6th Biennial Conference of the Asian Association for Lexicography

ASIALEX 2009
“DICTIONARIES IN EDUCATION”

The Imperial Queen’s Park Hotel
Bangkok, Thailand
August 20–22, 2009

FINAL PROGRAM AND ABSTRACTS
http://www.kmitl.ac.th/asialex/

Organizers
The Asian Association for Lexicography (ASIALEX)
King Mongkut’s Institute of Technology Ladkrabang (KMITL)
WELCOME MESSAGE FROM ASIALEX PRESIDENT

On behalf of the ASIALEX Committee, it is our great pleasure to welcome you to ASIALEX 2009. This is the 6th biennial conference of the Asian Association of Lexicography following the previous one held in Chennai, India in 2007. The conference theme is “Dictionaries in Education”. The purpose of this conference is to promote activities in various areas of lexicography by providing a forum in exchange of ideas, presentations of research achievement, and discussions of future directions.

This conference is supported by King Mongkut’s Institute of Technology Ladkrabang, ASIALEX, Commission on Higher Education (Thailand), Australian Education International and Thailand Convention and Exhibition Bureau. This year, a total number of 49 papers have been submitted from 18 different countries. All submitted papers have been peer reviewed by the Program Committee. The program begins on Thursday (August 20, 2009), and comprises 54 sessions spanning a period of three days. The proceedings are provided in CD-ROM version. We are grateful to have six distinguished speakers: Professor Pam Peters (Australia), Professor Gregory James (UK/Hong Kong), Professor Yukio Tono (Japan), Professor Gilles-Maurice de Schryver (Belgium and South Africa), Professor Theraphan L-thongkum and Professor Suwilai Premsrirat from Thailand. The highlights of the Social Program include a conference reception on Thursday evening (August 20), and the Bangkok temple and city tour on Saturday (August 22).
Apart from attending the conference, you are warmly invited to explore the colorful and dynamic city of Bangkok. It offers not only the cosmopolitan amenities, but also a unique treasure trove of cultural attractions such as temples, palaces, traditional ways of life along the Chao Phraya River.

Finally, we would like to express our sincere gratitude to everyone involved in making this conference a success. It is our great pleasure to have you with us at the conference where we hope new ties are made and existing ones strengthened.

With our best wishes for a pleasant stay in Bangkok.

Jirapa Vitayapirak
President of ASIALEX (2008-9)
WELCOME MESSAGE FROM KMITL PRESIDENT

It is with great pleasure that the Asian Association for Lexicography (ASIALEX) joins the Department of Languages and Social Sciences, King Mongkut’s Institute of Technology Ladkrabang (KMITL) in organizing this International Conference on “Dictionaries in Education”. This Conference is especially important since Asia is home to many diverse languages and cultures. They are important both in themselves and in the role they can play in education. The purpose of this special event is to promote activities in diverse areas of lexicography, i.e. dictionaries in any languages, the practice of dictionary-making, research on dictionary use and users, corpus lexicography, terminology, dictionary writing systems, software tools and dictionaries for natural language processing by providing a forum for exchange of ideas, presentations of research achievement, and discussions of future paths of research and studies.

King Mongkut’s Institute of Technology Ladkrabang is recognized as a famous public higher education institution in Science and Engineering by the Thai Ministry of Education. In this special occasion, KMITL is very pound to host ASIALEX 2009 in Thailand. As the President of KMITL, I would like to express my gratitude to the Organizing Committee which gives us a valuable opportunity to be a part of the great success.

We wish all participants a pleasant stay in Bangkok.

Assoc. Prof. Dr. Kitti Theraseth  
President of King Mongkut’s Institute of Technology Ladkrabang
CONFERENCE COMMITTEE

The conference is organized by the Department of Languages and Social Science, Faculty of Industrial Education, King Mongkut’s Institute of Technology Ladkrabang (KMITL), the Asian Association for Lexicography (ASIALEX), with support from The Commission on Higher Education, Thailand.

**Honorary Chairperson:**

Assoc. Prof. Dr. Kitti Theraseth
President of King Mongkut’s Institute of Technology Ladkrabang (KMITL)

**General Chairperson:**

Assoc. Prof. Peerawut Suwanchan
Dean, Faculty of Industrial Education, KMITL

**Program Committee Chairs:**

Assoc. Prof. Dr. Jirapa Vitayapirak
Department of Languages and Social Science
Faculty of Industrial Education, KMITL

Asst. Prof. Dr. Wirote Aroonmanakun
Department of Linguistics, Faculty of Arts
Chulalongkorn University

Dr. Pattama Patpong
Institute of Language and Culture for Rural Development
Mahidol University

Nongsamorn Pongpanich
Department of English, Faculty of Arts
Kasetsart University
Co-Chairs:

Prof. Zafar Iqbal (Vice-President), Pakistan  
Dr. Amy Chi, Hong Kong  
Prof. Varadarajan Jayadevan, India  
Prof. Yukio Tono, Japan  
Prof. Shigeru Yamada, Japan  
Prof. Gilles-Maurice de Schryver, Belgium and South Africa

Organizing Committee:

Assoc. Prof. Sukhuman Nilrat  
Asst. Prof. Benjawan Ubonsri  
Asst. Prof. Chakaj Rachaburi  
Asst. Prof. Dr. Jirasek Trimetsoontorn  
Asst. Prof. Kanitta Ounarat  
Miss Chatpirom Surachet  
Asst. Prof. Amphon Thongra-ar  
Asst. Prof. Piya Supavarasuwat
ACKNOWLEDGEMENTS

The conference and Organising Committee Chairs wish to thank all the international advisors and members of the organizing Committee for the contributions in organizing this conference. The Chairs also wish to acknowledge all the sponsors for their generous support and all others who have in one way or another contributed towards the success of this conference.

ORGANIZERS

The Asian Association for Lexicography (ASIALEX)
King Mongkut’s Institute of Technology Ladkrabang (KMITL), THAILAND

SUPPORTED BY

Chulalongkorn University, Thailand
Mahidol University, Thailand
Kasetsart University, Thailand
Thailand Convention & Exhibition Bureau (TCEP)
Australian Education International, Australian Government
Commission on Higher Education, Thailand
ASIALEX 2009
CONFERENCE INFORMATION

Conference Venue: The Imperial Queen’s Park Hotel
199 Sukhumvit Soi 22
Bangkok 10110, Thailand
Tel: (66)2261-9000
Fax: (66)2261-9530-4
Email: queenspark@imperialhotels.com

Opening/ Closing Ceremony: Queen’s Park 3, Level 2

Plenary Sessions: Queen’s Park 3, Level 2

Oral Presentation Rooms: Queen’s Park 3, 4, 5, Level 2

Secretary/ Internet Station: Benjasiri 1, Mezzanine Floor

Coffee/ Tea Breaks: Queen’s Park 3 Foyer, Level 2

Lunch: Parkview Restaurant, Ground Floor

Conference Dinner: The Terrace, Level 9
(Dress Code: Smart Casual)

Conference Contact: ASIALEX Conference Secretariat
Department of Languages and Social Sciences
Faculty of Industrial Education,
King Mongkut’s Institute of Technology
Ladkrabang, Bangkok 10520 Thailand
Tel/ Fax: (66) 2326-4321
Email: asialex@kmitl.ac.th
Abstract: Terminography presents exciting new frontiers for development in university and professional education, where the understanding of key terms and concepts is crucial in building disciplinary knowledge. Experimental online termbanks created at Macquarie University (Sydney, Australia) are combining aspects of pedagogical lexicography, such as the use of plain English definitions and authentic corpus-derived examples, with a variety of digitized features such as audiofiles of the pronunciation of each term, translation equivalents in selected languages, and also graphics and ontological diagrams. It seems likely that these various cognitive inputs enhance both L1 and L2 students’ understanding of disciplinary terminology, and good results have been obtained in testing a statistics termbank with a large cohort of Australian and international students. There are however open questions to be researched, e.g. whether students gain more from ontological diagrams which complement the definitions or ones that work in parallel. Whether the answer is the same for the terminology of an abstract discipline (such as statistics) and a highly taxonomic one (such as biology) remains to be seen. The uses of these and other mediostructures in the online medium will also be discussed – as a means of reducing the isolation of individual terms that happens in the macrostructure of print-based dictionaries. Multimodal termbanks play their part in tertiary education, building sets of terms into memorable knowledge structures for the ESP learner, and expanding the dimensions of terminography.

Key Words: terminography; online termbanks; bilingualized; ontological diagrams; mediostructures

1. INTRODUCTION
With the turn of the second millennium, changing practices in terminology are bringing it much closer to lexicography. No longer must it be confined to terms with a single meaning, since metaphorical usages of everyday words can easily become terminology in specialized domains (Temmerman 2000). Traditional terminologists emphasized the “univocity” of terms, and that each should be uniquely attached to a given concept in a specialized domain. This approach has proved increasingly difficult to maintain, because terms quite often function in more than one discipline with different senses. Our classic example (from working in the terminology of biology and statistics) is cell, which in the first refers to the essential organism of life, and in the second to one of the quantities in the grid layout of a statistical table. There are of course yet other meanings for cell in electronics, in ecclesiastical history etc. The different terminological senses of a single word put a premium on definitions to discriminate between them, and thus the art of defining becomes an essential tool in terminography as in lexicography.

This is a radical shift in terminology, one which opens terminography up to alternative ways of rendering the meanings of terms. The need for multidimensional representations of concepts/terms in specialized domains was indeed flagged in the late C20 by terminologists
including Bowker and Meyer (1993), to cover both hierarchical and nonhierarchical relationships among them. Beyond verbal explanations, the value of nonverbal illustrations of terms is newly recognized dimension of terminography, as demonstrated by Faber et al. (2007) in their discussion of terms in coastal engineering. The combination of verbal explanations of terms with diagrams and graphics which are not so language-specific also recommends itself as a resource for L2 tertiary training. It lends itself to the needs of international students in Australian universities, and it opens up exciting possibilities for us which we hope to codevelop with partner universities in Asia and in Europe.

2. TERMINOGRAPHY AND ESP
These new developments in terminography coincide with a long-felt need in teaching LSP (language for specific purposes), that is the importance of focusing on the terminology of the discipline or domain – be it business or medicine or linguistics (Dudley-Evans and St John 1998). Understanding specialized terms is essential for building disciplinary knowledge, and underpins the students’ competence as practitioners of the discipline. It is the prime factor in reading comprehension (Chung and Nation 2003), and probably also subfactor in listening comprehension (Spearritt 1962).

The more discursive aspects of C21 terminography also expose L2 students to volumes of ordinary English through the definition of terms, and therefore has a considerable role to play in language learning generally. Provided the definitions are worded in accessible English, they will not only mediate the term’s specialized meaning, but assist students to conceptualize it in the L2. The definitions thus help to operationalize their English resources to some very specific learning goals, and to enlarge their competence in using the language.

Contemporary learners dictionaries make considerable use of corpus data to illustrate authentic usage of general language, and to acquaint learners with aspects of grammar and collocation beyond whatever the microstructure indicates in formal terms. Specialized corpora can be compiled out of teaching and reading materials used in the discipline for which a termbank is being developed. Examples from corpus texts can be extracted and quoted along with the definition, to show how the term appears in authentic technical texts. The language of the corpus examples may well be more demanding than that of the definition. But ESP students often bring domain or subject knowledge to bear on such examples, because they are contextualized within the discipline (Fernandez and Peters, 2009). By the same token, their familiarity with the domain helps to motivate language-learning in the context LSP.

3. BILINGUALIZATION OF MICROSTRUCTURAL ELEMENTS
Additional support for L2 learners of terminology can be provided through bilingualization of English language material within the dictionary entries. For technical dictionaries, this has so far mostly been done in the context of research (Vitayapirak 2001; Fernandez 2005). But pedagogical research on bilingualization in general dictionaries has shown that it enhances production for average and good L2 dictionary users, and comprehension for all users (Laufer and Melamed 1994:575). The user’s pattern of reference was also shown to vary, in that unskilled users would refer first to the bilingualized material, rather than using it simply to check their understanding of the monolingual material. Yet this raises the question as to the appropriate level of bilingualization for such dictionaries. If the whole entry is bilingualized,
it does not encourage students to conceptualize terms in the L2. Rather it underlines their exposure to the L2 – like the conventional bilingual dictionaries which most ESL/EFL teachers try to dissuade their students from using.

Bilingualized termbanks can limit the amount of L2 to giving translation equivalents for the headword, without going so far as to provide L1 translations of the definition(s), of any illustrative examples given, and of labels in diagrams. The rationale is that providing a translation equivalent of the headword/term is enough to reassure the students that they are on the right track with the word, and to anchor their understanding of the term in the language networks of their L1. But this upfront use of L1 is too slight to divert them from the larger volume of L2 material in the entry, and its relevance to the discipline in which they are immersing themselves should help to ensure the necessary immersion in L2 as well.

Any level of bilingualization of an English learners dictionary necessarily serves to identify a particular target L2 learner (e.g. Chinese, as with the Longman Dictionary of Contemporary English 2004), which limits its potential value to other L2 learners. In the online medium, translation equivalents can be provided in multiple languages, from which students can select. In our large accountancy course at Macquarie University, the largest cohort of international students comes from China, and so our termbank presently provides Chinese (Mandarin) translations of all headwords. But we plan next to add Thai translations to this termbank, in a joint project with King Mongkuts University, Ladkrabang, so that Thai ESL students of accounting can make the most of the technical content while also adding to their English resources.

Providing translation equivalents for the headwords does not (in our case) imply allowing students to search for terms in their L1, as in a bilingual (L1 to L2) dictionary. Taking this extra step would probably help unskilled ESP students, for whom the bilingualized dictionary proved less helpful in Laufer and Melamed’s (1994) research. It might however vitiate the value of the termbank for intermediate and advanced language learners, who are better able to use the L2 resources, and to advance their knowledge of English thereby, as well as their understanding of technical concepts.

