>
<th>No.</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicate</td>
<td>91</td>
<td>70.00%</td>
</tr>
<tr>
<td>Adverbial</td>
<td>21</td>
<td>16.15%</td>
</tr>
<tr>
<td>Attribute</td>
<td>9</td>
<td>6.92%</td>
</tr>
<tr>
<td>Object</td>
<td>5</td>
<td>3.85%</td>
</tr>
<tr>
<td>Subject</td>
<td>4</td>
<td>3.08%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>130</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Table 2 indicates that being a predicate is its main function, followed by being an adverbial and attributive. It can occasionally be an object and subject. According to this result, when compiling a dictionary, it is better to give priority to providing sentences with it as a predicate. In this way, students can master the most frequent usage first. It can help students to reduce the learning burden and have a more effective learning experience.

4.1.3 Establish the Semantic Network

Though there is no consensus about the exact number of Chinese idioms, it is commonly agreed that Chinese has a large number of them. Some recently published dictionaries collected about 15000 (Dictionary Compilation and Research Group 2014) or even 24800 (Wang 2010) idioms. How to effectively help students memorize so many idioms? Some existing App lists the synonyms, antonyms of an idiom, but these isolated idioms are very difficult to remember effectively. One of the special part of our App design is to create idiom semantic network through meaningful connections and design games to help learners consolidate them. This can enrich learners’ mental lexicon as well.

阳春三月, 春光明媚, 正是游玩赏花的好时节。正逢植物园开花展, 只见姹紫嫣红、春意盎然。芬香扑鼻的迎春花摇曳生姿, 似在欢迎客人的到来; 冰肌玉骨的梅花, 历经风霜, 仍傲然挺立; 含苞待放的海棠花, 羞羞答答, 不愿让人看见它的真容; 又怎能少了那雍容华贵、富丽端庄、芳香浓郁的“花中之王”——牡丹。兴致勃勃地赏完了所有娇媚的花朵, 已是华灯初上, 只觉不虚此行, 最后恋恋不舍的离开了植物园。³

“In March of Spring, in these bright spring days, it was a good time to go sightseeing and admire the beauty of flowers. It was just the time that the botanical garden was holding a flower exhibition. It had different color of flowers and the spring was evident everywhere. The fragrant jasmines were swaying, like welcoming guests; the plums were like a beauty with flesh of ice and bones of jade, weathered so many storms, but still stand proudly; the begonia flowers were in bud, were very shy, and did not want people to see their true faces; how can (a spring) lack the "king of flowers" —— peonies which were elegant and poised, richly dignified and aromatic. After (I) finished the tour with full of enthusiasm, the colorful lights were lit. It was a worthwhile trip; finally (I) left the botanical garden reluctantly.”³
This short paragraph used many idioms to introduce a trip to a botanical garden. There are different ways of connecting them in a network. For example, according to the objects they describe, there are (i) idioms that describe flowers: 姹紫嫣红 chàzǐyānhóng “brilliant purples and reds”, 摇曳生姿 yáoyè shēngzī “swaying”, 傲然挺立 àorántǐnglì “stand proudly”, 含苞待放 hànbāo dài fàng “in bud”. 冰肌玉骨 bīngjī yùgǔ “flesh of ice and bones of jade”, 雍容华贵 yōngróng huáguì “elegant and poised”; (ii) idioms that describe the feelings of a person: 兴致勃勃 xìngzhìbóbó “full of enthusiasm”, 恋恋不舍 liànlèibùshě “have great attachment for ... and unable to part from ...”; (iii) idioms that describe the feelings of a person: 春光明媚 chūnguāngmíngmèi “bright spring days”, 春意盎然 chūnyì àngrán “the spring was evident everywhere”. According to the figure of speech used, there are idioms that are used anthropomorphically, such as 冰肌玉骨 bīngjī yùgǔ “flesh of ice and bones of jade”, 雍容华贵 yōngróng huáguì “elegant and poised”, and idioms that have metonymy, such as 历经风霜 lìjīng fēngshuāng “weather many storms”, 华灯初上 huádēngchūshàng “the colorful lights are lit”. 风霜 fēngshuāng “wind and frost” can refer to “hardships of a journey or of one's life”. 华灯初上 huádēngchūshàng “the colorful lights are lit” refers to the drawing of the evening.

In our learning App, after the learner has read or listened to a small story of idioms and learnt a few new idioms (or remind them about the idioms they have known before) within the related semantic network, they can play some games as well. For example, (i) select the idioms that describe flowers; (ii) select the idioms that can describe a person; (iii) select idioms that have a different meaning from the surface. Pictures can also pop up to remind students.

Using these approaches, students are not learning one isolated idiom, they can establish a semantic network, remember this idiom and other related idioms by using a story/picture/audio. This meaningful way of learning helps them master idioms more efficiently.

In addition to the above mentioned functions, the App also has bilingual translation, which enables most foreign learners to easily understand the text. No matter how much Chinese they have already learnt, anyone who is interested in idioms can use it.

4.2 Multimedia Functions
Different from the dull exhibition of PDs, EDs can have vivid multimedia functions. Since it combines Chinese traditional knowledge with modern technology, our App is also designed to provide students with most vivid and effective learning environment. The App applies sound and image to stimulate the auditory and visual sense of students. The vast majority of idioms have historical allusions, so when showing them, the setting options can use comics (conditions permitting, in the form of animation) and true voice explanation in order to reach a very good learning and the memory effect. In addition, for most foreign learners, the main purpose of learning Chinese is for oral communication. Therefore when they are learning idioms, the App has the setting option of pronunciation, which allows students to practice pronunciation through following and imitating the reading.

4.3 Interactive Learning Environment
The learning environment is designed to allow learners set speed according to their own arrangement, do peer assessment, and play online games. Such interactive settings make learners self-motivated and have fun in learning.

4.3.1 The Self-testing System
First, the self-testing system allows users to set self-learning speed. Students can learn according to their own schedule and arrangement. Therefore they will not have the problem of not being able to catch up. Second, new idiom learning can be divided into units. There’s a self-assessment system after each unit. The methods of assessments include: (a) choose the correct character of an idiom by listening; (b) select a proper sense of an idiom through various hints; (c) read the allusion and then choose the right idiom; (d) choose the right idiom to fill in the blank and complete
the sentence. Third, the system is designed to remember every error rate after a unit of evaluation, automatically classify the wrong words, and remind students to review them. According to the theory of Ebbinghaus Memory Curve, repeated learning helps students remember idioms from short-term memory to long-term memory.

4.3.2 Peer Assessment
(i) Online Play with Peers
With the popularity of mobile devices, students can now learn anytime and anywhere by using an electronic device. Besides, since the internet has been ubiquitous, students can not only study on their own, but also study together through the internet with other “players” who are using this App at the same time.
(ii) Understand or Collect Idiom Photos
On the one hand, the App can provide photos for students to guess which idioms fit them. On the other, “players” can take photos from real life situations that are related to the learned idiom and then upload them. This is a very effective way to check whether the students have grasped the meaning and the use of the idioms. For example, give students this sentence, ask them to find out which is an idiom, and take a photo in a situation that meets its meaning.

常常是演出还没开始，广场上就已人山人海。(CCL)
Chángcháng shì yǎnchū hái méi kāishǐ, guǎngchǎng shàng jiù yǐ rénshānrénhǎi.
“It is often the case that the show has not started, (but) the square has been packed.”

In addition, students can share their interesting experience in life through making sentences with their learned idioms. For example, one student may create this sentence:

如今，一些大型商场还在风风火火地建，不能不叫人担心。(CCL)
Rújīn, yīxiē dàxíng shāngchǎng hái zài fēngfēnghuǒhuō de jiàn, bùnéng bù jiào rén dānxīn.
“Now, some shopping malls are still rushed to be built; people cannot help but worry.”

Other learners can then leave their comments under this sentence and communicate with each other. Such design means that learners can chat and interact with each other. This is a way that largely stimulate students’ motivation for using the idioms.

4.3.3 Online Games
The setting of games is also very important. In our App, the idiom learning games can be arranged after each unit for the purpose of reviewing and consolidating what students have learned. After several units, it can have a review section. In addition, learners can still use the double or multiplayer mode, so that they can compete with each other to promote learning.

5. Conclusions

The Chinese idiom learning App in this study was designed to be based on both the features of Chinese idiom and the drawbacks of existing resources. The study focuses on three issues: whether there is enough useful idiom information, whether there are multimedia functions, and whether there is interactive learning environment.

Regarding useful information of idioms, this App aims to enable learners (i) to grasp the senses through different relations and choose the most suitable one for them (historical development, logical relation, frequency), (ii) to learn the most common grammatical function of an idiom first, and (iii) to establish a semantic network through learning related idioms in a meaningful paragraph. All these will have Chinese-English translations. Regarding multimedia functions, this App enables audio-video functions to enrich learning experience. Regarding the interactive learning environment, it is built in by self-testing system, peer assessment function and online games, which combine the new media and idiom study together. The study of these new functions makes this idiom-learning App a use-friendly, interesting and effective one, and can help a learner of Chinese as a second language grasp idioms more easily.

* Corresponding author.
Acknowledgement: This research is supported by the start-up research grant of The Hong Kong Institute of Education. The reference number is RG 71/2014-2015R.

References


An Electronic English Frequency Dictionary of Age Groups

LI Hanhong John  
The Open University of Hong Kong, Hong Kong  
typetoken@gmail.com

Abstract
Age has been an important factor in linguistic studies, especially in second language acquisition, psycholinguistics and sociolinguistics. Previous researches show a pattern of the relative frequency of using vernacular or nonstandard linguistic forms in different age groups: the frequency of vernacular forms are high in childhood and adolescence, steadily reduced in middle age, and gradually increase again in old age (Holmes 2013, Downes 1998). This indicates that age plays an important role in choosing different styles of words. For English learners of foreign or second language, it is vital to know a word’s stylistic features and usage. In order to capture how words are used in different age groups of native speakers, the current research makes full use of the age group information in the corpora of BNC XML (2007), CHILDES (2003), LCCPW(1996), POW (1984) to build up a balanced Age Group Corpus of 6-million-word tokens. This corpus is composed of six sub-sections, each representing one of the six age groups: Ag0 (aged 0-14), Ag1 (aged 15-24), Ag2 (aged 25-34), Ag3 (aged 35-44), Ag4 (aged 45-59), and Ag5 (aged 60+). For each word type in the corpus, the following information will be computed: its distributed frequency in different age groups, its total frequency in the whole corpus, its range (how many sub-sections of a corpus a word appears in), and its dispersion index (D-Value) and usage efficiency (U-value) (Carroll 1970). All of this information will be displayed in the shape of bar charts for English learners to explore each word’s stylistic usage in different age groups. Such a dictionary will not only provide one more approach to distinguish English synonyms but also examples for studying the relation between age and language in sociolinguistics.

Keywords: frequency dictionary, distributed frequency, age group, corpus and style

Introduction
With the development of computational lexicography, corpora have become an indispensable source in building a variety of dictionaries especially English learner’s dictionaries and frequency dictionaries. Word frequency information has been used to select headwords and defining vocabulary, order senses, and decide collocation, core vocabulary and language style. The richer metalinguistic information (i.e. text types, genres, author and speaker’ age, gender, social background) is annotated in the header of a corpus file, the higher possibility we can research into the relation between word frequency and these metalinguistic factors.

Distributed frequency, which computationally measures how a word is dispersed in different sections of a corpus, contributes to a deeper understanding of the lexical property of a word (Li 2013). Distributed frequency and dispersion index of a word in different sections of a corpus have been used to build frequency dictionaries (Davies and Gardner 2010). Distributed frequency combined with other factors such as text genres and writers and speakers’ age have been used to select better core vocabulary for learners (Li 2010, Li 2011, Li & Fang 2011). Furthermore, distributed frequency of a word in different genres can help to distinguish the differences between synonyms in terms of style (Li 2010).

In addition, age has been an important factor in linguistic studies, especially in second language acquisition, psycholinguistics and sociolinguistics. Previous researches show a pattern of the relative frequency of using vernacular or nonstandard linguistic forms in different age groups: the frequency of vernacular forms are high in childhood and adolescence, steadily reduced in middle age, and gradually increase again in old age (Holmes 2013, Downes 1998).
Making use of the previous study of distributed frequency, the current study attempts to combine age information tagged in corpora and a word’s frequency in different age groups so as to build an *Electronic Frequency Dictionary of Age Groups*. With this new dictionary, we offer a database for English learners to explore how a word is used in different age groups in terms of stylistics and sociolinguistics.

**Age and vocabulary style**

Will speakers of different age-grading show preference to different styles of words? Downes (1998) found a relationship between the use of vernacular forms and age as displayed in the following figure.

![Figure 1 Relationship between use of vernacular forms and age](source)


The figure suggests the relative frequency of vernacular forms used in different age groups (Holmes 2013). People in childhood and adolescence tend to use more vernacular forms of language. When they are approaching middle age, the use of vernacular forms drops “when societal pressure to conform is greatest” (Holmes 2013: 177). “Vernacular usage gradually increases again in old age as social pressures reduce, with people moving out of the workforce and into a more relaxed phase of their lives” (177).

This sociolinguistic model indicates that speakers of different age groups prefer different linguistic forms including different styles of words.

**Aims**

Inspired by the above sociolinguistic theory of language style and age, this research aims to build an *Electronic English Frequency Dictionary of Age Groups* with the help of modern English corpora. English learners can use this dictionary to explore how a word is frequently used by different age groups. In this case, we can study how a word is preferred by different age groups, and how the factor of age influences the usage of a word. Such a dictionary will not only provide one more approach to distinguish English synonyms but also examples for studying the relation between age and language in sociolinguistics.

**Construction of Age Group Corpus**

In order to build the *Frequency Dictionary of Age Groups*, we need to build a corpus according to the factor of age. According to the annotation of authors’ and speakers’ age
information in the header of corpora, we tried to select corpus files and build a balanced Age Group Corpus.

**BNC XML and Age Group**

“The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written” (Burnard 2009). Since BNC headers offer the tagging of age group information for each writer and speaker, it offers an opportunity to sum up the vocabulary usage according to age groups.

The vocabulary summary according to age groups in BNC XML (2007) is as follows:

**Table 1** Word tokens of each age group in the whole BNC XML (Li & Fang 2011: 163)

<table>
<thead>
<tr>
<th>age group</th>
<th>age range</th>
<th>spoken tokens</th>
<th>written tokens</th>
<th>total tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ag0</td>
<td>0-14</td>
<td>385,234</td>
<td>59,559</td>
<td>444,793</td>
</tr>
<tr>
<td>Ag1</td>
<td>15-24</td>
<td>594,400</td>
<td>542,578</td>
<td>1,136,978</td>
</tr>
<tr>
<td>Ag2</td>
<td>25-34</td>
<td>1,120,516</td>
<td>2,267,123</td>
<td>3,387,639</td>
</tr>
<tr>
<td>Ag3</td>
<td>35-44</td>
<td>1,075,749</td>
<td>6,726,931</td>
<td>7,802,680</td>
</tr>
<tr>
<td>Ag4</td>
<td>45-59</td>
<td>1,638,364</td>
<td>7,230,715</td>
<td>8,869,079</td>
</tr>
<tr>
<td>Ag5</td>
<td>60+</td>
<td>1,137,433</td>
<td>5,156,077</td>
<td>6,293,510</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>5,951,696</td>
<td>21,982,983</td>
<td>27,934,679</td>
</tr>
</tbody>
</table>

Based on the above information, a program was compiled to select corpus files randomly from different age groups of BNC XML. For each age group, it aims to collect written texts of half a million tokens and spoken texts of half a million tokens. The result is as follows:

**Table 2** Word tokens sampled from each age group of the BNC XML (Li & Fang 2011: 163)

<table>
<thead>
<tr>
<th>age group</th>
<th>age range</th>
<th>sampled spoken tokens</th>
<th>sampled written tokens</th>
<th>total tokens</th>
<th>total type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ag0</td>
<td>0-14</td>
<td>302,618</td>
<td>59,559</td>
<td>362,177</td>
<td>16,080</td>
</tr>
<tr>
<td>Ag1</td>
<td>15-24</td>
<td>500,397</td>
<td>500,868</td>
<td>1,001,265</td>
<td>24,932</td>
</tr>
<tr>
<td>Ag2</td>
<td>25-34</td>
<td>510,863</td>
<td>513,666</td>
<td>1,024,529</td>
<td>25,219</td>
</tr>
<tr>
<td>Ag3</td>
<td>35-44</td>
<td>525,850</td>
<td>501,106</td>
<td>1,026,956</td>
<td>26,572</td>
</tr>
<tr>
<td>Ag4</td>
<td>45-59</td>
<td>501,575</td>
<td>506,572</td>
<td>1,008,147</td>
<td>26,582</td>
</tr>
<tr>
<td>Ag5</td>
<td>60+</td>
<td>505,803</td>
<td>507,623</td>
<td>1,013,426</td>
<td>24,796</td>
</tr>
</tbody>
</table>

In this step, we have built up subcorpora of The Age Group Corpus for the following age groups in BNC XML: Ag1 (aged 15-24), Ag2 (aged 25-34), Ag3 (aged 35-44), Ag4 (aged 45-59), Ag5 (aged 60+). Each age group contains 1 million tokens, half of which are spoken tokens and the other half written tokens.