4. GRAPHICS AND ONTOLOGICAL DIAGRAMS

Technical dictionaries, bilingual or otherwise have been rather slow to provide graphic illustrations of terminology. This is partly a matter of publishing economics as well as the relentless march of technology, which quickly makes technical visualizations out of date. Illustrations of scientific concepts have been found to be less helpful than expected to novice students (Hill 1988). The shapes and proportions of objects represented in line drawings can be misconstrued, and the placement of labels on organic structures, e.g. those used to name different components of the human brain, may leave it unclear whether the label refers to the whole structure or just the closest part of it.

In their paper on civil engineering terminology, Faber et al. (2007:41) sought to “mesh” the wording of the definitions with visual information in the graphic images provided. Beginning with a definitional template using the familiar parameters of terminology (agency, constituency, location, process, function etc.), they then established a typology of images, in terms of their iconicity relative to the concept, its relative abstraction (the “cognitive effort required for [its] recognition”, p.51) and/or dynamism (i.e. movement). Thus the iconic
image could show what the object, e.g. a sea wall, was made of (i.e. its material constituency), whereas its location could be indicated by means of an abstract black-and-white map. A process such as *dredging* calls for a dynamic (before and after) diagram, as do concepts where a particular function is salient and needs to be explained. Since each image is intended to capture just one of the salient characteristics, it takes 4-6 different images to support the definitional template. The downside then is a miscellany of drawing styles for a single concept, each presenting its own kind of abstraction. It is an open question whether this verbal and graphic matching of each parameter of a concept would – cumulatively – mediate it more effectively to the reader, rather than integrating them in a holistic image. The paper does not provide any experimental evidence of their reception.

Outside terminology per se, alternative approaches to graphic illustration allow that they may relate in other ways to verbal text, for example in a complementary relationship, rather than one which seeks to match the graphic with the parameters of the concept, as defined. Unsworth (2006) found this happening naturally in textbooks where illustrations accompanied a discursive treatment of a concept. But once again there was no experimental testing of how readers might apprehend the combination. With all the preceding and surrounding text to enlarge their understanding of the term, it would be impossible to test the interplay between the graphic and verbal explanation of a term/concept on a single page of the textbook, or their interplay with any glossary at the back of the book.

These approaches to linking terms and graphics raise multiple issues about the representations of terminology, whether they should be atomistic or holistic, and the relative value of matching or complementary illustrations. Both approaches were based on paper presentations of the material and bound by the physical limitations of hard copy. An online termbank breaks free of these constraints, so that the content and quantity of verbal text that goes with each graphic can be managed, and thus the interplay between particular types of graphics and definitions (matching and complementary) is more easily tested. In practice, termbank definitions often contain both kinds of relationships. In our biology termbank, where *RNA* is defined as “one of the two main types of nucleic acid, the other being DNA”, but detailed graphics present the different structures of *RNA* and DNA side by side. In life sciences, a single holistic diagram with its labels often adds to the definition of the organism or a biochemical processes, and is not confined to matching it. Likewise in social sciences such as statistics, where disciplinary concepts are constructed via abstract relations, the concept of *outlier* can be effectively communicated by means of a statistical graph. In both kinds of discipline, complementary illustration in a holistic image clearly has its place.

One other kind of graphic which has a role to play in online termbanks is the ontological diagram. These are analogous to the hierarchical structures (also called ontologies) which are widely used in computational linguistics for the purposes of information retrieval. They suit the computational need for top-down intelligent searching, and lend themselves to parallel treatments of conceptual structures and terminology in different disciplines. Specialized subject ontologies have a place in tertiary and professional education (Dicheva 2008) as a “semantic backbone” for teaching structured sets of terms, showing how they cluster under larger, inclusive headings which represent the higher conceptual levels or organic unit. Ontologies may represent hyponymic relations (that of belonging to the same class), as in the classification of *nitrogenous bases* below; or meronymic ones (i.e. part-whole relationships), e.g. the three components of the *balance sheet* in accounting. Either way, these relationships can be diagrammed by means of an electronic tree diagram, in the vertical or horizontal
plane. Though the vertical diagram seems more intuitive, it is hard to label the nodes to explain whether they represent hyponymy or meronymy, and the ontology is ultimately ambiguous.

```
balance sheet

asset  liability  equity

nitrogenous bases

adenine  cytosine  guanine  thymine  uracil
```

In horizontally displayed ontologies, labels identifying the two different types of relationship can be built into the diagram itself, although they do not then provide a nonverbal means of explaining the semantic relationships among the sets of terms:

```
balance sheet contains asset
contains liability
contains equity
```

Of course the definition of each term may explain how it relates to the higher level concept (“is a type of”/“is a part of”), and the diagram (at least the horizontal one) becomes redundant. Better complementation can be achieved by using vertical diagrams which are differentially labeled or colored, according to whether they display hyponymic or meronymic relations. Either way the opportunity to make ontological diagrams a regular feature of the microstructure of a termbank is a great advantage of the online medium.

5. MEDIOSTRUCTURAL ELEMENTS IN ONLINE TERMBANKS

Between the macro- and microstructure of a dictionary, there is typically a mediostructure of crossreferences that take readers from one entry to another, or from one part of the dictionary to another. They may also reference the reader to sources external to the dictionary per se, as Nielsen (2003) argues, although their contribution to the mediostructure of a particular publication is then rather abstract. In the electronic medium, the distinction between external and internal crossreferences is less apparent because both involve linking to other pages on the internet, and users are not always aware of when the link has taken them outside the website on which they began.

The notion of mediostructure predates electronic lexicography, in the various type of crossreferencing used in print dictionaries to help the reader find related words. However the motivation for the crossreference is often unstated, and readers may be surprised at where they are taken. Ideally, as Nielsen (2003) comments, the cross-reference is “function-driven”, i.e. it anticipates where the termbank-user is coming from and why they might need to go laterally to find relevant information. Since the terms of a specialized domain are
related ontologically, the termbank maker can predict which other members of the same set are relevant to the search, and propose them for further inquiry. The mediostructure within online termbank is thus a good deal more concrete than that of the print dictionary, especially when illustrated by means of ontological diagrams, like those we have just seen.

While ontological diagrams display the semantic mediostructure of terms within each specialized termbank, other kinds of lexical relations are also designed into the TermFinder mediostructure. Morphological relations are an important way of expanding lexical resources, and thus Biology students who are aware that *transcribe* is the operative verb relative to *transcription* not only have two inputs to the same concept, but also a useful model for the numerous such verb/noun pairs in English academic vocabulary. Speakers of European languages which have borrowed from Latin are usually aware of this, but they are less apparent to L2 students with Asian mother tongues. For them, it is probably helpful to point out these “related words”, i.e. words related by inflection or derivation to the term being defined on that page. If the inflected form of the term is also quite frequent, we usually provide an example of it, alongside the uninflected one. These alternative forms are interconnected through a hidden network of annotations, to supplement the main query system, and they greatly expand its reach over and beyond the macrostructure of terms for which pages have been developed, and the entry points to information about a particular term. Thus students who type in a query for the biological process *blebbing* (for which there is no page), will be taken to a redirection page which provides a link to the page for *bleb*, and explains that their query “has the same stem as *BLEB*”

In online termbanks, the notion of function-related mediostructure becomes a real rather than abstract aspect of dictionary design, customized for the novice by providing a structured approach to the concepts and terms of the domain. Ontological diagrams and the network of query-specific annotations both play their part in the crossreferencing system – to meet tertiary students at their point of inquiry, and explain how it relates to the disciplinary terms and concepts they will need.

6. EXPERIMENTAL TESTING OF MULTIMODAL TERMBANKS

The features of terminography discussed above have all been designed into the TermFinder termbanks at Macquarie University. They provide multiple cognitive inputs to the learning of terminology and disciplinary concepts, and the combination seems to be effective. We can say this with some confidence, on the basis of large-scale testing of termbanks containing these facilities (apart from translation equivalents), which was carried out in 2007 and 2008 with full classes of students in first year Statistics and Biology. The value of the Statistics termbank can be seen in the improved exam scores of students who used it during the semester, by comparison with those from the previous cohort in 2007 who had no access to it. The results showed uniformly higher scores for the experimental group on the same set of exam questions, as shown in the table below:

<table>
<thead>
<tr>
<th>GROUP</th>
<th>No. of students</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007 cohort</td>
<td>1178</td>
<td>4.477</td>
<td>2.182</td>
<td>0.065</td>
</tr>
<tr>
<td>2008 cohort</td>
<td>1460</td>
<td>5.743</td>
<td>2.228</td>
<td>0.058</td>
</tr>
</tbody>
</table>
The difference between the mean student scores in 2007 and 2008 is highly significant, with a p-value of < .0000. The lecturer-in-charge (1) noted also that in 2008 almost all students actually tackled the terminological questions, in contrast with 2007, when quite a few did not attempt them. Evidently the availability of an enriched termbank gave them greater confidence in dealing with such questions. This also suggests that their language resources were sufficient to at least understand the question.

These experimental results are gratifying, but since they are based on holistic testing of the termbanks, they do not shed light on the kinds of questions raised earlier (#4) about the relative importance of the termbank’s various components. Fortunately the modular structure of elements within each termbank allows us to create experimental versions with or without certain components, and to research their effectiveness with L1 and L2 tertiary students. The contribution of the different components to their understanding of technical terms/concepts can be tested and their interplay. It may be that complementary forms of illustration are less satisfactory for L2 students than for L1s, whereas matching illustrations are rather redundant for L1s. The cognitive inputs of the various terminographic elements are easier to analyse in the context of specialized domains, where terms provide a sharper focus than general words of the language. Because of its highly structured pages, TermFinder termbanks lend themselves to comparative testing of the individual components of microstructure, such as:

i. Definitions and example sentences. Do the example sentences help to enlarge students’ understanding of the term as intended? Or is their generally more difficult language too challenging?

ii. Parallel or complementary forms of illustration. Are graphics that match up with the wording of the definition more or less useful than those which are complementary? Does this vary with individual disciplines?

iii. Use/nonuse of translation equivalents. Do these work better in some disciplines than others, e.g. “hard” sciences rather than social sciences?

The findings from this further experimental research will provide a basis of evidence on which to optimize structure and contents of termbanks for tertiary students (L1 and L2), according to the discipline they are working in. We hope they will also contribute to some of larger debates about how people respond to multimodal inputs, and whether they are more likely to produce cognitive overload when too closely matched (Sweller 2002).

7. CONCLUSION
Terminographers are able to capitalize on many recent advances in lexicography, and combine them in online medium to optimize learning and teaching of terminology in university education, especially for L2 students. Multimodal termbanks also provide enriched language experience – vital for language learning as well as building the conceptual knowledge needed for university disciplines.

NOTE
(1) The quantities shown in this table are extracted from data kindly supplied by Dr Jenny Middledorp, who was convener of the STAT170 unit in the Department of Statistics, Macquarie University in 2007 and 2008. Her report on early testing of the statistics termbank is contained in Peters et al (2008). Her participation in the TermFinder project is most gratefully acknowledged.
REFERENCES


Emeritus Professor Pam Peters
Macquarie University, Australia
Pam.Peters@ling.mq.edu.au

Emeritus Professor Pam Peters’ work with the Dictionary Research Centre and Style Council Centre at Macquarie University in Sydney has made her one of the country’s leading authorities on Australian style. Her major publications on writing and usage issues are the Cambridge Guide to Australian English Usage (2007) and the Cambridge Guide to English Usage (2004). She recently coedited an anthology of corpus-based studies in Australian and New Zealand lexicogrammar, to be published in July 2009 by John Benjamins, Amsterdam. At Macquarie University's new Centre for Language Sciences, she currently leads the TermFinder project in pedagogical terminography, creating termbanks to assist second-language students in understanding the terminology of their university disciplines, such as accounting, biology, geology, statistics. Online termbanks can support students in many more ways than the traditional glossary.
PLENARY SESSIONS

Plenary I

MULTILINGUAL DICTIONARIES FOR LINGUISTIC AND ACADEMIC PURPOSES

Theraphan L-Thongkum
Chulalongkorn University, THAILAND
Theraphan.L@chula.ac.th

Many languages of the ethnic minorities inhabiting the Southeast Asian region are exclusively spoken or, in other words, only the aural medium is used for communication among the members of each ethnic group. This is due to the fact that traditional orthographies and writing systems do not exist. Until recently, these languages have been analysed and described by linguists and Christian missionaries for their specific purposes. Thus, orthography inventing and dictionary making can be regarded as a practical and useful start for linguistic research, bible translation and literacy projects.