However, for Ag0 (aged 0-14), BNC XML does not have enough data to build up a 1-million-token sub-corpus. Therefore, extra corpus files are needed.

**Supplementary Corpora for Ag0**

For the Ag0 in BNC XML, we combine it with other corpus data so as to collect texts of 1 million words for Ag0 (aged 0-14). We include all words from the Polytechnic of Wales Corpus (POW⁴, Souter 1989). In addition, we include words from the following three components of the British English Corpora in the Child Language Data Exchange System

---

⁴ The Edited POW can be downloaded here after application [http://ota.ox.ac.uk/desc/1514](http://ota.ox.ac.uk/desc/1514)

The composition of the first age group (Ag0) can be summed up as follows:

Table 3 Composition of the first age group (Ag0) in the Age Group Corpus (Li & Fang 2011: 164)

<table>
<thead>
<tr>
<th>component</th>
<th>tokens</th>
<th>mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ag0 in Spoken BNC</td>
<td>302,618</td>
<td>Spoken</td>
</tr>
<tr>
<td>The Polytechnic of Wales Corpus (POW)</td>
<td>60,717</td>
<td></td>
</tr>
<tr>
<td>The Howe, Manchester, and Wells sub-corpora from the Child Language Data Exchange System (CHILDES)</td>
<td>530,416</td>
<td></td>
</tr>
<tr>
<td>Ag0 in Written BNC</td>
<td>59,559</td>
<td>Written</td>
</tr>
<tr>
<td>The Child Writing sub-corpus of the LUCY Corpus</td>
<td>25,995</td>
<td></td>
</tr>
<tr>
<td>The Lancaster Corpus of Children’s Project Writing (LCCPW)</td>
<td>82,396</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>1,061,701</td>
<td></td>
</tr>
</tbody>
</table>

Age Group Corpus

With BNCXML and the supplementary corpora, we finally collect all the needed corpus data for constructing a balanced Age Group Corpus. The composition of this corpus can be summed up as follows:

Table 4 Composition of the Age-Group Corpus (Li & Fang 2011: 165)

<table>
<thead>
<tr>
<th>age group</th>
<th>age range</th>
<th>spoken tokens</th>
<th>written tokens</th>
<th>total tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ag0</td>
<td>0-14</td>
<td>893,751</td>
<td>167,950</td>
<td>1,061,701</td>
</tr>
<tr>
<td>Ag1</td>
<td>15-24</td>
<td>500,397</td>
<td>500,868</td>
<td>1,001,265</td>
</tr>
<tr>
<td>Ag2</td>
<td>25-34</td>
<td>510,863</td>
<td>513,666</td>
<td>1,024,529</td>
</tr>
<tr>
<td>Ag3</td>
<td>35-44</td>
<td>525,850</td>
<td>501,106</td>
<td>1,026,956</td>
</tr>
<tr>
<td>Ag4</td>
<td>45-59</td>
<td>501,575</td>
<td>506,572</td>
<td>1,008,147</td>
</tr>
<tr>
<td>Ag5</td>
<td>60+</td>
<td>505,803</td>
<td>507,623</td>
<td>1,013,426</td>
</tr>
</tbody>
</table>

In this corpus, there are 1 million tokens for each age group. Basically, it is a balanced corpus of spoken and written language, containing around 6 million words.

Data processing for the Frequency Dictionary of Age Groups

For the established Age Group Corpus, we compute each word type’s frequency in each age group, its range, total frequency, coverage, dispersion index (D-Value) and usage efficiency (U-value) (Carroll 1970).

Figure 2 Computation of the Frequency Dictionary of Age Groups

Hw=headword
frqAg0=frequency in Age Group (0-14)
frqAg1=frequency in Age Group (15-24)
frqAg2=frequency in Age Group (25-34)
frqAg3=frequency in Age Group (35-44)
frqAg4=frequency in Age Group (45-59)
frqAg5=frequency in Age Group (60+)
FrqSum=total frequency
C=coverage
CC=cumulative coverage
D = Dispersion Index by John Carroll (1970)
U = Usage Efficiency by John Carroll (1970)

Dispersion index (D) can indicate how widely a word is distributed. If a word occurs in each age group, D will be 1. If a word occurs in only one age group, D will be 0. Carroll’s Usage Efficiency (U) is another parameter to measure how widely a word is distributed in different age groups in our research. It not only considers a word’s total frequency in a corpus but also its frequency distributed in different corpus sections (Carroll 1970).

The formula of calculating Carroll’s D and U is as follows: (Carroll 1970: 63)

Let $n$ = the number of categories

\[
f_j = \text{the frequency (number of occurrences) of a given word-type in category } j \quad (j = 1, 2, ..., n)
\]

\[
F = \text{the total frequency (number of occurrences) of the word in the corpus} = \sum_j f_j
\]

\[
s_j = \text{the number of tokens in category } j
\]

\[
N = \text{the total number of tokens in the corpus} = \sum_j s_j
\]

\[
p_j = \text{the proportion of tokens in category } j \text{ that are instances of the given word-type. Thus, } p_j = f_j / s_j
\]

\[
P = \sum_j p_j \quad (P \text{ may take any positive value; it is not in general equal to unity})
\]

Then:

\[
H = \log P - \frac{(\sum_j p_j \log p_j)}{P} \quad (p_j \log p_j = 0 \text{ for } p_j = 0)
\]

\[
D_2 = H / \log n = \text{the index of dispersion}
\]

\[
f_{min} = (\sum_j s_j f_j) / N
\]

\[
U_m = (1,000,000 / N) [FD_2^+ (1-D) f_{min}]
\]

Data visualization and electronic frequency dictionary of age groups

In order to build the Electronic English Frequency Dictionary of Age Groups\(^5\), the computed data in Figure 2 will be visualized into different bar charts and embedded with a search function via Tableau software\(^6\). For example, Figure 2 shows how "funny" appears in different age groups. According to the search result of the dictionary, R (Range) is 6 and D (dispersion index) is 0.96, which indicate that "funny" is a word widely used by different age groups. According to the frequency distributed in each group, "funny" tends to be used more frequently among young people. Its frequency continues to drop when speakers grow older. This word is

\(^5\)This Electronic English Frequency Dictionary of Age Groups is available online
https://sites.google.com/site/typetoken/home/english-frequency-dictionary

\(^6\)Tableau software http://get.tableau.com/trial/
most frequently used by people of Ag0 (aged 0-14). It might tell us that children and teenagers tend to be happier and have more funny things.

Figure 3. Electronic Frequency Dictionary of Age Groups

Application of the electronic Frequency Dictionary of Age Groups

In order to test the newly built Electronic Frequency Dictionary of Age Groups, we apply it to the following aspects: 1) to identify typical words in certain age group, 2) to distinguish synonyms, 3) to find sociolinguistic patterns.