As an experimental phonetician and, at the same time, a field linguist who has worked on approximately 30 unwritten SEA languages spoken in Thailand and its neighbouring countries, I quite often have to play a “duel role”, i.e. as compiler and user of our multilingual dictionaries. Consequently, other specialists of SEA linguistics and linguistic students can obtain some language data from these dictionaries for their further serious linguistic investigations. The focus of my presentation will be on the methodology, process and problems of making multilingual dictionaries both without and with the aid of computer, based on my direct experience from 1979 up to the present and from the view point of an amateur lexicographer.
Prof. Theraphan L-thongkum, Ph.D.
Chulalongkorn University, Thailand
Theraphan.L@chula.ac.th

Prof. Theraphan Luangthongkum is a professor in linguistics at the Faculty of Arts, Chulalongkorn University. She received her PhD. in Phonetics from University of Edinburgh, UK (1977). She was the Dean, Faculty of Arts, Chulalongkorn University. She received many awards such as distinguished professor award of the year 2006 from the Eakin Laugesen Memorial Fund and outstanding research award of the year 2007 from the Thailand Research Fund (TRF).
Plenary III

DICTIONARIES IN EDUCATION – EDUCATION IN DICTIONARIES

Gregory James
Hong Kong University of Science and Technology, HONG KONG
lcgjames@ust.hk

OED2 defines the dictionary as “a book dealing with the individual words of a language (or certain specified classes of them), so as to set forth their orthography, pronunciation, signification, and use, their synonyms, derivation, and history, or at least some of these facts…” The focus on “individual words” can, however, be misleading. In the context of a second language, what learners need is help with the use of words in context, information that is often missing, or scantily treated in print dictionaries. The move from print to the electronic medium allows great scope for responding to learners’ differing needs simultaneously, and for affording convenient access to a range of contextual sources for determining modes and styles of expression, and the appropriate choices of words and phrases - in short, putting the “diction” back into the dictionary. This presentation will exemplify some of the work which is being accomplished in this domain in Hong Kong.
Prof. Gregory James
Hong Kong University of Science and Technology, Hong Kong
lcgjames@ust.hk

Gregory James is a professor in Linguistics and the Director of the Language Centre at Hong Kong University of Science and Technology, Hong Kong. He is an expert in Tamil lexicography. He has been researching and writing many books in Lexicography. His major publication with R.R.K. Hartmann is Dictionary of Lexicography.
Plenary IV

DICTIONARY OF ENDANGERED LANGUAGES IN THAILAND

Suwilai Premsrirat
Mahidol University, THAILAND
lcspm@mahidol.ac.th

Ethnic minority languages around the world are in crisis. In Thailand at least 14 languages are severely endangered. Other languages are at different stages of endangerment. The loss of language is the loss of knowledge systems and local wisdom recorded in the lexicon. Therefore lexical documentation for making a dictionary of endangered languages is urgent. This paper discusses data collection in endangered ethnolinguistic communities and types of dictionaries developed by linguists and by villagers. The characteristics of dictionaries of endangered languages will be discussed in terms of form and content. Dialects are recorded as much as possible. Lexical items belonging to various semantic fields related to the way of life and natural environments of the speakers are alphabetically arranged within each semantic field. The structure of the lexical entry includes the single word as well as phrase. Information about the speakers and their language is provided as well as the geographical distribution of the speakers, cultural and pragmatic information about the use of the languages and the influence of major languages on the development of minority languages. The word-finder lists for the vocabulary in the dictionary are alphabetically arranged and rhyming lists are provided for further work on language development and language standardization in the language revitalization process.
Suwilai Premsrirat is a professor in Linguistics at Mahidol University, Thailand. She received her Ph.D in Linguistics from Monash University, Australia. She received the best researcher awards from National Research Council of Thailand in 2006 and CIPL (Comite International Permanent des Linguistes) for The Endangered Languages in 2008. She has been researching and writing on ethnic minority languages in Thailand and mainland SEA since 1975 specializing in Austroasiatic languages. Her major publication includes a *Thesaurus and Dictionary Series of Khmu in SEA. The Ethnolinguistic Mapping of Thailand* is an important work conducted by Professor Suwilai and her research team during 1990-2000. She was the founder of the Resource Center for Revitalization and Maintenance of Endangered Languages (in 2001). At the moment she has been involved with 11 ethnolinguist group working on language revitalization program at different stages.
Featured Speaker I

SELECTING DICTIONARY INFORMATION BASED ON CRITERIAL FEATURES IN L2 LEARNER CORPORA

Yukio Tono
Tokyo University of Foreign Studies, JAPAN
yukio.tono@gmail.com

This study will report an on-going project of re-designing the selection of information in the entries of pedagogical dictionaries based on criterial features in L2 learner corpora. By criterial features, we mean those lexico-grammatical features that distinguish one proficiency level from another. In our project, a corpus of Japanese EFL learners' free compositions called the JEFLL Corpus was compiled. More than 10,000 students participated in the data collection, which resulted in a corpus ranging from the first year of junior high schools to the third year of senior high schools, thus a pseudo-longitudinal design. All the composition data was corrected by a native speaker, which produced the two aligned versions: the original and the corrected versions of the corpus. Then we performed an automatic extraction of differences between the two versions, using the Dynamic Programming algorithm. This output was exported to further multivariate analysis, identifying the relationship between those features and the school years. Based on these potential criterial features, I argue that the entries in a learner's dictionary should be re-organized in such a way that a graduated presentation is to be made for the information needed at different levels of proficiency.
Featured Speaker II

THE LEXICOGRAPHIC TREATMENT OF IDEOPHONES
AN IN-DEPTH CASE STUDY FOR ZULU

Gilles-Maurice de Schryver
Ghent University, BELGIUM
University of the Western Cape, South Africa
TshwaneDJe Human Language Technology, South Africa
gillesmaurice.deschryver@UGent.be

Ideophones, a wordclass that only exists in some of the world's language families, is a striking feature in Bantu. While the unique morphology, semantics and syntax of ideophones have all received considerable attention in the scientific literature, lexicographic aspects have unfortunately largely been ignored. This is surprising as ideophones are a lexicographer's worst nightmare. In a recent corpus-driven Zulu–English dictionary project, their compilation took an average three times longer than the compilation of entries in any other word class. Ideophones in Zulu are not particularly frequent (only five make it into the top 1500 lemmas, with about 100 in a dictionary covering the 5000 most frequent lemmas), but even when translating into Zulu, mother-tongue speakers feel the need to introduce them (cf. gqwa, qatha). Lexicographically speaking, meanings assigned to ideophones should not be too broad neither too specific: the art is to get the right level of generalization as far as the semantic import is concerned, with the examples functioning (true to their core function) as possible instances only (cf. do, khumu, mpo, ngqi). The result for the bilingual dictionary compiler is that no fool-proof translation equivalents can be provided, rather, each sense of an ideophone is merely a paraphrase of the semantic import, with the examples not random but hand-picked instances to illustrate the range of possible uses. Compared to monolingual dictionary making, the absence of a corresponding word class in the non-Bantu language, forces the bilingual dictionary maker to come up with various strategies to 'translate/transpose' the examples. When the ideophone is onomatopoeic-like, an English 'sound word' can be inserted (cf. nsi, pho), or recourse can be taken to an English 'exclamation' (cf. nci, qhu). In most cases, however, the Zulu ideophones
need to be translated with English 'verbs', and sometimes also as 'noun phrases' (cf. *dw*ī, *qu*), 'adjectives' (cf. *kli*wi, *nse*), 'adverbs' (cf. *gelekeqe*, *ngqa*), 'phrases' followed by clauses (cf. *phecelezi* > "in short, also known as"), and at times the ideophone simply 'disappears' as it is rendered by the English translation in full (cf. *juqu*, *shu*). Ideophones also tend to 'stress' what is already there, meaning that their function at that point in the sentence is to give more weight to the action expressed by the verb, the pronoun, etc. (cf. e.g. *qhwa*ba). For exemplification purposes in (bilingual) dictionaries, such instances are avoided, but not skipped altogether. While some ideophones could be said to stand on their own, corpus evidence clearly shows that one out of five typically combines with specific words (cf. *bha* > *emini kwabha*, *ngci* > -*gcina ngci*, -*vala ngci*), and still others only appear in so-called fixed expressions (cf. *bhug*e > -*myama bhug*e, *nhlo* > *selokhu kwathi nhta*). This corpus evidence, of course, needs to be recorded in the dictionary.

In addition to the corpus-driven lexicographic treatment briefly illustrated above, the paper will also compare the resulting articles with the treatment in existing (non-corpus-driven) dictionaries. In this regard, frequency considerations will be discussed on both macro- and microstructural level, and the differences highlighted. One aspect where both lemmatization approaches have – surprisingly – resulted in an analogous treatment, is the carving up of the homonym-sense continuum. An explanation why this is so will be attempted.
CORPUS-BASED GENERATION OF SUGGESTIONS FOR CORRECTING
STUDENT ERRORS

Richard Watson Todd
King Mongkut's University of Technology Thonburi, THAILAND
irictodd@kmutt.ac.th

Abstract: Since the COBUILD project, corpus analysis has become central to dictionary
development. In learners' dictionaries, however, there appears to be one component not based
on corpus analysis, namely, suggestions for correcting common learner errors. This paper
presents an attempt to identify corrections for learner errors automatically based on corpus
analysis. Initially, a tagged corpus of 100 million words and 200 sample student sentences
containing local errors were collected (e.g. It contains (word A) into (word B) three main
parts). Based on z-scores of word-word and word-word-word co-occurrences in the corpus,
suggestions for corrections were generated using six methods: 1. delete word A; 2. delete
word B; 3. switch words A and B; 4. replace word A with a different word; 5. replace word B
with a different word; 6. insert a word between words A and B. The most appropriate of these
generated suggestions was identified by a native speaker for each sentence. Using chi-square,
eight rules for identifying the most appropriate suggestion based on parts of speech were
identified. An example rule is: use method 3 when word A is a noun and the replacement
word-word combination has a high frequency of co-occurrence. These rules were then
validated against a further 100 student sentences. The validation showed that some rules are
reliable while others require adaptation. The new rules have now been incorporated into a
grammar checker program to generate suggestions for over 40 million potential word-word
errors.

1. INTRODUCTION
This paper examines procedures for the automatic corpus-based identification of corrections
for errors in student writing. Although the work reported here was completed as part of a
project to create a grammar checker for non-native speakers, it has implications for the
design of learners' dictionaries. Since the publication of the COBUILD dictionaries, the use
of corpora has become central to dictionary development. This influence can be seen clearly
in the examples of use and typical grammatical patterns included in learners' dictionaries.
However, there appears to be one component of learners' dictionaries that is not based on
corpus analysis, namely, suggestions for correcting common learner errors. Such suggestions
appear in dictionaries of common errors (e.g. Alexander, 1994; Intaravitak, 2004) and in
usage notes in some dictionaries (e.g. Longman Active Study Dictionary). While the growth
in use of corpora of learner English (Hunston, 2002) has resulted in research identifying
common learner errors, there has been almost no corpus-based work on how these errors can
be corrected. This paper examines this issue.

2. CORRECTING ERRORS
In contexts where a writer needs to self-correct writing, feedback from a superior language
system is needed (Watson Todd, 2001). After all, if a writer has made an error, there is little
reason to expect them to be able to self-correct without some external feedback. For
correction of local errors in writing, there are four types of superior system typically
available. In language teaching contexts, teachers are the most usual source of feedback with peers also being available. In situations of independent language use, however, neither teachers nor peers may be available and so writers need to find other sources of feedback for corrective revision of their work. One possibility is to use a dictionary – a bilingual dictionary to check translations of words known in the first language, a monolingual dictionary with grammatical patterns of word use that can act as models for self-correction, or a dictionary of common learner errors. A further possibility is computer programs such as grammar checkers that provide feedback on grammatical and lexical errors.

2.1 Using Dictionaries for Self-correction of Writing
While dictionaries may be used most frequently for purposes of translation (Chan, 2005) and for receptive purposes more than for productive purposes (Harvey and Yuill, 1997), they can still be a useful resource for written production of language. Research into self-correction of written work found that checking in a dictionary was the most frequently used strategy with a success rate of more than half of attempted corrections (Kubota, 2001). To be effective in self-correction, dictionaries, especially monolingual ones, need to give more than just definitions. Examples of use (Summers, 1988) and grammatical information (Bogaards and van der Kloot, 2001), such as the coded syntactic information in the COBUILD dictionary (Harvey and Yuill, 1997), are particularly helpful. However, effective use of such information depends on the awareness and analytic ability of the user. An alternative available for less capable users is to provide directions for how errors can be corrected.

The main dictionary-based source of suggestions for error correction is dictionaries of common errors. For example, both Alexander (1994) and Intaravitak (2004) provide pairs of sentences where one exhibits an error commonly made by learners and the other shows the correct form of the same sentence. To use these dictionaries for self-correction, learners need to realise that they have potentially made an error, find the keyword in their sentence for the error, look up the keyword in the dictionary, compare the example sentences to their own sentence, and then try to apply the repair suggested to their writing. Clearly, some awareness and analytic ability is required to do this, but it is much less than is required to correct a sentence based on the generalised part-of-speech patterns found in many dictionaries.

There are, however, drawbacks with using dictionaries of common errors for self-correction. Most obviously, a learner needs to be aware of the potential for making an error, but there are also drawbacks inherent in the dictionaries themselves. First, the number of errors that can be included in such a dictionary is limited, and many learner errors do not fit predictable patterns. Thus, the extent to which a dictionary of common errors can lead to self-correction is unclear. Second, the errors and suggestions in dictionaries of common errors are linked to specific words and are not generalisable. Again, this limits the number of errors that can be dealt with through the use of such dictionaries. Third, dictionaries are based around lexis and thus the errors that can be treated through a dictionary are limited to lexical errors. Although several errors which are traditionally regarded as being grammatical errors can be reconceptualised as lexical errors, several frequent error categories, such as subject-verb agreement, cannot be dealt with through a dictionary-based approach.

2.2 Using Grammar Checkers for Self-correction of Writing
An alternative approach which overcomes the limitations of dictionaries is grammar checkers. A grammar checker is a program that analyses written input, identifies errors (typically only local grammatical and lexical errors) and gives suggestions for correcting the
errors. Ideally, the program should have zero overflagging while identifying and giving valid suggestions for as many errors as possible (Tschichold, 1998).

There are three main approaches to designing a grammar checker (Dodigovic, 2005; Su et al., 1996). First, a grammar checker can be based on pattern matching (e.g. Hu et al., 1999; Lawley, 2004; Moré et al., 2004). Second, it can parse an input text and then match this against rules induced by human experts (e.g. Liou, 1991; Liu et al., 1993; Rider, 2005; Wojcik et al., 1993). Third, it can take a probabilistic approach by checking the likelihood of certain bigrams and trigrams being correct against a corpus (e.g. Chodorow and Leacock, 2000; Kann, 2002; Sjöbergh, n.d.).

Grammar checkers can cover both grammatical errors (such as the effective identification of errors and suggestions for subject-verb agreement in the Microsoft Word grammar checker) and lexical errors. For the latter, a probabilistic approach to designing a grammar checker is needed. The main advantage of a probabilistic approach to identifying learner errors is that a vast number of potential errors can be dealt with – many orders of magnitude more than can be dealt with in a dictionary of common errors. The disadvantages are that only a small number of these errors are likely to occur in learner writing, and that how suggestions for correction can be given for the errors identified is problematic. This paper examines the issues involved in using a probabilistic approach to identify suggestions for correcting learner errors.

3. IDENTIFYING SUGGESTIONS FOR CORRECTION THROUGH A PROBABILISTIC APPROACH

A probabilistic approach is based on the likelihood of immediate co-occurrence of pairs of words (bigrams) and trios of words (trigrams) ascertained through analysis of a large corpus. In such an approach error identification is relatively straightforward. The corpus can be analysed to identify frequency of occurrence of all bigrams and trigrams and also to calculate z-scores for their co-occurrence. Z-scores measure the probability of two (or three) words co-occurring given the overall frequency of each of the words in the corpus. The most common use of z-scores in corpus analysis is to identify strong collocations with Barnbrook (1996) recommending that z-scores greater than 3.0 represent collocations. Taking the opposite perspective, negative z-scores can be seen as representing unlikely bigrams and trigrams indicative of errors. Taking both frequencies and z-scores into account, we can set levels at which a given bigram or trigram should be considered an error. In the grammar checker project discussed in this paper, using a 100 million-word corpus, any word-word bigram occurring with an overall frequency of 10 or less and with a z-score of –3.0 or less is taken as an error. For the 6,700 most common words in English, roughly 40 million of the potential 44 million possible combinations of co-occurrences can be categorised as errors.

Such a straightforward analysis cannot be used for identifying valid suggestions to correct the erroneous bigrams and trigrams for three main reasons. First, it is not clear how an error should be corrected. Second, there are a massive number of potential corrections for any error. Third, the semantic and pragmatic context is a major influence on which of the potential corrections is most appropriate. An example should help make these points clearer. Let us suppose that the bigram the at has been identified as an error. On what basis should a correction be attempted? Should one of the words be deleted? Should one of the words be replaced by another word? Or should a word be inserted between the and at? Whichever of these bases is used, a large number of potential corrections are available. For example, if it is
decided to attempt a correction based on inserting a word, there are tens of potential corrections with a relatively high frequency in the corpus. Choosing which of these is the most appropriate normally depends on the wider context of the sentence. Since accounting for context is almost impossible to automatise, a different approach is needed. In this study, large numbers of possible suggestions for corrections were generated for 200 errors taken from a learner corpus. These possible suggestions were then evaluated against a range of factors to identify potential automisable bases for identifying valid suggestions. Finally, these automisable bases were evaluated using further learner errors.

3.1 Procedures for Identifying Suggestions for Correction

To be able to automatically generate potential corrections for learner errors, first these errors need to be identified. This study considers only word-word bigrams (although word-part of speech bigrams, part of speech-word bigrams, part of speech-part of speech bigrams, and word-word-word trigrams were also dealt with as part of the overall project to create a grammar checker). Using a 100 million-word tagged corpus, frequencies and z-scores for all potential word-word bigrams for the most common 6,700 words in English were generated and cutoff points for identifying errors were identified resulting in 40 million possible word-word bigram errors.

To generate suggestions for correcting errors, six methods were used. These are explained using the example sentence i where the word-word bigram error is underlined.

\[ \text{i} \quad \text{He drives a red colour car.} \]

A B C D

In this sentence, the word-word bigram error is the co-occurrence of \textit{red} (Word B) and \textit{colour} (Word C). These words occur in the context of Words A and D. Any words added to the sentence in generating a suggestion are termed Word E. The six methods are:

Method A: Delete Word B.
Method B: Delete Word C.
Method C: Switch Words B and C.
Method D: Replace Word B with another word (Word E) based on highest z-scores of resulting word-word-word trigrams.
Method E: Replace Word C with another word (Word E) based on highest z-scores of resulting word-word-word trigrams.
Method F: Insert a word (Word E) between Words B and C based on highest z-scores of resulting word-word-word trigrams.

Methods A, B and C each result in a single suggestion for correction. For methods D, E and F, the 7 highest rated trigrams were considered. Therefore, each error could have between 3 and 24 suggestions for correction.

These methods were implemented using 200 example sentences taken from a learner corpus (with the expected errors being identified in 191 sentences). The resulting possible suggestions for corrections were evaluated on three bases: 1. a linguist identified the single most appropriate suggestion for the context (for sentence i, method B); 2. the linguist rated the appropriacy of each suggestion on a 3-point scale; and 3. the average z-score rating for all word-word bigrams involving words A to E in the revised sentence was calculated.

To identify patterns of appropriate suggestions based on these three methods of evaluation, three automatisable factors were considered. First, for Words B, C and E where appropriate in revised sentences, relationships between the parts of speech (grouped into 6 categories)
and evaluation ratings were investigated. Second, relationships between the logarithmic word frequency score of Words B, C and E and the evaluation ratings were examined. Third, whether Word E occurred elsewhere in the same sentence was compared against the evaluation ratings.

To investigate which methods were related to which factors with high evaluation ratings, a series of chi-square tables were constructed. High chi-square scores in these tables were taken as being indicative of a strong possible relationship between the method and the factor in determining the evaluation rating.

4. RESULTS
4.1 Identifying Contexts for Choosing between the Six Methods
From the investigation of ways of correcting the 200 example sentences, the first finding was that evaluating suggestions based on the linguist’s judgment of the most appropriate suggestion was the most valid basis for evaluation. The other two bases for evaluation resulted in a high proportion of grammatically correct but semantically inappropriate corrections being deemed acceptable. Using this basis for evaluation only, it was found that the clearest distinctions between different aspects of the automatisable factors were found for part of speech information. The following discussion therefore examines the six methods by looking at the linguist’s evaluations of the most appropriate suggestion and the parts of speech of the words concerned.

For the 200 sentences, acceptably appropriate suggestions were generated for 141 sentences. Of these method B was the most common (for 46 sentences), with method C occurring rarely (for only 6 sentences). However, from this it cannot be concluded that method B should be the basis for generating suggesting for correcting word-word errors, since this method would only generate appropriate suggestions for 23% of all errors. Rather, the most appropriate method to use appeared to depend on the parts of speech of the words concerned.

For method B, of the 46 sentences where this method provided the most appropriate suggestion, in 24 of the sentences Word B was a verb, and in 29 sentences Word C was a preposition. These findings, however, again cannot be taken at face value since we need to know the overall frequency of verbs as Word B and prepositions as Word C. Therefore, taking the proportions of overall frequency as the expected values, chi-square values were calculated for each method compared with each part of speech. It was found that the part of speech of Word B was important only for method A, whereas the part of speech for Word C was important for 5 of the 6 methods and the part of speech of Word E for 2 of the 3 relevant methods. Identifying the matches between parts of speech and methods that were responsible for the high chi-square values allows algorithms to be written for automatic generation of appropriate suggestions for word-word errors. An example algorithm is:

Use method B (i.e. delete Word C) when the part of speech of Word C is a preposition.

In total, eight such algorithms were written, and these accounted for 85 of the original 200 sentences.

The goal of the grammar checker project is to produce a program that can identify at least 50% of local errors made in writing with under 10% overflagging (i.e. identifying correct language as an error) and that can give appropriate suggestions for correcting errors in most instances. For the 200 example sentences containing word-word miscollocations, the errors
were identified in over 95% of cases. Following the eight algorithms for generating corrections, appropriate suggestions were made for around 45% of the identified errors – a figure not quite reaching the target, but acceptable when considering that other aspects of the program achieve 90% appropriate suggestions. However, the procedures used for generating corrections need to be validated against a further set of sentences.

4.1 Validating the Algorithms for Generating Corrections
While the initial figures for evaluating the procedures for generating corrections were acceptable, their applicability to other word-word errors needs to be examined before they can be considered valid. Therefore, a further 100 example sentences containing word-word errors were collected from student writing and the algorithms for generating suggestions were evaluated based on a 3-point scale of appropriacy (where the linguist's judgments of most appropriate suggestion was rated 2, an appropriate (but not the most appropriate) suggestion was rated 1, and an inappropriate suggestion was rated 0).

The relevant errors were identified in 96 of the 100 sentences, and 88 of these sentences took forms that made it possible to generate suggestions by all methods. The sentences for which suggestions could not be generated were ones where the problematic point was at the start or end of the sentence. To generate suggestions following method E above requires z-scores for trigrams consisting of Word B, Word E (replacing Word C) and Word D. For sentences where, in the original, the last word is Word C, such trigrams cannot be generated and so suggestions cannot be made.

For the 88 sentences where suggestions could be generated by all methods, 78 suggestion sets following the algorithms taking parts of speech into account were generated, and 43 of these were deemed appropriate (with 14 being the most appropriate possible correction). Of the 35 inappropriate suggestion sets following the algorithms, 17, while not providing appropriate corrections, did not actually produce inaccurate sentences. However, for 18 of the 64 sentences, following the algorithms for generating suggestions would result in replacing the original error with another equally serious error. The overall appropriacy of the suggestions generated by following the algorithms and the 18 problematic sentences need to be examined in more detail.

For the suggestions generated following the algorithms, the mean level of appropriacy on the 3-point rating scale was calculated for each of the six methods. While 5 of the 6 methods resulted in a mean rating greater than 1, for method C (switching Words B and C) the mean rating was 0.31 for the 13 sentences where it was applied. This is clearly unsatisfactory and probably results from the fact that in the original 200 sentences method C was applied in only 6 cases, a number too low to allow for generalisable patterns to be identified. It was therefore decided to delete the algorithm for generating suggestions from method C.

The second issue concerned the 18 problematic sentences. 14 of these concerned using method D to generate suggestions. From the initial analysis of 200 sentences, three of the eight algorithms concerned applying method D and these algorithms were the most frequently applied in analysing the further 100 sentences. It therefore appeared that method D was being over-applied and additional context constraints on its application were needed. The original algorithms for applying method D concerned the part of speech of Word C. Comparing the appropriate suggestions derived from applying method D (involving 19 sentences) with the 18 problematic sentences, it appeared that the parts of speech of Words B
and C also needed to be taken into account. The algorithms for method D were therefore rewritten with an example of a resulting algorithm being as follows.

Use method D (i.e. replace Word B) when the part of speech of Word C is [a noun or a verb or a preposition] and [Word B is an adjective and Word A is an adverb].

Reapplying the refined algorithms to the 100 sentences did not affect the frequency of appropriate corrections, but reduced the number of sentences for which inappropriate and inaccurate suggestions would be given to 3 – an acceptable figure.

5. IMPLICATIONS

From this discussion of a small part of the whole project to create an effective grammar checker, it should be clear that automatising corrections for written errors is an extremely complex process. It is complex because of both the inherent complexity of the English language and the complexity and unpredictability of student errors. Despite this complexity, the grammar checker program produced from this project has reached its target of accurately identifying at least 50% of local errors with under 10% overflagging, and the part of the program devoted to word-word miscollocations has achieved a reasonable, if not high, rate of producing effective suggestions for errors. The target reached, however, is nowhere near as good as what a competent human evaluator can achieve. I believe that there are three main reasons for this shortfall in error identification and correction.

First, the grammar checker program does not take meaning into account in analysing input text. Rather, it focuses purely on surface forms and parts of speech, with the result that output text, while accurate, may be semantically inappropriate. This is an area where dictionaries perform more effectively than grammar checkers, especially for polysemous words. It would be interesting to examine the possibilities of integrating the meaning-distinguishing features of dictionaries with the accuracy-oriented aspects of grammar checkers in future work.

Second, even though the purpose of the grammar checker is to correct errors, in the project discussed here no consideration was made of corrected texts or the process of text correction. Rather, the identification of errors was based on low probabilities of co-occurrence of pairs of words in a model corpus, and correction of errors was based on evaluating potential corrections generated by following automated rules. An alternative approach of training a neural network using a corpus of teacher-corrected student work may be worth investigating.

Third, the data considered in this paper may not be sufficient given the probabilistic approach underlying the design of the grammar checker. A corpus of 100 million words may be the norm in linguistic studies these days (and is the size of the British National Corpus which is often used as a benchmark in corpus research), but it may be too small to provide reliable probabilities of use, especially for word-word-word co-occurrences. Similarly, the initial collection of 200 example sentences may, at first sight, appear sufficient for identifying patterns, but the mismatches of analysis between this collection and the further collection of 100 sentences suggests that this is not the case. Effective probabilistic analyses may require data sets an order of magnitude larger than those considered here. With present computers, this presents a problem in research situations where funds are not extensive. The analysis of word-word-word co-occurrences in the 100 million word corpus in this study required several weeks of continuous dedication of a server to the analysis. It seems likely that valid large-scale probabilistic analyses, especially those involving the training of neural networks, can
best be undertaken using grid computing, and thus may become restricted to a few dedicated localities.

The study reported in this paper may not be of direct relevance to current research into dictionaries, but I believe that it points to issues important for future lexicography. Historically, dictionaries have continuously expanded the number of linguistic aspects covered from an initial coverage of spelling and meaning only to current learner dictionaries which may include information on register, synonyms, antonyms and variants as well as a wide range of other aspects. If this trend of expanding coverage of linguistic aspects continues as seems likely with the growth of electronic dictionaries (Jackson, 2002), lexicographers in the future may need to learn from research in natural language processing such as that discussed in this paper.