Typical words in certain age group
We search piggy in the dictionary and its distributed frequency in different age groups is displayed in Figure 4. Judging from the search result, it is obvious that piggy is most frequently used in age group Ag0 (aged 0-14). This indicates that piggy is a typical word used among children.
Differences between synonyms

We try to explore whether the dictionary can help learners to distinguish synonyms. We test the dictionary by searching get and obtain. The results are displayed in Figure 5 and Figure 6. Figure 5 indicates that obtain, as a formal word, is used less frequently in Ag0 (aged 0-14), Ag1 (aged 15-24) and Ag2 (aged 25-34) while it is most frequently used in Ag3 (35-44), Ag4 (aged 45-59) and Ag5 (aged 60+). It matches the sociolinguistic pattern discussed in section 2. Speakers at middle age tend to use more formal form. Obtain is more formal than go. Moreover, it is acquired at an elder age in the language development.
Figure 5 Distributed frequency of obtain
Figure 6 shows that *get* is a common word, frequently used by different age groups.

Sociolinguistic pattern for vocabulary use

As pointed out in section 2, young speakers tend to use more vernacular or non-standard form of language, and as they grow older they tend to use more standard form due to social pressure. We try to search swearing words such as *shit* and *damn* in the *Electronic Frequency Dictionary of Age Groups* with result displayed in Figure 7 and Figure 8. It turns out that these informal and swearing words are much more frequently used by youngsters in Ag0 (0-14), Ag1 (15-24) and Ag2 (25-34) than the middle aged people, Ag3 (35-44) and Ag4 (45-59). Middle-aged people tend to use more standard form or formal form of language. One reason is that they need to conform to social pressure. When employed, they pay more attention to their language style and try to conform to social norms. If they are parents of children, they pay attention to their language style and using swearing words will be bad examples for their children. *The Electronic Frequency Dictionary of Age Groups* thus helps us to understand better how a word is used in different age groups due to sociolinguistic factors.
Figure 7 Distributed Frequency of *shit*
Conclusion

The Electronic Frequency Dictionary of Age Groups makes full of modern corpora, computational method and sociolinguistic theories. It can show how a word’s frequency is distributed in different age groups. It offers a possibility for a learner to identify typical words used by certain age group, and contrast the differences of synonyms by observing how they are used among different age groups. Moreover, it can be used as a database for sociolinguistic studies to find out the relation between a word usage and social factors.

Due to the limitation of only one-million tokens in each age group, the frequency of some words is low in each age group. With the development to modern corpora and more annotation of age information in headers, it is hoped that a better version will be available.

References


LINE Dictionary: More than a dictionary

Tae-hoon Lee and Eun-joo Jeon
NAVER Co. Ltd, South Korea
paper.smell@navercorp.com and eunjoo.jeon@navercorp.com

Abstract
Due to the technical development of Internet and mobile phones, foreign learners, interpreters, and translators are able to use dictionaries with their PCs and smart phones easily. Since 1999, NAVER launched NAVER Dictionary for Korean users, now servicing dictionaries of 31 languages and 11 million headwords through PC Web, Mobile Web, Mobile APP, etc. currently in March, 2015. With various contents and experiences accumulated for 15 years, NAVER launched LINE Dictionary APP in 2014 for Chinese, Thai, and Indonesian users by cooperating with LINE. LINE Dictionary not only provides reliable dictionary contents with the most advantages of online dictionary, but also native-spoken word audio, self-developed TTS (Text to Speech), Translator, and word-learning program Words Up to make users easily learn languages. Beyond simple dictionary search service, LINE Dictionary presents comprehensive service model that can satisfy language learners, interpreters, and translators by sending a variety of helpful daily contents such as Today's Expression, Quotes/Proverbs, and Grammars through LINE messenger for language study.

Keywords: LINE DICTIONARY, translator, words up, mobile app, language learning

1. Naver Dictionary

Since 1999, NAVER has provided the service of NAVER Dictionary through PC Web, Mobile Web, Apps. As of May 1, 2015, it is servicing 11 million headwords of 31 languages currently, and for the month of April 2015, it reached 11 million monthly UB (Unique browsers) and 300 million page views. Almost 50% of them were English-Korean Dictionary. In addition, Multilingual Translator/Interpreter, Multilingual phrasebook, user-participating translation, Word Study and Quiz services, etc. have also been getting positive reviews from the users.

It has been decided to develop and manage dictionary service targeting global users with the accumulated contents and experience earned from NAVER Dictionary as a base. Under the globally familiar brand name of LINE Dictionary, it released English-Indonesian dictionary service in June, 2014, and English-Thai dictionary service in October, 2014. Then, it combined nciku Chinese-English dictionary that was running from 2007 with LINE Dictionary.
Table 1. Naver Dictionaries

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Headwords</th>
<th>Dictionary</th>
<th>Headwords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Korean</td>
<td>522,964</td>
<td>Italian</td>
<td>85,504</td>
</tr>
<tr>
<td>English</td>
<td>5,134,445</td>
<td>Portuguese</td>
<td>170,210</td>
</tr>
<tr>
<td>Hanzi</td>
<td>398,533</td>
<td>Georgian</td>
<td>15,848</td>
</tr>
<tr>
<td>Japanese</td>
<td>582,730</td>
<td>Ukrainian</td>
<td>16,681</td>
</tr>
<tr>
<td>Chinese</td>
<td>1,854,884</td>
<td>Rumanian</td>
<td>34,802</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>112,819</td>
<td>Swedish</td>
<td>71,709</td>
</tr>
<tr>
<td>Indonesian</td>
<td>73,194</td>
<td>Polish</td>
<td>33,070</td>
</tr>
<tr>
<td>Thai</td>
<td>90,573</td>
<td>German</td>
<td>203,643</td>
</tr>
<tr>
<td>Arabic</td>
<td>21,991</td>
<td>Russian</td>
<td>414,683</td>
</tr>
<tr>
<td>Cambodian</td>
<td>9,339</td>
<td>Latin</td>
<td>42,899</td>
</tr>
<tr>
<td>Mongolian</td>
<td>16,893</td>
<td>Turkish</td>
<td>13,420</td>
</tr>
<tr>
<td>Hindi</td>
<td>40,801</td>
<td>Albanian</td>
<td>12,231</td>
</tr>
<tr>
<td>Persian</td>
<td>31,699</td>
<td>Uzbek</td>
<td>36,468</td>
</tr>
<tr>
<td>Swahili</td>
<td>15,470</td>
<td>Dutch</td>
<td>121,393</td>
</tr>
<tr>
<td>French</td>
<td>1,624,388</td>
<td>Hungarian</td>
<td>57,922</td>
</tr>
<tr>
<td>Spanish</td>
<td>429,943</td>
<td>Czech</td>
<td>26,433</td>
</tr>
</tbody>
</table>

2. LINE Dictionary

2.1 About LINE
LINE has been serviced worldwide with the mobile messenger app, and it is deriving a lot of interests from Japan, Thailand, Taiwan, Indonesia and other Southeast Asian countries especially. Apart from being a messenger app, LINE has been operating various Family Apps that are useful to the mobile users such as LINE Dictionary, LINE TV, LINE Webtoon, LINE SHOP, LINE Antivirus, LINE Camera, etc.
Requests from Southeast Asian users for the Dictionary App

In order to plan the features and build contents for the LINE Dictionary, in-depth interviews with 14 people were done in Hanoi, Vietnam, Bangkok, Thailand, and Jakarta, Indonesia from January 1st to 18th, 2014. We selected the interviewees who were university students/postgraduate students/office workers learning English, using the Dictionary App, studying English more than three hours a week and using Dictionary App more than three times a week. We asked interviewees to sort 41 cards related to Dictionary App features in order of importance and then recorded their opinion regarding usage of Dictionary application in detail while showing them the prototypes of LINE Dictionary. The common important factors that interviewees picked follows:

Contents
Number of headwords
Example sentences and translations of various fields
Learning information such as phonetic alphabets (International Phonetic Alphabet, IPA), conjugated forms, thesaurus, etc.
Learning contents such as expressions, wise saying, etc.

Features
Translator
Offline app support
Auto-completion of searched words
Listening to the native audio pronunciation
Enabling to search the conjugated forms support
Learning function support such as memorizing words, sentences, etc.

In particular, for the functions, there were lots of requests regarding offline app support due to poor Internet and mobile infrastructure environment and also there were high demand for willingness towards combining translator and Dictionary Apps in one.

In addition, overwhelming demand for offline support and translator was observed through the results of online surveys conducted with 4,500 LINE Dictionary Indonesian users on September 16, 2014.

![Pie chart showing dictionary preferences](image)

Figure 2 LINE Dictionary online survey. 2015.
Therefore, we have planned LINE Dictionary to solve those users’ requests of using dictionary search, translator, learn foreign language at once and has been getting positive reviews from the users.

2.3 LINE Dictionary Contents
LINE Dictionary has been providing English-Indonesian, English-Thai, and Chinese-English dictionary services from June 2014. It started with the mobile app service but also released PC Web, Mobile Web versions to use seamlessly in a variety of environments.

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Headwords</th>
<th>Example Sentences</th>
<th>Audio Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-Chinese</td>
<td>1,479,194</td>
<td>7,191,803</td>
<td>65,305</td>
</tr>
<tr>
<td>Chinese-English</td>
<td>1,364,251</td>
<td></td>
<td>240,416</td>
</tr>
<tr>
<td>English-Indonesian</td>
<td>159,935</td>
<td>7,248,847</td>
<td>58,693</td>
</tr>
<tr>
<td>Indonesian-English</td>
<td>55,80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English-Thai</td>
<td>142,673</td>
<td>7,299,329</td>
<td>40,447</td>
</tr>
<tr>
<td>Thai-English</td>
<td>51,132</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It had also been providing dictionaries such as SE-ED and Oxford with high reliability and allows users to search a simple meaning on the offline status. It has been planned to get the headwords from research center and publishing company which have public confidence, and latest Internet open dictionary continuously and update.

Apart from main dictionary, it provides various contents that are helpful in learning foreign languages such as today’s expression, wise saying/proverb, quiz, etc. and they are getting positive reviews from the users. Also, listening to the American/British English pronunciation of native speakers, and listening to the standard Chinese pronunciation are essential contents using dictionary.
Figure 4 Variety of learning contents such as Today’s expression, wise saying, quiz, etc.

LINE Dictionary Features

2.4.1. Auto-Completion (Supports offline mode)
It supports the auto-completion feature that recommends relevant search results if users only type the part of the word to search. In addition, user can check out the main meaning of the word and the phonetic information on the auto-completion result page and it also operated on the offline mode so users can use it more conveniently.

2.4.2 Phonetic Search
User can search the word by phonetic search even if users do not know the exact spelling of the English words.

2.4.3 Pop-up Dictionary
If users touch the word that used on the example sentence, user can see the meaning of the word from the pop-up dictionary so users do not have to jump to another screen.

2.4.4 Example Sentence Search and User Participated Translation
It provides the example sentence search so it allows users to see the various examples of specific word and can get the help for the writings. Also users can participate in translation and upload the translated sentences if there are no translated ones.
2.4.5 Translator
It includes self-developed translator so users can use both dictionary search and translate the sentence together. In order to improve the translation quality, we are continuously looking for the good example sentence source and improving the translator by reviewing the users’ translated results every day.

2.4.6 Words Up Program for the Word Study
Users can seize on the words searched from LINE Dictionary through the Words Up function. All the words that users searched are stored in Words Up, and if the users run the Words Up, the variety of programs like revise, arrange words in order, selecting best word, etc. will be available to memorize and understand the words best. It also sends push-alarm periodically to encourage users to study words continuously.

Words Up allows users to study words with the words that LINE Dictionary prepared apart from the words that users searched and it is planned to expand as a better program through the partnership with other publishing companies later.

3. Conclusion
We have created LINE Dictionary in a situation of wondering how to make everyone satisfied as the dictionary users have been gradually decreasing and translator users have been getting increased. Finally, we have developed LINE Dictionary that allows users to have dictionary,
translator, and study functions in one app through the results derived from user interviews and surveys, and it has been getting positive reviews from the users.

Obviously the functions and contents as a Dictionary App basically will be improved and we will try to propose the model of digital dictionary in the upcoming future.

References


*Kamus Inggris (Kamusku)*: Available at https://play.google.com/store/apps/details?id=com.dridev.kamusku

LEXiTRON Thai-English Electronic Dictionary: Available at http://lexitron.nectec.or.th/2009_1

LINE Dictionary Introduction: Available at http://linedict.naver.com/about

LONGDO Dict: Available at http://dict.longdo.com/

NAVER Dictionary: Available at http://dic.naver.com/
Spelling Variation Problem in Persian:

Automatic Development of a Spelling Database for a Digital Dictionary

Saghar Sharifi
Karaj Branch, Islamic Azad University, Karaj, Iran
saghar.sharifi@kiau.ac.ir

Marziye Sanaati
Research Center of Cultural Heritage and Tourism Organization, Tehran, Iran
msanaati20@gmail.com

Masood Ghayoomi
Freie Universität Berlin, Berlin, Germany
masood.ghayoomi@fu-berlin.de

Abstract:
In developing digital dictionaries, data-driven approaches are given a particular attention whose advantages being their easy updating and reusability for other advanced language processing applications. In corpus linguistics, when there are different spellings of words in the texts, the task would be difficult, because the search result of one particular word in the corpus without taking its spelling variations into consideration might be incomplete. This problem is severe in Persian and is partly rooted in the nature of the Persian alphabet in terms of joined-up writing and space application, the Hamzeh character and its base, among others, and partly due to its phonological and morphological properties, all of which leading to the appearance of ambiguous forms.

This paper aims at proposing a data-driven model to develop a database that contains the spelling variations of a word in the corpus along with their relevant frequencies to represent the degree of the agreement among the native speakers to spell a word. In our proposed method, we conduct a weighted algorithm using the Levenstein distance to find various spellings of a word in the Persian corpus via assigning low weight to particular characters that are substituted or inserted. The prepared data is manually error-analyzed to find the ambiguous cases. Towards the goal, we also propose to classify the spelling variations into orthographic and morpho-phonological classes.

Keywords: the Persian Language, Spelling Variation, Corpus Linguistics, Classification, Spelling Database, the Levenstein distance

1. Introduction

Writing is a medium through which language is represented via using graphic signs and symbols. It is a complement to speech and is a system that should be acquired. Although Persian is an Indo-European Language, it uses the Arabic Alphabet, also called Perso-Arabic, which is a modified version of Arabic script.

The Persian alphabet, has four letters more than Arabic, namely "ژ/ž", "پ/p," "چ/č," and "گ/g". Moreover, some letters in the Persian Alphabet have a different sound from their Arabic counterpart. For instance, in Arabic, "ط/t" represents the pharyngealized dental stop, while in Persian, it represents the same phoneme as "ط/t", i.e. the dental stop.

The specific properties of the Persian script bring about difficulties in writing Persian in such as way that Moshiri (1997: 9) argues that there is hardly anyone who has not been in trouble with the Persian spelling. The inconsistency in Persian writing is a major factor giving rise to various spelling forms for words. This is found in written texts, both formal and informal,
such as novels, textbooks, business correspondence, the press and in short text messages as well. Researchers and those having a hand in writing Persian have much discussed the problem and referred to the points which are mainly orthographic. In order to solve some of these problems in the field of studying the Persian language, we need to have access to a corpus which is not biased. Creating such a corpus gathered from different linguistic sources, we should meet the challenges: determining word-internal boundary, different writing styles, use of some specific Arabic letters and Latin alphabet in Persian texts, to name a few (Ghayoomi et al 2010).

In this paper, we investigate the automatic development of a spelling database as a dictionary from a new perspective. Then, we propose a fine-grained classification of the factors leading to spelling variation. Furthermore, we aim at organizing the data as a database to be analyzed and used in language teaching, corpus linguistics, and various computational linguistics applications.

This paper contains six sections. In Section 2, we refer to the previous studies in the field. Section 3 introduces the corpora, and the method used in our research is described in Section 4. In Section 5, the classification of the factors giving rise to the spelling variation is explained, and additionally, we discuss the analysis of the data. And finally, the paper ends with concluding remarks in Section 6.

2. Review of literature

Spelling variation in Persian has been much discussed in the literature. Moshiri (1997: 9) regards spelling problem a prevailing one even among the educated. She refers to invasion of foreign spelling, especially Arabic, into Persian and the existence of homophones in the Persian language, the words "قلم" /qāleb/ 'mould' and "غالب" /qāleb/ 'conqueror', as an example.