6. REFERENCES


Abstract: Bilingual dictionaries between limited distribution languages are uncommon, in particular when the languages they cover are geographically and culturally remote. Even so, a few such dictionaries exist between Finnish and Thai. Initially is examined why an interest in Finnish–Thai dictionaries has evolved over the past two decades. The main reasons are the growth of tourism, trade and communications between Finland and Thailand, coupled with recent Thai immigration to Finland. The rest of the paper basically is concerned with the so far largest Finnish–Thai–Finnish dictionary, edited by S. Seppänen and W. Tookirii and published in 2006. Parameters such as distribution structure, macrostructure and microstructure are analysed, complete with a statistical survey of the Finnish–Thai part of this dictionary and a discussion on its lemmata selection and directionality. Finally a brief outlook on future prospects for Finnish–Thai–Finnish lexicography is given.

Key Words: Bilingual dictionaries; limited distribution languages; Finnish and Thai; present situation; future prospects.

1. INTRODUCTION

Bilingual dictionaries between two languages of somewhat equally limited distribution are unusual. As was pointed out by the present writer on an earlier occasion: “Bilingual dictionaries are mostly commercial products and thus subject to the constraints of feasibility and marketability” [Sundström 2005, 268]. In consequence, bilingual dictionaries between lesser-used languages are rarely found; even more so when the languages concerned are vastly separate, from a geographical and/or cultural point of view.

Against this background it may be assumed initially that even today there should be no bilingual dictionaries at all between Finnish and Thai. To begin with, Finnish assuredly is a language of limited distribution. In fact it stands out as a kind of curiosity among most European languages, belonging, as it does, not to the Indo-European language family but to the Finno-Ugric group of what is known inter alia as the Uralic languages (and also includes some Asian languages, in central Siberia). The majority national language of Finland it is spoken as a mother tongue by some five million people of that country, to which may be added (although numbers are notoriously inexact) a few hundred thousand in neighbouring Sweden, perhaps one hundred thousand in Russia and a few thousand in the very north of Norway. – In other words there are more than twice as many people living in the city of Bangkok than there are native speakers of Finnish in the whole world! Then, on a European scale, Finland is a young nation (attaining her independence only in 1917) and up until the beginning of the past century there were virtually no contacts between Finland and Asia. Indeed, Finland never had neither an East India trading company nor any colonies anywhere.

Also, in brief, the Finnish language for most of its history used to be relegated to a subservient status. Before attaining independence, Finland spent some six hundred years under Swedish rule, and so, Swedish was long the language of administration and higher
education. Finland's period as a Russian grand duchy (1809–1917) initially did little to change this situation. Only during the latter half of the nineteenth century did Finnish eventually start its rise to the position of major national language of Finland. The first few decades of independence saw a marked development also in the field of lexicography with bilingual dictionaries compiled between Finnish and most major European languages. To Asian languages, however, dictionary coverage was not extended.

2. CONTACTS BETWEEN FINLAND AND THAILAND: A SHORT SURVEY

On October 9 1919, the then Kingdom of Siam extended diplomatic recognition to Finland, actually the first Asian country to do so. Over the subsequent years, though, contacts remained minimal. In 1956, a Finnish diplomatic mission opened in Bangkok but again, the countries remained almost literally poles apart.

2.1 The nineteen seventies: two events heralding closer Finno-Thai interaction

In November 1976, the Finnish national airline company inaugurated non-stop flights from Helsinki to Bangkok. Certainly, flying to Thailand was out of economic reach for most Finns in those days, not to mention Thais coming to Finland. All the same, as regards travelling time, Helsinki was suddenly brought as close to Bangkok as to the town of Rovaniemi on the Arctic Circle by overnight train.

Two years later, in 1978, the Finnish Missionary Society (today the Finnish Evangelical Lutheran Mission, FELM), closely affiliated with the major national church (the Evangelical-Lutheran Church of Finland) took up missionary activities in Thailand. Through church services and various church sponsored activities thousands of Finns became acquainted with the word "Thailand", an acquaintance reinforced by illustrated missionary magazines, which in Saarelma's words "became 'important educators for internationalism' for the Finnish people" (Saarelma 2008, 49).

2.2 The opening-up era of lexicography between the Finnish language and Asian languages

The nineteen eighties saw the opening up of another phase in Finnish bilingual lexicography, with the broadening of dictionary scope to Asian languages. In those years work started on the two projects, destined to become the largest bilingual dictionaries between Finnish and two Asian languages. First came the “Suomi–Vietnam sanakirja” ['Finnish–Vietnamese dictionary'], published in 1997 by Laurent Tran-Nguyen and followed in 2002 by the companion volume “Vietnamilais-suomalainen yleissanakirja” ['Vietnamese–Finnish general dictionary'] by the same author. Both dictionaries contain over sixty thousand headwords and for several years they provided the largest bidirectional lexicographic coverage between Finnish and any Asian language. Why then precisely Vietnamese? The reason is, in the closing years of the past century, owing to refugee resettlement, there were several thousand Vietnamese living in Finland, forming by far the largest community of Asian origin in that country. Subsequently, in 2005, the Suomi–kiina suursanakirja, ['Finnish–Chinese General Dictionary'] came out in print, as the result of more than thirty years of work by Chinese scholar Li Guangyun, featuring eighty-five thousand headwords. The above works stand out as the incomparably largest Finnish–Asian languages dictionaries to this day.
2.3 The past fifteen years: Finnish presence in Thailand, Thai immigration to Finland

The past fifteen years saw not only the incipient steps of Finnish–Asian languages lexicography but also an exponential growth of prosperity and internalisation in Finland. Living standards went up by almost half in just about ten years' time and as is well known, a formerly obscure Finnish manufacturing company rose to become the world's biggest cellular phone producer. As a result, Finnish orientation towards the external world took a quantum leap forward, illustrated notably by Finland's accession to the European Union and the vastly expanded Finnish presence in the world's most dynamic powerhouse of economic development, i.e. Asia. This presence also appeared at grassroots level, in the shape of tourism. According to the Finnish embassy in Bangkok, tourism from Finland to Thailand these days annually exceeds one hundred thousand travellers.

The opening years of the 21st century saw yet another development that has so far attracted but little notice: a surprisingly strong influx of Thai immigrants to Finland. From some 1 700 people in 2003, the Thai community numbered some three thousand in 2006 and the tendency is still growing. In today's Finland there are about as many Thai nationals living as there are Britons and Germans. –Also, according to data from Finland's embassy to Thailand, an estimated one thousand Finnish nationals are living in Thailand at present. Investigating the reasons for these migrations, fascinating though it may be, would not, however, fall within the remit of this presentation.

With human contacts interlingual contacts followed. In 2002 the Finnish–Thai Association was formed, a structure whose programme includes Thai language courses for Finns.

All in all, this evolution created an enabling environment for Thai–Finnish lexicographical projects. The initial steps were taken as early as 1986 (incidentally, the year when Finland's diplomatic mission in Bangkok was upgraded to full embassy status) with the appearance of Jarmo Talasmaa's Suomi-thai matkaajan sanasto ['Finnish–Thai traveller's word list'], published in Helsinki 1986. It was followed ten years later by Suomi-thai matkaajan sanakirja ['Finnish–Thai traveller's dictionary'], edited by a person of Thai origin, Changklung Cebu, and containing seven thousand headwords. Finally, a bi-directional dictionary was published in 2004 by the Finn Veijo Hoikka and his Thai-born wife Punnea. For all their merits, however, these works were only modest beginnings. And so, the stage was set for the epoch-making work of Seppo Seppänen.

3. FINNISH-THAI-FINNISH DICTIONARIES TODAY

Today, by far the most comprehensive and authoritative Finnish–Thai bilingual dictionary is the “Suuri Suomi–Thai–Suomi”, meaning 'Large Finnish–Thai–Finnish [Dictionary]', which appeared in 2006, edited by a Finn, Mr Seppo Seppänen and co-authored by a Thai, Ms Wanidaa Tookiri. According to its back cover presentation, it contains 12,000 Finnish headwords and 10,500 Thai headwords. Three years later, in 2009, it was followed by a shorter version, the “Uusi Thai sanakirja” ['New Thai Dictionary'] by the same author. Mr Seppänen thus stands out as the incontestable authority in today's Finnish–Thai bilingual lexicography and from now on, this paper will focus on his work.
For brevity's sake, the “Suuri Suomi–Thai–Suomi” will be referred to below as the SSTS whereas the “Uusi Thai sanakirja” will be abbreviated as the UTS.

In discussing the SSTS we shall first proceed to an overview of that which the Swedish lexicographer and terminologist Svensén called dictionary distribution structure. To Svensén the distribution structure is primarily composed of three parts: the lemma list (i.e. the totality of headwords), then overview articles and also independent dictionary components either in the front matter or elsewhere in the dictionary, to which possibly a fourth component may be added: information external to the dictionary such as other dictionaries, manuals etc to which references are made in the dictionary (Svensén 2004, 98).

Below, first the distribution structure of the SSTS and then its macrostructure and microstructure will be discussed.

3.1 The SSTS: dictionary distribution structure

The SSTS consists of a double lemma list. First comes the Finnish–Thai part and then follows the Thai–Finnish part. The lemma lists are preceded by a surprisingly extensive front matter and also followed by a back matter of considerable length.

The front matter contains the following:

1. Instructions to the user (one page, in Finnish only).
2. Table of Thai characters with Latin alphabet equivalents and pronunciation guidance for speakers of Finnish.
3. A basic overview of Thai phonetics, including tonemes (five pages, in Finnish only).
4. Information about the technique for presenting illustrative phrases and about abbreviations used.
5. A page containing miscellaneous information, written in Finnish and displaying a very disjointed structure, containing as it does, comments on so-called naughty words included in the lemma list, followed, for no apparent reason, by a few lines about the proper way of using the wai greeting in a Thai cultural context and finally advice to foreigners wishing to practice spoken Thai.

The most surprising part, then, awaits the reader in the no fewer than fifty pages of back matter (in a dictionary totalling 347 pages). At first, there are passages resembling a miniature thesaurus, or rather a tourist phrase book, where the user is informed about words and phrases related to given themes and situations (furniture and interior decoration, shopping for clothes, toiletries etc), followed by an appendix about food and drinks with 262 items translated, mostly from Finnish but some also only from English into Thai. – In passing could be noted that on two occasions, no equivalent has been provided, as is the case with the Finnish fish name *kuha* ['pike-perch'] and the English *cod*, probably owing to a pure oversight during the editing process. Then follows a trilingual phrase list for visits to health professionals, with headwords in English, followed by translations, first into Finnish and then into Thai and here, exceptionally, the Thai language is written both in Latin and Thai letters. After that are found thirty-two Finnish food recipes translated into Thai (sic!) and finally, two pages intended for Finnish readers and containing information about numerals and time expressions in Thai.

The distribution structure, in sum, amply testifies to a certain multi-tasking by the SSTS. Parallel to providing information about Finnish and Thai lexical items and their respective
translations, this dictionary also assumes the roles of phrase book, source of culture-related information and even cook book! Useful though this might certainly be for the tourist user, a clearer delimitation of what the dictionary truly attempts to be might well have opened up possibilities for considerably enlarging the lemma lists, without changing the size of the book.

3.2 The SSTS: dictionary macrostructure

In discussing the macrostructure (and also the microstructure) attention will be focused only on the Finnish–Thai part of the dictionary, owing to the present author's total unfamiliarity with the Thai language. – Owing to technical difficulties, the Thai lexical items will be represented only in the Latin alphabet.

In the Finnish–Thai part of the dictionary, the articles are arranged in two columns per page, the headwords appearing in boldface type. In the Thai–Finnish part, only the Thai alphabet is used for the Thai headwords and the articles are arranged in three columns per page.

Among the Finnish headwords, the explanatory metalanguage, seldom used, is almost invariably Finnish (never Thai). The articles are interspersed with nineteen photographic illustrations, sometimes directly relating to the text, sometimes seeming to be the fruits of random selection. To give just one example: next to the Finnish adjective harmaa, rendered into Thai as sii-thau is enclosed a photo of the dictionary editor, a greying man, and the young female co-editor, turning their backs to the camera.

On the subject of dictionary macrostructure, a striking tendency of blurring the border lines between macrostructure and microstructure can be noticed throughout the Finnish–Thai part of the SSTS. A graphic illustration is provided by the treatment of Finnish verbs like olla ['to be' or 'to have'] and tulla ['to come', 'to become']. Instead of grouping together the various semantic fields as moments forming part of the same article, the author has preferred listing them as separate headwords, as is shown below. Thus, for instance, such an essential Finnish verb as olla is distributed over a full twenty articles

This tendency shall be further illustrated below with examples containing the Finnish verb tulla. In the illustrations below, the Thai equivalents to the Finnish headwords are shown in the same typographical representation (boldface, hyphens) as they appear in the SSTS. – The comments in square brackets are my own and do not form part of the dictionary text.

**tulla** ['to come', 'to become'] maa

**tulla** johonkin ['arrive somewhere'] maa-thyng

**tulla** joksikin ['turn into something'] klaai-pen

**tulla ulos** ['come out'] ook

(The Finnish explanatory metalanguage notes johonkin and joksikin mean 'somewhere' respectively 'something' in English).

---

1 Uralic languages, such as Finnish, are well known for using the same verb both to mean *to be* and *to have.*
In the above examples the metalanguage is characterised by a marked lack of consistency. Sometimes double quotation marks are used; sometimes single quotation marks, and once, the metalinguistic comments are even made in English rather than in Finnish. Again, arguably, these different articles could very well have been joined into one.

Yet another example is provided by the Finnish adjective **paha** ['bad', 'evil'], which is presented in the following manner:

```
paha _hoot+raai
paha + raii
paha leeu
```

(The plusses and minuses refer to various tonal aspects of the Thai equivalents.)

This may be informative to the Thai reader who knows beforehand which different semantic aspects are covered by the different Thai equivalents, but the Finnish reader is left at a loss with respect to which Thai equivalent to choose.