Nasri (1936), Ashuri (1987), Tayyeb (1993), AmirJanloo (2003), Eftekhari (2006) and Mortezayi (2002) have discussed the most important points about the spelling problem in Persian which are described as follows:

a) In Persian writing, short vowels (/a/, /e/ and /o/) and gemination (ّ) are not shown in the alphabet. This may allow different readings of a word. For example, the form "در" might be read either /dar/ 'door' or /dorr/ 'pearl'.

b) In some cases, one phoneme is represented with more than a grapheme and vice versa. For instance, /l/ is represented both with "ئ" and 'ئ' that represents both the consonant /l/ and the vowels /u/ and /o/.

c) Sometimes for the letter 'ئ', which represents the glottal stop word-initially, the tildes (ˆ) is not written on Alef (ا), consequently two word forms for one such as اب 'آبان' /ʔābān/ 'name of the 8th month'.

d) Hamzeh (ء), the letter signifying the glottal stop, may be written on various bases. This results in to have various word forms for one, such as 'مسؤول' /masʔuł/ 'responsible'.

e) Letters have different shapes with respect to the position they appear, at the beginning of a word, in the middle, or at the end. For example, /g/ is represented by "گ" word-initially and "گ" word-terminally, as in /گل/ 'flower' or /مرگ/ 'death'. There are also some letters that join their adjacent letters from both sides, while others join only from the right (See Section 5.1.1 for the examples).
f) In some words, there might be a letter which is not pronounced. In "خواستن"/xāstan/ 'want', for instance, the letter "و" is not pronounced.

g) At times, complex words are written continuously in a flowing line without following a logical rule. For instance, /mišavad/ 'becomes' can be written in three forms: joined-up as میشود, disjoined with a white space as می شود, and disjoined with pseudo-space, as می شود.

h) Some Arabic words and expressions have entered Persian texts with their own native spelling, such as حتی الامکان/hattalʔemkan/ 'as far as possible'.

i) In the Persian alphabet, some diacritics are used to make a distinction among some letters. This leads to the similarity of some of these letters, such as 'ز' /z/, 'ژ' /ž/ and 'ج' /j/, 'چ' /č/ and 'ح' /h/.

Sadeghi and Zandi Moghadam (2006) studied different styles in writing simple and derived words, such as 'جدای' /jodāʔi/ 'separation', and they refer to lack of uniformity in joined-up writing and space application in complex words as in 'تجارت' /tejārat-xāne/ 'commercial office' and 'چاپخانه' /čāpxāne/ 'printing house'. They address the problem to the difference between Arabic and Persian scripts and believe that the Academy of the Persian Language plays a leading role in unifying different writing styles (Sadeghi and Zandi Moghadam, 2006: 5-8). Besides the Academy of the Persian Language, other institutions, for instance Iran University Press, have much contributed the case through publishing style guides (Sami‘i Gilani, 1988).

In addition, several papers have been written on studying the challenges and introducing automatic methods to resolve the orthographical problems in a Persian corpus created from the Web, among which Ghayoomi et al (2010) could be mentioned.

As can be seen, much of what is reviewed focus on orthographic properties of the Persian writing system. In this study, we aim at presenting a new classification via investigating the factors causing the spelling variation. To this end, we use a data oriented approach that is discussed in the next section.

3. The corpora investigated

In a data driven approach, a corpus is required. To this end, we gather a collection of data from three corpora. The BijanKhan Corpus7 (BijanKhan 2004) contains 2.5 million word tokens and it is accessible online for free. Produced by the Research Center of Intelligent Signal Processing8, this corpus constitutes a subsidiary of a larger one with 100 million words. It is comprised of various formal written texts in which the words have been tagged both semantically and morpho-syntactically.

The second corpus is the Persian Language Dependency Treebank9 (Rasooli et al 2013) created by Dadegan Group10. It contains over 500,000 words and can be accessed online for free. The corpus is well-balanced with a good textual variety. In addition to POS tags, syntactic structure of the sentences has been given.

---

7 http://ece.ut.ac.ir/dbrg/bijankhan/
8 http://www.jsdp.reisp.ac.ir
9 http://dadegan.ir/catalog/perdt
10 http://www.dadegan.ir
Talkbank is another corpus with a volume of over 500 million words gathered from Persian websites and blogs. This data set is collected by the research group of Schlomo Argamon at the Illinois Institute of Technology. The advantage of this corpus comparing to the other corpora is that the data is user-generated and is the outcome of the interaction between the user and the Web. As a result, unlike the other two, it has the variety of register and style and is far from being standard.

4. Computational method used to find the spelling variations

As mentioned in Section 1, we aim to provide a framework for identifying spelling variation which is a big challenge in corpus creation and corpus linguistics in Persian. To do this, we propose a computational method and also a new classification. Ghayoomi et al (2015) have introduced an extended version of the Levenstein Distance algorithm (Levenstein 1966) to automatically extract those variations from the integrated corpora. Using the algorithm, one can measure the distance between two words in terms of 'substitution' and 'insertion'. By default, the weight for each distance is measured 1. In the extended algorithm by Ghayoomi et al (2015), weighting is changed to 0.1 for the distance between certain characters that cause spelling variation (See Section 5 for these variations), while in other cases, the weight remains 1. The change in weighting is done to make searching the data easier for creating a database of the spelling variations.

5. Spelling variation and data analysis

In this section, we study the proposed categories and show the factors giving rise to spelling variation in Persian. In the following of this section, we present the analysis statistically and discuss about it.

5.1 The Causes

Investigating the data from the integrated corpora revealed that spelling variation in Persian may be classified under 6 groups as follows:

5. 1. 1 Joined-up Writing and Space Application

Being an orthographic property of the Persian script, joined-up writing and space application, seems to be a major problem in spelling a word. Some Persian characters are joined to the adjacent letters from both sides "س" in "سر"/sar/ 'head' and in "ساس"/bas/ 'enough', while others are joined only from the right, as "ر" in "كر"/kar/ 'deaf'. This may lead to the emergence of various and sometimes obscure or even meaningless spelling forms. In all words (both simple and complex) with the letters from the latter group in the initial and middle position, a disjoined form may appear, such as in "؟اگر"/ʔagar/ 'if' and "ناد"/dād/ 'he/she gave'. Obviously, in the above examples, there is no need to apply space after Alef, the first and second letter respectively.

In complex words (including derivatives, compounds and derivative-compounds), in addition to the above-mentioned cases, disjoined forms emerge mostly at the boundary between morphemes, regardless of the fact that which group the final letter of the morpheme belongs to, as in Examples (1) and (2):

(1) می شد

mi šod

It should be mentioned that joined-up writing could be a matter of personal preference. Previously, it was a general trend.
In these words, the component parts, i.e. the morphemes, can be written in three different ways: joined, disjoined (with a space) or disjoined without a space (with pseudo-space), as is the case in 'میشود' and 'می شود', and 'می شود.' Inconsistency in choosing each of these forms could be attributed to the writer or the typist’s hastiness, carelessness or the absence of a writing style for applying space rules correctly and consistently. Besides, the choice may be influenced by some morpho-phonological factors, i.e. substitution and deletion (See Sections 5.1.4 and 5.1.5).

5.1.2 Absence of Tilde on Alef
There are examples in which the absence of tilde on Alef leads to the spelling variation, such as "ابن" instead of "ابن.

5.1.3 Hamzeh
Writing Hamzeh on various bases, such as "أ", "إ", "ؤ", "ئ", its replacement with "ی" (representing /y/) or its omission could lead to the emergence of different forms for one word, such as Examples (3) and (4):

(3) a) آبین
?āyin
tradition
b) آبین
inʔā
tradition

(4) a) مسؤول
masʔul
responsible
b) مسؤول
masʔul
responsible

5.1.4 Morpho-phonological Factors: Phonological Changes
At times, due to the application of some phonological processes, especially in informal speech, the spelling of a word changes. These changes occur rather on a regular basis in the form of substitution of a letter with another. Various forms resulting from phonological changes can be divided into the following sub-categories:

a) Word Form
This type of change is found word-internally in all word classes. Phonetic change often causes a change in style or register of a word mainly from standard to colloquial or slang and rarely to literary. For instance, "اگر" /ʔagar/ 'if' changes into "اگه" /ʔage/ and "راه" /rāh/ 'way' becomes "ر" /rah/.