All in all, the tendency to split up the same headword over several articles is dubious at best. Access at first glance may improve, but from a lexicographic point of view it is hardly satisfactory.

### 3.3 The SSTS: dictionary microstructure

As for microstructure, the SSTS almost exclusively features what may be called a Minimal Article Structure; MAS for short. By MAS is meant the barest minimum structure of an article in a bilingual dictionary, i.e. an article consisting of a headword followed by one single target language equivalent (whether in the shape of one or several words, but without labels and/or explanatory metalanguage). This is true notably of the latter section, with headwords in Thai and equivalents in Finnish. Such an approach is understandable, notably against the background of space constraints in a dictionary of fairly limited size. Even so, the practice lends itself to some questioning, in view of this statement by P. Sharpe: “For language pairs that do not share the same cultural traditions the notion of one-to-one equivalence is likely to become less and less reliable” (Sharpe 1989, 315). – Very few articles in the Finnish–Thai part of the SSTS, however, give even two equivalents in Thai for a given Finnish headword and the largest number of Thai equivalents ever attested was three (only once, in conjunction with the Finnish headword **varovainen** ['careful']).

One of the most peculiar features of the SSTS microstructure is the widespread tendency to create what I have chosen to call multiple headword units. In them, several Finnish lexical items are grouped together to form, as it were, one single headword. Consider the following example

```
järjestää, asettaa, asentaa, pystyttää, perustaa tang
```

Most probably, the intention is to show that the Thai verb **tang** can function as an equivalent to all of the five Finnish verbs constituting the multiple headword unit. In English **järjestää** means 'organise', **asettaa** means 'establish', **asentaa** means 'install', **pystyttää** means 'set up', 'build' and **perustaa** 'found', 'lay the foundation of'. Now, for a Thai dictionary user it is certainly important to learn that the Thai verb **tang** may take at
least five different equivalents in Finnish, depending on its shades of meaning, but surely this is the kind of information one would expect in the Thai–Finnish lemma list and not here? And generally speaking, it is obviously not common lexicographic practice to structure dictionary headwords like this.

What we have is in fact a kind of directionality role confusion. The vagueness of a given Thai lexical item is reflected, not by including this Thai lexical item in a number of different articles, as an equivalent of several Finnish different lexical items, but rather by lumping together into one multiple headword those Finnish lexical items that could take this Thai lexical items as an equivalent. – In this context, the notion of vagueness is defined, in the words of inter alia the Thai linguist K. Thepkanjana, as “a case in which non-distinct meanings are associated with a phonological form” (Thepkanjana & Uehara 2005, 309).

Sometimes, even lexemes with distinct meanings are brought together to form a multiple-headword unit:

käyttää, tehdä, lähettää, suorittaa, maksaa +tzai

Here the Finnish verbs forming part of the MHU mean respectively 'use', 'do/make', 'send', 'carry out' and 'pay'. Thus the evidently polysemous nature of the Thai equivalent is illustrated by grouping together the Finnish lexical items generating this equivalent into one multiple headword unit!

Examples such as this one abound in the Finnish lemma list. As will be shown in the statistics below, almost a fourth of the Finnish headwords display this structure.

Also, there is a tendency to mix metalanguage and headwords, in other words letting metalinguistic explanations form part of the headwords themselves. Cf below:

uskomaton, epäiltävä, saa nähdä selviyttyykö tehtävästä  naa+phit_sa-wong

Here uskomaton means 'unbelievable', epäiltävä means 'suspicious' whereas saa nähdä selviyttyykö tehtävästä is a verbal phrase meaning 'remains to be seen if he can handle the task'. This practice in fact illustrates H. Adhikari's remark about dictionaries committing errors in letting “non-lexical units figure as main entries” (Adhikari 2007, 52), as does the article below:

huoltomies, joka paikan höylä peruskoulussa phaan-rong

Here huoltomies means 'maintenance man' whereas joka paikan höylä peruskoulussa is a nominal phrase meaning 'a Jack of all trades at comprehensive school'. – The exact meaning of this phrase remains uncertain to the present writer.

Some further microstructural aspects lead us to conclude that a fine-tuning touch would no doubt have been to the advantage to the end product. Again, reference may be made to the use of metalanguage. Consider the case of kuusi, with its Thai equivalent _hok. Now, there are two homonymous words kuusi in Finland, the one referring to a fir tree, the other being the numeral six. However, there is only one article kuusi in the Finnish lemma list, with no explanatory metalanguage.
Another example can be seen below, with the homonyms rikki ('broken' as an adjective, 'sulphur' as a noun):

riikki, murtunut _hak
riikki, irti _tääk_ook
riikki (aine) kam-ma-than

Here, the two homonyms form separate articles. What is less comprehensible, though, is why the adjective rikki should occur in two articles, depending on certain minor differences in its shades of meaning. Also, it may again objected that the placing of the Finnish metalanguage murtunut ['broken'] and irti ['loose'] is potentially confusing, as the Thai user may be mislead into thinking they form a part of the Finnish headword itself. Arguably also, the metalinguistic comment (aine) ['element'] could preferably have been reproduced in smaller print, as is usually done with metalanguage in the SSTS.

3.4 A statistical survey of the Finnish–Thai lemma list in the SSTS

For statistical purposes, all articles on every page in the Finnish–Thai lemma list were counted. Then the percentage of articles with a minimal article structure (MAS, for short) was calculated per page, along with the percentage of multiple headword units (MHU, for short) and the number of illustrative phrases per page.

<table>
<thead>
<tr>
<th>Pages</th>
<th>Alphabet interval</th>
<th>Number of articles</th>
<th>Percentage of MAS articles</th>
<th>Percentage of MHU articles</th>
<th>Illustrative phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>17–26</td>
<td>aa–au</td>
<td>452</td>
<td>74.1</td>
<td>21.9</td>
<td>24</td>
</tr>
<tr>
<td>27–36</td>
<td>au–fa</td>
<td>472</td>
<td>85.9</td>
<td>28.8</td>
<td>9</td>
</tr>
<tr>
<td>37–46</td>
<td>fa–hu</td>
<td>491</td>
<td>89.0</td>
<td>24.4</td>
<td>10</td>
</tr>
<tr>
<td>47–56</td>
<td>hu–ja</td>
<td>482</td>
<td>84.2</td>
<td>22.4</td>
<td>15</td>
</tr>
<tr>
<td>57–66</td>
<td>ja–ka</td>
<td>500</td>
<td>78.4</td>
<td>28.0</td>
<td>13</td>
</tr>
<tr>
<td>67–76</td>
<td>ka–ko</td>
<td>522</td>
<td>84.5</td>
<td>19.9</td>
<td>10</td>
</tr>
<tr>
<td>77–86</td>
<td>ko–ky</td>
<td>513</td>
<td>91.6</td>
<td>27.7</td>
<td>9</td>
</tr>
<tr>
<td>87–96</td>
<td>ky–lo</td>
<td>550</td>
<td>86.2</td>
<td>22.6</td>
<td>5</td>
</tr>
<tr>
<td>97–106</td>
<td>lo–me</td>
<td>497</td>
<td>90.1</td>
<td>22.3</td>
<td>11</td>
</tr>
<tr>
<td>107–116</td>
<td>me–ni</td>
<td>491</td>
<td>88.8</td>
<td>29.3</td>
<td>12</td>
</tr>
<tr>
<td>117–126</td>
<td>ni–pa</td>
<td>526</td>
<td>85.9</td>
<td>24.0</td>
<td>17</td>
</tr>
<tr>
<td>127–136</td>
<td>pa–po</td>
<td>561</td>
<td>82.0</td>
<td>23.9</td>
<td>7</td>
</tr>
<tr>
<td>137–146</td>
<td>po–re</td>
<td>500</td>
<td>84.0</td>
<td>24.8</td>
<td>12</td>
</tr>
<tr>
<td>147–156</td>
<td>re–si</td>
<td>530</td>
<td>83.8</td>
<td>24.7</td>
<td>8</td>
</tr>
<tr>
<td>157–166</td>
<td>si–sä</td>
<td>489</td>
<td>84.1</td>
<td>22.9</td>
<td>12</td>
</tr>
<tr>
<td>167–176</td>
<td>sä–to</td>
<td>528</td>
<td>83.1</td>
<td>29.6</td>
<td>11</td>
</tr>
<tr>
<td>177–186</td>
<td>to–us</td>
<td>500</td>
<td>81.8</td>
<td>25.6</td>
<td>17</td>
</tr>
<tr>
<td>187–196</td>
<td>us–va</td>
<td>466</td>
<td>79.4</td>
<td>29.2</td>
<td>12</td>
</tr>
<tr>
<td>197–206</td>
<td>va–yh</td>
<td>525</td>
<td>85.7</td>
<td>28.6</td>
<td>5</td>
</tr>
<tr>
<td>206–211</td>
<td>yh–öl</td>
<td>184</td>
<td>78.3</td>
<td>5.4</td>
<td>10</td>
</tr>
</tbody>
</table>

All in all, the Finnish–Thai lemma list featured 9 779 articles. The average percentage of MAS articles for this lemma list as a whole was 84.0 per cent. In other words, only 16.0 per cent of the articles in the Finnish–Thai part, on average, featured anything more than one Thai equivalent to the Finnish headword (such as, either metalinguistic comments,
style labels or several equivalents). The peculiar MHU structure was present in a full 24.3 per cent of the articles. – This may in part explain the discrepancy between the 9,779 articles counted and the 12,000 given as a number for the Finnish headwords: in MHU:s the individual headwords may have been counted separately. Finally, there were 229 illustrative phrases.

3.5 Directionality and lemmata selection

At this point, attention will be focused on the problem of dictionary directionality. In the instructions to the user, it is pointed out that the dictionary is a service to the Finns spending their winters among Thai people and particularly to Thai persons who do not always have an easy time adjusting to society and family life in Finland (Seppänen 2006, 7). Thus, the dictionary clearly comes out in favour of being bidirectional. – Regrettably, however, these intentions conflict with lexicographic realities. As was pointed out by M. Ahmad “one source of trouble in most existing bilingual dictionaries is that they try to meet the needs of both [source language] and [target language] speakers at the same time. As the needs of these two kinds of users are different, it is impossible to pay equal attention to both in one and the same volume” (Ahmad 2007, 93). This opinion is further reinforced by K. Nakao, who unequivocally stresses that bidirectionality is even more impossible in dictionaries “in which one language is a Western and the other a non-Western language” (Nakao 1989, 296).

Ultimately, as regards directionality, the SSTS is manifestly geared to Finnish-speaking users. The focus on Finland and things Finnish can be seen by the somewhat trivial fact that there are no fewer than five headwords relating to lumi ['snow'] and the names of common trees in Finland like mänty ['pine'] and koivu ['birch'] are found among the lemmata, as are also words with such a distinctly northern flavour as hiihtää ['go cross-country skiing'], kelkka ['toboggan'], suket ['ski'] (although for reasons unknown, presented in the plural rather than in the singular suksi). Further evidence is given by headwords such as hirvi ['elk', alternatively, in American English, 'moose'] and mustikka ['blueberry']. We even find suomuurain, ['cloudberry'], a berry growing in subarctic regions only; in Newfoundland/Labrador (one of the few English-speaking areas where they are found) known as bakeapple2. – Generally, the lemmata selection focuses noticeably on everyday life, with few technical terms and even fewer words relating to law, science etc – understandably enough for a dictionary of such a relatively modest size. For instance, such buzzwords in public debate these days as ilmastonmuutos ['climate change'] or kasvihuonelimö ['greenhouse effect'] are not present among the headwords.

Otherwise the modernity of the lemmata chosen is amply illustrated by Finnish headwords such as digitaalikamera ['digital camera'], kännikkä ['cellular telephone'], mikrouuni ['microwave oven'] and tietokone ['computer']. The only obsolescent Finnish headword attested was lähetysaarnaaja for ['missionary (worker)'] which in present-day Finnish is replaced by lähetystööntekijä.

As a curiosity may finally be mentioned the comparatively wide range, again for a dictionary of this size, of lexemes relating to human sexuality.

---

2 The Thai translation given is ton+maai_dang_klaau. Perhaps this is a neologism coined by Thai seasonal workers employed in berry-picking in northern Finland?

In early 2009 the most recent development in Finnish–Thai lexicography was published, the UTS. At this juncture, a kind of miniature paradigm shift could be observed. Smaller than the SSTS, the UTS concentrates on its primary role as a dictionary, as opposed to its secondary roles of phrasebook, traveller’s handbook or even encyclopaedia. This development is underlined by its markedly improved access structure: the Thai headwords are presented, not only in Thai script but also in Latin letters. A further economy of space is achieved by omitting all photographic illustrations.

The number of headwords in the UTS is not explicitly indicated, but on the basis of a statistical sample, containing ten per cent of its pages, should amount to an estimated five thousand in either direction.

4. FUTURE PROSPECTS FOR FINNISH-THAI LEXICOGRAPHY

The dictionaries edited chiefly by S. Seppänen will certainly form the cornerstone of Finnish–Thai lexicography for several years to come. Few indeed are Finnish nationals, who like Seppänen, have attained a mastery of Thai and fewer still those who have tried their hand at lexicography between Finnish and Thai. Now the time appears ripe for an expanded product. In view of the ever increasing contacts between Finland and Thailand and the status of Thai as a vehicle for culture we would need a bilingual Finnish–Thai dictionary of a size similar to existing dictionaries between Finnish and such Asian languages as Chinese and Vietnamese (cf above). A dictionary of this kind might also spur interest in academic study of Thai in Finland, a language in which courses are not offered for the time being at Helsinki university (Colliander 2009), and thus indeed promote the implementation of the conference motto “Dictionaries in Education”. Here indeed is a venue for future cooperation between Finland and the Kingdom of Thailand and on a plea for such cooperation I should like to conclude my presentation.

5. REFERENCES


(3) Colliander, Maria. (2009) E-mail message sent in response to author's inquiry about the status of Thai at Helsinki University today.


DISCUSSING EXAMPLES IN LEARNER’S DICITONARIES
FOR FOREIGN LANGUAGE PRODUCTION

Mei Xue
University of Århus, DENMARK
meix@asb.dk

Abstract: Modern functional theory of lexicography emphasizes the distinction between reception and production and centralizes the specific user’s needs arising in specific user situations. Admittedly, foreign language learners’ needs for lexicographic assistance will vary depending on the learning contexts they are involved in and the data inside dictionaries are assumed to adapt to such specific needs. This paper is intended to have a close look at one type of data—examples in learner’s dictionaries. The aim of this paper is to emphasize the value of the data on examples for foreign language learners in production and explore what characterizes examples to facilitate foreign language production. This explorative study will focus on two aspects: first, the language difficulty characterizing Chinese undergraduates in English writing will be identified based on the analysis of typical errors made in their compositions and their specific needs for assistance from examples in dictionary. Second, qualitative research will be made on examples of some headwords in three dictionaries, namely, Oxford Advanced Learner’s Dictionary of English (7th), Longman Dictionary of Contemporary English (4th) and Longman Language Activator (2nd). The first two dictionaries are chosen for they are the most popular learner’s dictionaries with Chinese learners of English, and the last one for it’s production-oriented. The qualitative research and analysis is based on functional theory of lexicography.

Key Words: functional theory, example, production

1. INTRODUCTION

Examples have been an integral part of learner’s dictionary since the early efforts of Michael West and A.S. Hornby in the 1930s and 40s. When Hausmann and Gorbahn (1989:45) state that

“Any learner’s dictionary stands or falls by the number and quality of its examples, for it is the examples above all which help the foreign learner to understand a word (decoding function), learn it (learning function) and use it (encoding function).”

They speak out the partial truth concerning the relationship between examples and learner’s dictionaries, as examples are assumed to fulfill various functions, namely, “attestation of the existence of a word and a sense, elucidating the meaning of the word, illustrating the contextual features, such as syntax, collocation, register, etc” (Atkins & Rundell, 2008: 453-454). The understanding about the functions of examples have been discussed and classified more or less in the same way by some researchers (Cowie, 1989; Creamer, 987; Drysdale, 1987; Laufer, 1992; Humblé, 2001; Rundell, 1998; Xu, 2006). However, the assertion of any learner’s dictionaries does not necessarily depend on one type of data inside, but to what extent the types of data inside the dictionary fulfills the intended functions of the very dictionary through satisfying the targeted users’ needs arising in the particular user situations.
Despite the acknowledged importance of examples in learner’s dictionary, the relevant research is relatively rare and sporadic, only slightly dabbling the functions, types and properties of examples, until the advent of corpora in 1980s gives rise to the discussion of respective merits and demerits of authentic and made-up examples (Atkins & Rundell, 2008; Drysdale, 1987; Fox, 1987; Hausmann & Gorbahn, 1989; Humblé, 1998 & 2001; Laufer, 1992 & 1993; Nesi, 2000; Rundell, 1998; Sinclair, 1987; Xu, 2006 & 2008). Hausmann and Gorbahn (1989:45) state examples function differently as assistance for foreign language learners “decoding, learning, and encoding”, but they do not explain what difference existing between learning and decoding as well encoding. Some studies confirm the positive and useful feature of examples for production purpose as examples are assumed to demonstrate the typical behavior of words in typical contexts (Cowie, 1989; Drysdale, 1987; Laufer, 1993; Marello, 1987; Rundell, 1998). Other studies (Laufer, 1992; Humblé, 1998) have explored the use of authentic and made-up examples for foreign learners in production, and come to the expected conclusion that both types of examples maybe pedagogically beneficial without clear delimitations. Laufer’s study (1992) indicates that the lexicographer-made examples are pedagogically more beneficial than the authentic ones in comprehension and production of new words. Humblé’s study (1998) points out the interrelations between examples and learners’ proficiency level of foreign language without proposing ways to define learners’ proficiency level. Despite the limitations and respective focuses, these studies show positive values of examples in learner’s dictionary for production purpose either implicitly or explicitly, as Laufer (1992:71) states “a correct and natural use of a word in a sentence, or several sentences, will necessarily bring out the grammatical, semantic, pragmatic and collocational characteristics of the word”.

The enlightening study on example by Humblé (2001) to some extent echoes Hausmann and Gorbahn’s (1989) claim that dictionary examples can function for different purposes, and goes further to discuss criteria guiding examples intended for encoding (production) purpose, challenging the possibility of “one example to satisfy learners’ divergent needs and serve different types of users” (2001:62). However, this study is too ambitious to discuss examples in monolingual as well as bilingual dictionaries used by learners with different language and culture backgrounds. Furthermore, the profile of the learners is rather vague and has been categorized arbitrarily into beginners and advanced learners without any defining parameters. As learning takes place in certain socio-economic context, learners’ needs for information in examples are to be shaped by their native language and background culture. Few researches on examples available in English literature explore the specific needs of learners with specific language and cultural background. Xu (2006:154-158) attempts to fill this lacuna by investigating the reference needs a specific group of dictionary users (e.g. Chinese EFL learners) have for examples and the effective methods of identifying the needs, based on questionnaires. This study identifies the category of information that Chinese EFL learners expect to extract from illustrative examples, and claims that a questionnaire survey is a useful approach to investigating users’ explicit needs for illustrative examples (ibid:172). However, the conclusion is unconvincing as the study fails to indicate in which situation Chinese EFL learners have needs for the specific types of information conveyed in examples and it is also doubtful that learners can explicitly define their needs for certain types of illustrations without regard to the specific situation in which their needs arise.

These studies shed some lights on the exemplification in learner’s dictionaries, but still cherish a lot of hope on lexicographer’s craftsmanship, which are partially attributed to the
fact that they lack of the guidance from lexicographic theories in real sense. Modern functional theory of lexicography (Bergenholtz, 2003; Bergenholtz&Tarp, 1995, 2003; Tarp, 2004a, 2004b, 2008, 2009) establishes theoretic framework for lexicography and lays down general principles to guide the lexicographic practice by emphasizing the real function of dictionary and centralizing dictionary users in the process of dictionary consultation. Considering the fact that dictionary is a utility tool, the types of data inside dictionary are to actualize the function of such a practical tool. It goes without saying that dictionary examples are to realize the intended function of dictionary determined by the certain types of users and their specific needs in specific situations. The types of data and their presentation inside dictionaries are subject to the specific user’s needs in particular situations. No exception will go to dictionary examples. This paper is intended to have a close look at one type of data—examples in learner’s dictionaries with attempt to explore value of examples for foreign language learners in production and what characterizes examples to facilitate foreign language production. This explorative study will focus on two aspects: first, the language difficulty characterizing Chinese undergraduates in English writing will be identified based on the analysis of typical errors made in their compositions and their specific needs for assistance from dictionary examples. Second, qualitative research will be made on examples of some headwords in three dictionaries, namely, the seventh edition of Oxford Advanced Learner’s Dictionary of English (OALD7), the fourth edition of Longman Dictionary of Contemporary English (LDOCE4) and the second edition of Longman Language Activator (LLA2). The first two dictionaries are chosen for they are the most popular learner’s dictionaries with Chinese learners of English, and the last one for it’s production-oriented. The qualitative research and analysis is based on functional theory of lexicography.

2. CHINESE LEARNERS’ LEXICOGRAPHIC NEEDS IN ENGLISH WRITING

Two kinds of needs have been clearly defined for learners in text production of foreign language within the framework of functional theory: function-related and usage-related needs (Tarp, 2008: 152). Their function-related needs are to be realized by different types of data inside the dictionary while their usage-related needs are to be actualized by the structure of the data. Examples are a type of data necessary for learner’s in text production of foreign language, but its usefulness are to be defined by the specific users in concrete situation in terms of information accessibility and retrieval at minimum information cost. Therefore, it is essential to achieve understanding of the specific group of dictionary users before discussing the criteria of selecting and presenting examples in learner’s dictionaries. Chinese learners’ needs for example in English writing will be shaped by the socio-economic contexts in which the foreign language (i.e. English) learning takes place. The most salient variables influencing Chinese learners in English writing are their mother tongue (L1), prior knowledge in L1 culture, their foreign language (L2) proficiency level, and their learning contexts (Wen&Johnson, 1997; Wang & Wen, 2002;), which are to characterize their particular needs for information conveyed through examples to facilitate their English writing.

In the mainland of China, English learning is compulsory in curriculum from secondary to tertiary level and in cities, starting from elementary school. When reaching undergraduate level, most Chinese learners have been studying English for at least 6 years with age ranging from 17 to 20, and are expected to continue English study in classroom for the first two years of their university life. Despite the wide range of varieties of their dialects, they speak fluent Putonghua (Mandarin Chinese) and share the same writing system. According to the syllabus
set by China Ministry of Education, these post-secondary learners are supposed to have command of 3,300 core vocabulary and be able to produce a piece of English writing at least 100 words within 30 minutes. However, it should be noted that English is a foreign language only consciously learned and used in formal educational contexts, and learners’ exposure to target language is rare. Chinese learners mainly assimilate the knowledge about vocabulary and grammar of English in textbooks by memorization. So, Chinese learners’ skills in using English are developed after their assimilation of certain knowledge of English vocabulary and grammar, and lag behind their knowledge about English lexis and grammar. In English writing, they may refer to Chinese to formulate their ideas and figure out the lexical equivalents and possible sentence structures when they have difficulties to express themselves directly in English (Wang & Wen, 2002).

The drastic differences between Chinese and English inevitably result in salient negative transfer in English writing by Chinese learners, as text production demands heavier loading pressure in cognition compared with that in comprehension. Language learning takes place within certain social-economical contexts, and the distance between source language and target language is sure to leave imprint on foreign language learners (Cortazzi & Jin, 1996: 169). Actually, a lot of researches have confirmed the “existence of negative transfer in from Chinese to English at the phonological, lexical, syntactic, semantic or discoursal level” (Chan, 2004:57). English, as most Indo-European languages, matches verb conjugations according to subject pronoun and time reference (i.e., present, past, and future). In contrast, Chinese does not differentiate verb forms regarding the number of subjects, tenses, mood and voices and uses adverbs (e.g., always, yesterday) to identify time frame. The almost uninflected nature of CHinese leads to several “errors” in verb conjugations in Chinese EFL learners’ writing with reference to person, number, part of speech, tense, mood, agency, and the copula. It is common for Chinese learners to produce “I very happy,” instead of "I am very happy"; “He feel very happy” instead of “He feels very happy”; “He visit his parents last week” instead of “He visited his parents last week”. Semantic inflections (i.e. prefix and suffix) are also absent in Chinese. For instance, xingfu (幸福) in Chinese correspond to different parts of speech of happy in English. Moreover, the conceptual understanding of countability of nouns and transitivity of verbs are different in Chinese and English. Chinese learners have to memorize uncountable nouns and intransitive verbs as well as their specific usage. Furthermore, the absence of finite and infinite article system in Chinese makes Chinese learners uncertain or unaware of the presence of articles in English writing. Besides the difficulties caused by negative transfer from L1, Chinese learners’ incompetence in active use of their knowledge about the foreign language aggravate their difficulties in writing.

It is beyond this paper to discuss the contrastive rhetoric between Chinese and English, but examples generally are assumed to be effective ways to raise Chinese learners’ consciousness of these syntactic constrains to help them express their ideas accurately and appropriately in writing, when they resort to examples for the usage of words, including its semantic lexical, collocational, syntactic and pragmatic properties in different contexts (Huang, 2006; Huang, 2005; Xu, 2006; Xu & Luo, 2005; Yang, 2001). More importantly, all the information intended to be conveyed through examples is to be accessed, retrieved and reapplied with minimum efforts by Chinese learners. Thus, the required information in examples is to be explicative and explicit for Chinese learners to access and retrieve easily. Accessibility here is to be understood at two layers: the example itself is comprehensible and the required information is explicative. That is, the examples must be comprehensible to Chinese learners.
before they can extract the required information to complete their production tasks. On the other hand, the information concerning the usage of the word shown in examples is indicated implicitly for retrieval. Considering the profile of Chinese learners under this study, the vocabulary in examples should be controlled and context is transparent for Chinese learners to identify and interpret. For those examples greatly shaped by English culture, usage notes should be provided for quick access. Skeleton examples and typography can also be used for easy identification of the important usage information. The following will make a qualitative analysis of some example articles from learner’s dictionaries to examine the possible lexicographic assistance Chinese learners in text production can obtain in terms of information accessibility and retrieval.

3. QUALITATIVE ANALYSIS OF THE EXAMPLE FOR PRODUCTION PURPOSE

It goes without saying that dictionary users appreciate examples and examples are helpful in text production. The criteria of selecting examples and policies to make examples for dictionaries have been proposed (Atkins & Rundell, 2008; Cowie, 1989; Humblé, 2001; Xu, 2008). When Atkins and Rundell claim “[t]he nature of examples will vary according to the types of dictionary and the needs and expectations of its users”, they are disrupting the interdependence between the function of dictionary and users’ needs and expectations, as examples in their understanding are assumed to negotiate sort of balance between types of dictionary and users’ needs. Actually, the type of dictionary is determined by its intended functions which are actualized by different types of data (e.g. examples) inside the dictionary, and the function of dictionary is to satisfy the needs of certain types of users’ in certain types of situations. Therefore, the principles guiding selection and presentation of the types of data are to be in accordance with the function of the dictionary. Examples are expected to act either independently or conjunctively with other types of data to help realize the function of the dictionary, and their efficacy is to be defined in terms of accessibility and retrieval.