---

13 Imperfect
14 Plural Affix
In verbs (except the present tense of the linking verb 'بودن' /'budan/ 'to be'), the changes may occur in personal and number endings, such as Example (5):

\[
\begin{array}{ccc}
\text{می خورد} & \rightarrow & \text{می خوره} \\
\text{mi.xor.ad} & \rightarrow & \text{mi.xor.e} \\
\text{IMPF.eat.3SG} & \rightarrow & \text{IMPF.eat.3SG} \\
\text{'he/she eats'} & \rightarrow & \text{'he/she eats'}
\end{array}
\]

It should be noted that only the verbs belonging to the standard level undergo this type of change and those having a literary register, such as "روییدن" /ruyidan/ 'to grow', "گشتن" /gaštan/ 'to become' or "رهانیدن" /rahānidan/ 'to save', with no application in informal settings, are not subject to such changes.

b) Multi-Unit Word Form
In this case, the phonological change crosses the word boundary and converts it into a combination. Here, a free morpheme changes into a bound morpheme and is necessarily attached to another word as its base, as in Examples (6) and (7).

\[
\begin{array}{ccc}
\text{منو} & \rightarrow & \text{منه} \\
o.man & \rightarrow & e.man \\
I/me.and/PAR & \rightarrow & is.me/I
\end{array}
\]

Accordingly, instead of 'من است' /'man ʔast/ 'is mine', we have "منه" /man.e/ 'is mine' in which the free morpheme "است" /ʔast/ 'is' changes into a bound one (ـه) due to a phonological change.

c) Word Meaning
At times, phonological change leads to homography, polysemy or ambiguity. Thus, for example, the linguistic form 'بَره' /bere/ could signify both /bere/ 'if he/she goes' and /barre/ 'lamb'.

5.1.5 Morpho-phonological: Deletion
Similar to phonological changes, deletion of a letter from a word may allow the following sub-categories:

a) Word Form
In verbs, the deletion may occur in the stem, in personal and number endings or in both, as in Example (8) to (10):

\[
\begin{array}{ccc}
\text{می آورم} & \rightarrow & \text{می آرم} \\
\text{mi.āvar.am} & \rightarrow & \text{am.ár.mi} \\
\text{IMPF.bring.1SG} & \rightarrow & \text{IMPF.eat.1SG} \\
\text{'I bring'} & \rightarrow & \text{'I bring'}
\end{array}
\]

\[
\begin{array}{ccc}
\text{می شوید} & \rightarrow & \text{می شین} \\
\text{mi.šav.id} & \rightarrow & \text{in.š.mi} \\
\text{IMPF.become.2PL} & \rightarrow & \text{IMPF.become.2PL}
\end{array}
\]

\[\text{Particle ار /rā/ as the direct object marker}\]
'you become'  'you become'

(10) 

\(\text{شوي} \rightarrow \text{شي}
\)

\(\text{šav.i} \rightarrow \text{iš}
\)

\(\text{become.2SG} \rightarrow \text{become.2SG}
\)

'you become'  'you become'

b) Multi-Unit Word Form

Here a due to deletion of a letter from the spelling of a word, it changes into a bound morpheme and attaches into another word as its base. For example, "/hā/ (the plural suffix) becomes '/ā/.

\((11)\)

\(\text{درختها} \rightarrow \text{درختا}
\)

\(\text{tree.PL} \rightarrow \text{tree.PL}
\)

'trees'  'trees'

\(\text{آن} \rightarrow \text{آنست}
\)

\(\text{it} \rightarrow \text{it.is}
\)

'it is'

c) Word Meaning

In such words, in addition to deletion, polysemy and meaning ambiguity may occur. For example, "شي may be pronounced either /ši/ 'you become' or /šei/ 'thing'.

5.1.6 Others

Studying the data, we came across word types which seem to have no sense in Persian, such as '

\(\text{هها}
\)

\(\text{دای}
\)

\(\text{میک}
\)

Some forms are likely to be part of a Quranic verse as '' or be used as a trademark not so familiar in the language, such as '' and ''. Obviously, these forms do not enter dictionaries, thus they have been classified under 'Others'.

5.2 Analysis

In this paper, we investigated and classified 1000 word types spelled variously from the integrated corpora. The sum of the frequency of this number of word types is 36,652,829 word tokens. The output of the extended Levenstein Distance algorithm was checked and analyzed qualitatively from the linguistics point of view. After giving a fine-grained classification of variation types, a statistical analysis is provided.

Table 1 presents the coarse-grained classification of the spelling variation in Persian in two main categories: orthographic and morpho-phonological. The third category, Others, includes word types having spelling error or being meaningless.

Each of the forms may be put in more than one category, such as "منو which has undergone a change in style due to a phonological change and is read either /man.o/ (a combination of 'من' and 'و') or /meno/ 'menu'. In most categories, a sub-category of ambiguous forms can be observed. It was also revealed that Orthography had the most and Phonology the least distribution, regardless of others.

Table 1 Coarse-grained Classification of Variation Types

<table>
<thead>
<tr>
<th>Coarse-Grain Category</th>
<th>Relative Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthography</td>
<td>71.35</td>
</tr>
</tbody>
</table>
The fine-grained classification of the categories for orthographical reasons is presented in Figure 1. According to the figure, the three major factors giving rise to spelling variation are wrong application of space word-internally or at the morpheme boundary, absence of tilde on Alef, and failing to follow the rules of joined-up writing. As for Morpho-phonological factors, ambiguity and change in register are two main reasons for spelling variation, as presented in Figures 2 and 3. The rate of these two factors seems to be equal in both figures.

![Figure 1](image1.png)
Figure 1 Fine-grained Classification of Spelling Variation due to Orthographical Reasons

![Figure 2](image2.png)
Figure 2 Fine-grained classification of spelling variation due to phonological reasons
6. Concluding remarks

As was previously mentioned, cursive nature of Persian script and also application of some phonological processes in informal speech, cause spelling variation to be a major challenge in this language, with joined-up writing and space application standing at the top. Thus, in complex words, if the space is applied at the morpheme boundary, unification in choosing each of the forms, joined or disjoined with space or without space, can be reached via preparing a good style guide. In this case, the corpus needs to contain morphological information of the words. If space is not applied at the morpheme boundary, unification can be gained through teaching the properties of Persian script and, thus, the correct writing style. Through identifying and classifying the phonological changes, we can make a list of such various forms and incorporate them into teaching materials for non-native language learners. Based on the analyzed data, issues related to the Persian script are the main sources for spelling variation. This indicates the importance of making the corpus usable and applicable for corpus linguistics and computational linguistics applications.

This research was based on studying 1000 word types. Obviously, investigating on more tokens will result in having a better picture on the topic to find the major problems that cause spelling variation in Persian.

References


Ghayoomi, Masood, Saghār Sharifi and Marziyeh Sanaati. 2015. Tannvo’e negāreshi dar zabāne Farsi va tahiyeye khodkāre dādegane emlāyi az peykareye zabānie mobtani bar veb (Spelling variation in the Persian language and automatic development of a spelling database from a Web-based corpus). In Proceedings of the 1st International
Conference on Web Research (ICWR). University of Science and Culture, Tehran, Iran, March 15-16.
Nasri, Musa. 1314. Bā khatte konunie Farsi che bāyad kard? (What to do with current Persian orthography?). Mehr 3(26): 200-205.
Towards Automatic Translation from Sanskrit to English

Chitresh Soni, Girish Nath Jha, Naveen Kumar Jha
Mohanlal Sukhadia University, India
soni.chitresh@gmail.com

Abstract
The paper is a reflection of ongoing research on building a Sanskrit-English Translator through statistical methods. In this work the help of Microsoft Translator Hub for training of system is taken. The main contribution of the research is to address complexities involved in making such system and ways to handle those challenges. For the purpose of training the system both parallel and monolingual data of Sanskrit and English as well as dictionaries of terms is used. Parallel data is manually translated from Sanskrit classical texts into English. After the initial phases of tuning, testing and evaluation we analyzed the issues and challenges arriving in it such as Named Entity Recognition [NER] and case markers which are discussed in the paper. The objective of the paper is to suggest the ways to better equip the statistical machine to learn and handle such ramifications with some sample analyzed data. Here to supplement the result a simple yet innovative way of morphological transformations to increase the efficacy of dictionary entries and assigning probabilities to the system is used. Both of that lead to improvements in the translation quality.

Keywords: Sanskrit-English translator, Microsoft translator hub, Sanskrit classical texts, named entity recognition, case markers, morphological transformations

1. Introduction:
The massive literature in Sanskrit language contains knowledge helpful for the well being of humanity. This vast knowledge should be made available to the masses. If a Machine Translation System (MTS) from Sanskrit to English is created then Sanskrit connects with many Indo-European languages because MTS from English to those languages are currently available. Sanskrit-English translation is also crucial because of the literary superiority of both the languages and also English being the lingua-franca of actual, scholarly and computer worlds.

2. Role of computers and field of MT
Why choose a Machine Translation system over a human translator? Although there is no such goal like substituting a human translator, yet sometimes we need quick and handy translations without worrying for pinpoint accuracy. Cost- effectiveness is another driver for Machine Translation researches because an MT system can translate for you at free of cost or at cheap license and installing charges. The factor of confidentiality also aides to the utility of an MT system just in the case of some important law draft or secret military docs. Universality is always an advantage with the MT systems as they can be accessed through the net at any point of time without the botheration to find a particular person. Moreover MT systems possess more consistency because they will always translate a input into a same way, so it is another example of viability of an MT system. And most importantly its neutrality and impartial nature, because biased human scholars sometimes more or less start interpreting the texts in an incorrect manner while translating them.

MT is a sub-field of computational linguistics that tries to use the computer software to translate text or speech from one natural language (source) to another (target). In lenient terms a MTS decodes the meaning of the source text of the given language and re-encodes this meaning in the target language. But behind this simple procedure there lies a complex cognitive operation. To decode the meaning of the source text in its entirety, computer must
interpret and analyze all the features of the text, a process that requires in-depth knowledge of the grammar, semantics, syntax, idioms, etc. of the source language, as well as the culture of its speakers. The system needs the same in depth knowledge to re-encode the meaning in the target language. Therein lay the challenge of machine translation; how to program a computer that will "understand" a text as a person does, and that will "create" a new text in the target language that "sounds" as if it has been written by a person.

3. Existing automatic translation systems in Indian languages

3.1. Anglabharati (1991)
Developed by Indian Institute of Technology, Kanpur, this system translates from English to Indian languages, primarily Hindi, using a rule-based transfer approach. The primary strategy for handling ambiguity is post-editing.

Originated at IIT Kanpur and later shifted to the Centre for Applied Linguistics and Translation Studies (CALTS), Department of Humanities and Social Sciences, University of Hyderabad, the Anusaaraka project’s goal is on Language Access between Indian languages. Anusaaraka system maps local word groups between the SL and TL, using principles of Paninian Grammar (PG).

3.3. The Mantra (MACHI Ne assisted TRAnslation tool) (1999)
The Mantra (Machine Assisted Translation tool) Project has been developed by C- DAC, Bangalore. Mantra translates English text into Hindi in a specified domain.

3.4. The Matra system (2004)
MaTra is an ongoing project at C-DAC, Mumbai. It aims at machine-assisted translation from English into Hindi, essentially based on a transfer approach using a frame-like structured representation.

Anubharti, another recent project at IIT Kanpur, is a template-based MT from Hindi to English, using hybrid example-based model for MT (HEBMT).

Two machine translation systems from English to Hindi, developed jointly by Carnegie Mellon University USA, Indian Institute of Science, Bangalore, India, and International Institute of Information Technology, Hyderabad. Shakti machine translation system has been designed to produce machine translation systems for new languages rapidly. Shakti system combines rule-based approach with statistical approach whereas Shiva is example based machine translation system. Some modules also use semantic information. Currently Shakti is working for three target languages, Hindi, Marathi and Telugu.

4. Difficulty in translating Sanskrit into English

The grammar of the Sanskrit language has a complex verbal system, rich nominal declension, and extensive use of compound nouns which create issues and challenges while translation to English. Other than grammar issues translation of cultural terms such as dharma etc. is not possible and also should not be forcibly done. So, teaching the computer that not to translate them and how to handle them is pretty necessary.

Now, divergence is a major issue in translation between two natural languages. The language divergence arises when lexically and syntactically similar sentences of the source language do not translate into sentences of identical structure in the target language. Depending upon the direction of translation, divergence is of two sorts: unilateral and bilateral. If divergence
occurs when we translate from Sanskrit to English or vice versa then this type of divergence is
fallen into unilateral type of divergence. If divergence occurs in round trip then this type of
divergence is called as bilateral type of divergence. The language divergences between
Sanskrit and English can be considered as representing the divergences between Subject -
Object - Verb and Subject - Verb - Object classes of languages. This phenomenon may occur
in any pair of languages for machine translation.

5. Undergoing research on SET System

For making Sanskrit to English Machine Translator, the support of Microsoft translator hub
was crucial. We used MST hub for training of data for making of system. In the process of translation the language to translate from is called Source Language and the
language to translate to is called target language.

Here the translation system we can make either may be domain specific or domain
independent. We are in a process of making a domain independent translation system.
To build a translation system for our project, we needed to assemble a set of documents. These documents were of two types

5.1. Parallel documents

This refers to a pair of documents in our set, both containing the same information. One
document in the pair contains sentences in the source language and another document in the
pair contains sentences in the target language. Parallel documents are used by the system to learn how words, phrases and sentences are
commonly mapped between the two languages, to learn how to process the appropriate
context depending on the surrounding phrases, and as a result a particular word may not
always translate the same way.

5.2. Monolingual documents

This refers to target language documents in our set for which we do not have a corresponding
source language document containing the same information. Monolingual documents help a
translation system decide which of the considered alternative translations is more accurate,
more natural-sounding, and more in-tune with the context in the examples of the target
language. The quality of the resulting translation system will depend on the no. of sentences in the
document set and the quality of the sentences. The more examples the documents contain on
diverse usages for a word, the better job it can do during the translation of something it has
not seen before.
The system requires a minimum of 10,000 parallel sentences to start with and then adding
more parallel content improves the quality of the translation system. When setting up
Trainings, Hub allows us to partition the documents under 3 mutually exclusive data sets.

5.3. Training data set

Sentences of parallel and monolingual documents included in this set are used by the Hub as
the basis for building the translation system.

5.4. Tuning data set

Sentences of parallel documents included in this set are used by the Hub to tune the
translation system for optimal results. The tuning set is used to set all parameters of the
translation system to the optimal values. The tuning set is selected manually, in order to
achieve the most representative selection of sentences. When we let the system choose the
tuning set automatically, it will use a random subset of sentences from the bilingual training
documents, and exclude these sentences from the training material itself.
5.5. Testing data set
Sentences of parallel documents included in this set are used to compute the BLEU (Bilingual Evaluation Understudy) score of the translation system.

Only bilingual documents can be part of the testing data set. Not more than 2500 sentences are needed as testing set.

When we let the system choose the testing set automatically, it uses a random subset of sentences from the bilingual training documents, and excludes these sentences from the training material itself.

![Figure 1 System architecture](image1)

**Figure 1 System architecture**

![Figure 2 Document upload](image2)

**Figure 2 Document upload**
6. Using dictionaries

We can specify a dictionary of terms that Microsoft Translator should use in translation, in addition to the training data for the translation system. Use of a dictionary has the potential of degrading the quality of the translations. Because training documents showing the terms used in context are better than a plain dictionary. Terms used in sentence form teach the system the correct inflection and agreement, better than a dictionary can. Also the dictionary maps the dictionary term or phrase exactly to the given translated form. So we tried to minimize the dictionary to the terms that are already in the training sets. The dictionary works well for compound nouns like product names proper names (“City of New Delhi”), or features of the product. It doesn’t work equally well for verbs or adjectives, because these are typically highly inflected. So we avoided dictionary entries for anything but compound nouns.

7. Initial challenges posed
In initial stages while evaluating and reviewing our system, we found many issues and challenges. Significant amount of them are difficulty recognising named entities, comprehending tense usage, understanding case markers and verbs. To improve on these divergent translations a large corpus of bilingual data is in process of being created including as many of varieties as possible of the source text which will help the system in handling these repercussions better. The platform is also improved time to time by Microsoft.

References