The entries for suggest from OALD7, LDOCE 4, and LLA2 will be examined carefully considering the research topic in this paper. Entry suggest is chosen for the fact that it is very unlikely that Chinese learners at tertiary level will look up suggest in text reception, as suggest is among the core vocabulary of English curriculum at junior school. Examples in articles from OALD7 and LDOCE4 are classified according to the sub-senses of the main entry. For the convenience of discussion, examples will be numbered in accordance with the sub-senses ordered in the respective dictionary entries. If there is more than one example under the same sub-sense, the example will be marked with letters alphabetically. For instance, 1a and 1b indicate that example a and b are listed under the same sub-sense 1 in the entry.

In the article suggest in OALD7, 16 examples sentences are presented and grouped according to the different contextual meanings and show the possible syntactic structure in the grammatical formulas set before examples, such as suggest sth (to sb), suggest sb/sth (for sth), etc. Examples 1c, 1e and 1f shows that suggest could be used in the structure “suggest + that” either in active or passive voice, and the connective that should not be omitted in that-clause in passive voice. However, they fail to convey one important message that when suggest acquires the contextual meaning similar to propose and is followed by that-clause, the verb in that-clause should take canonical form regardless of the tenses in the main clause. The difference shown in examples 1e and 1f is negligible for Chinese learners, since both syntactic structures are acceptable, and extracting the necessary information quickly to be
used in the immediate productive situations means more. To bracket the modal verb *should* will be more explicative. If the dictionary intends to help Chinese learners in production, the grammatical constraints should be explicitly indicated in the examples like *He suggested we (should) go out to eat. He suggested that the lecture (should) be given in English.* English syntactic structure is salient to convey different contextual meanings; while grammar gives way to meaning intended in Chinese. Below the sub-sense *recommend*, example 2d explicitly shows constraint on syntactic structure *suggest sb sth*. No substantial difference is shown in examples 2b and 2c, as both show *suggest* can be used in the structure *animate subject + verb+ inanimate object*. The grammatical code *wh* before example 2e is implicit, but the example 2e shows one possibility leaving other possibility (e.g. *how, where*, etc.) uncertain for learners. For native speakers, such combination of code and single example is sufficient, while learners are unconscious or uncertain of the grammatical extensions. In a word, the examples in article *suggest* in *OALD7* show the possible usage of *suggest*, but not the constraints except in example 2d. These examples cannot prevent Chinese learners produce sentences like *He suggested to go by plane. He suggested us to leave earlier. He suggested that we went out for walking*. Furthermore, examples 1a, 2a, 2c, 3a and 4a are in question forms and of oral style, which means more in reception and cognitive learning than production.

1a. May I suggest a white wine with this dish, sir?
1b. A solution immediately **suggested itself to me**.
1c. I suggest (that) we go out to eat.
1d. I suggested going in my car.
1e. It has been suggested that bright children take their exams early.
1f. (BrE) It has been suggested that bright children should take their exams early.
2a. Who would you suggest for the job?
2b. She suggested Paris as a good place for the conference.
2c. Can you suggest a good *dictionary*?
2d. Can you suggest me a good *dictionary*?
2e. Can you suggest how I might contact him?
3a. All the evidence suggests (that) he stole the money.
3b. The symptoms suggest a minor heart attack.
3c. What do these results suggest to you?
4a. Are you suggesting (that) I’m lazy?
4b. I would never suggest such a thing.

(Figure 1: article example from *OALD7*)

In *LDOCE4*, 15 examples are grouped according to subdivision of senses and 4 examples are put in a grammar box to indicate the syntactic constrains. Like examples in *OALD7*, example 1c, 1d and 1g demonstrate the usage of *suggest that* in active and passive voice, but the difference between examples 1c and 1g is superfluous, as they are more appropriate in showing different usage of word *before* rather than *suggest*. The same goes true with example 1a in which *keep doing* is as salient as collocation *suggest ways*. Example 1h and 6a itself is obscure to be understood because of the insufficiency of contextual information, and is more useful in text reception than in production. If they are presented for productive use, Chinese translation is necessary for their comprehensibility. Example 3a shows that *suggest* can be used in passive voice with an animate subject, but does not indicate whether it is acceptable to say *sb is suggested to do sth*. Example 1d in *LDOCE4* conflicts with example 1e in
OALD7 in terms of verb form in *that*-clause. After example 1d, it is natural for Chinese learners produce sentence like *It was suggested we performed at the party*. Most examples in entry *suggest* from *LDOCE4* are still reception-oriented. The examples in grammar box indicate the syntactic constrains, but both examples used for the syntactic structure *Suggest that sb do sth* emphasize the inaccurate usage *suggest sb to do* rather than the verb form in *that*-clause. In the example following the syntactic structure *Suggest doing sth*, salience is given to the forbidden usage of *suggest to do*. All these constrain will not prevent Chinese learners produce sentences like *He suggested that we left now. He suggested a book to me. He suggested me a book. He suggested me wearing something warm*. These sentences are very common in Chinese learners’ English writings as such syntactic structures are applicable in Chinese.

| 1a. They keep suggesting ways to keep my weight down. |
| 1b. She wrote to me and suggested a meeting. |
| 1c. I suggest you phone before you go round there. |
| **1d. It has been suggested that** the manager will resign if any more players are sold. |
| **1e. Joan suggested asking her father for his opinion.** |
| **1f. The therapist suggested how Tony could cope with his problems.** |
| **1g. May I suggest that you think carefully before rushing into this?** |
| **1h. No possible explanation suggests itself (=is able to be thought of).** |
| 2a. Trends in spending and investment suggest a gradual economic recovery. |
| 2b. Opinion polls suggest that only 10% of the population trusts the government. |
| 2c. The evidence suggests that single fathers are more likely to work than single mothers. |
| 3a. John Roberts has been suggested for the post of manager. |
| 4a. Are you suggesting my husband’s been drinking? |
| 5a. I’m not suggesting for one moment that these changes will be easy. |
| 6a. The stage was bare, with only the lighting to suggest a prison. |

**Grammar**

*Suggest that sb do sth*: *He suggested that we go (Not suggest us to go) for a drink.*

*You can miss out *that*: What do you suggest we do (Not suggest us to do)?*

*Suggest doing sth*: *I suggest wearing (Not suggest to wear) something warm.*

*Suggest sth*: *She suggested a walk before dinner.*

(Figure 2: article example from *LDOCE4*)

In *LLA2*, all examples in article *suggest* are grouped according to different syntactic structure under one definition “to tell someone your idea about what they should do, where they should go etc, or about what you and they should do together”. The examples here are concentrated on providing optional ways for learners to express their ideas with the meaning of “propose”, instead of exploiting all the possible contextual meanings word *suggest* acquires. Such practice helps learners in production if they consult dictionary just for the formal feature of *suggest*, but if learners come to confirm whether *suggest* means *indicate* and its usage, they will be discouraged. Besides, all the examples show the syntactic structure in active voice and subjects are animate, which maybe interpreted by learners passive voice is not acceptable nor the inanimate subjects. There is no explicit indication of the syntactic constrains on the usage of *suggest to do, suggest sb to do* and *suggest sb sth*. 

7
1. “Why don’t you come with us?” Alan suggested.
2. It was a sunny afternoon, and Jim suggested a trip to the beach.
3. My Dad suggested that I should apply for the job.
4. I suggest we take a break and finish this later.
5. It was raining heavily, and she suggested calling a taxi.
6. Can you suggest where we might be able to get a decent meal?

(Figure 3: article example from *LLA2*)

In conclusion, the examples in article *suggest* from three dictionaries demonstrate the possible syntactic structures and collocation of *suggest*, but all lack constrains on the usage of *suggest* in one way or another. As Humblé (2001:69) states “it is not the task of dictionaries to teach grammar in a comprehensive way, but the dictionary is the place where learners look for information on particular grammar items such as syntactic constraints.” Nesi (2000) supports this statement with her experiments on the examples used by EFL learners in productive activity. The vocabulary used in examples from the three dictionaries is reduced, but the information for production purpose conveyed through examples is far from explicative. *OALD7* and *LDOCE4* are claimed to be used for comprehensive purposes, but still give more weight to reception and cognitive learning than production in terms of examples. For Chinese learners in productive situation, they resort to example not only for what is possible (descriptive information) to express themselves but also what is constrained (prescriptive information) to express themselves correctly and appropriately. Other types of data will function together with examples to facilitate production tasks. Moreover, a clear structure and typography is called for easy access the explicative messages intended by the examples.

4. CONCLUSION

In production, learners expect to extract information about “syntactic behavior, collocational preferences and selectional restrictions, sociolinguistic features (including register and regional variety), semantic features and contextual effects” (Rundell, 1999: 37). Examples are long assumed to be an effective way to describe various aspects of lexical behaviors in different contexts for learners. The above analysis demonstrates that the discussion about selection and presentation of examples will be superfluous without considering the purposes for which the examples are used. The parameters affecting exemplification should be considered with learners’ particular needs arising in the specific user situations, which are greatly shaped by the learners’ mother tongue, background culture, their learning circumstances as well as their L2 proficiency. Furthermore, examples function with other types of data, such as definition, grammatical notes, word class, etc. to help learners in production.

REFERENCES


JAPANESE TERMS FOR ‘FATHER’:
SOME QUESTIONS OF MEANING AND USE

Anthony E. Backhouse
Hokkaido University, JAPAN
aebackhouse@gmail.com

The meaning and use of kinship terms is a topic of basic importance in lexical semantics, in Asian languages as elsewhere. This paper aims to contribute to the description of terms denoting male parents (‘father’) in modern Japanese. Existing lexicographical and pedagogical treatments of these terms tend to focus on a restricted range of items and uses, typically emphasizing the distinction between otoosan (as an honorific term used in address/reference to others’ fathers, as well as to one’s own father within the family) and chichi (as a somewhat formal term used to refer to one's own father with outgroup addressees). Such accounts leave a variety of other items and uses unattended. For example, in certain situations otoosama, rather than otoosan, is appropriate, indicating the need to recognize different types or degrees of ‘honorificness’. By contrast, we find that chichi is also used frequently in descriptions of racehorses’ pedigrees. Additional common items include chichioya (lit. ‘father-parent’): this is defined in many Japanese monolingual dictionaries simply as ‘chichi’, sometimes as ‘chichi (de aru oya)’ (‘(a parent who is a) father’), but this leaves differences in meaning and use between these two terms unclear. With the help of corpus evidence, we take up such questions and attempt to specify additional semantic and pragmatic factors in this area.
HOW ILLUSTRATIONS CAN MAKE AN ENCODING DICTIONARY EFFECTIVE

Mahmood Ahmad
B. Z. University, Multan, PAKISTAN
mahmood.szbf@yahoo.com

Abstract: Judged against the lexical needs of the target users, a dictionary should contain sufficient information which is helpful for them. It is a sad fact that Urdu-English dictionaries available on the market fall short of the needs and wants of the target users. Even a cursory browsing of these dictionaries reveals that they contain inadequate information. The greatest drawback of these dictionaries is lack of sense discrimination. In most cases they contain a string of translation equivalents. Given the low proficiency in the target language on the part of the target users it seems improbable that they will pick the appropriate translation equivalent. Further, the dictionary skills have been found to be unsatisfactory. As a result, the effectiveness of these dictionaries is reduced. In order to make an Urdu-English dictionary really effective, the use of illustrations, both verbal and pictorial, should be incorporated in the design of the dictionary. The illustrations must be authentic taken from real language data gathered in the form of corpora.

Key Words: Learner’s lexical needs, Text production, bilingual dictionaries, user-perspective, sense-discrimination, example sentences, pictorial illustrations, corpus.

1. Lexical Needs of Learners
   The lexical needs of learners fall into two categories: (a) decoding vs. encoding needs. In decoding activities (reading, listening and translation from L2 into L1) L2 texts have several contextual clues which help learners in figuring out the meaning i.e. translation equivalent in the L1. Learners do so by guessing and inferring. Imagine a Pakistani learner of English comes across the following sentence: ‘We keep our dog in the pen’. If s/he is not sure about the appropriate meaning of the lexical item ‘pen’ in this context, s/he will resort to guessing. The words ‘keep’ and ‘dog’ will help her/him infer that a ‘pen’ is a kind of ‘baaRa’ or ‘enclosure’. Consequently, s/he may choose not to consult a dictionary. However, if the learner needs reassurance, s/he will consult some EFL dictionary or more probably an English-Urdu dictionary. The meaning(s) offered in an L2-L1 dictionary are easy to follow for her/him as they are couched in her/his mother tongue. Moreover, s/he is familiar with the pragmatics of the L1 translation equivalent.

   On the other hand, in encoding activities (writing, speaking and translation from L1 into L2) learners face a tougher challenge. As s/he leaves the safe ground of the mother tongue, s/he finds herself/himself of the slippery ground of the foreign language, with numerous pitfalls. The L1 text contains no clue about the possible translation equivalent. Hence there is no chance of guessing the appropriate translation equivalent left. Let us look at following the Urdu sentence

   میٹر کی سوئی خراب ہے

   If a learner does not know the translation equivalent of the lexical items سوئی ‘sooi’, or خراب ‘kharab’ there is nothing in the text to help her/him out. The most familiar strategy used by learners in such a situation is to consult an Urdu-English dictionary.

2. Urdu-English dictionaries
   A number of Urdu-English dictionaries are available of the market. Some of these are as follows:
2.1. Inadequacies of the Urdu-English dictionaries

It is a sad fact that Urdu-English dictionaries available on the market fall short of the needs and wants of the target users. Even a cursory browsing of these dictionaries reveals that they contain inadequate information. (The paper does not deal with the macro-structure of the dictionaries under review).

2.1.1 Lack/absence of Sense Discrimination

The three most popular Urdu-English dictionaries available on the market contain the following information on the lexical items سوؤی ‘sooi’ and خرآب ‘kharab’:

<table>
<thead>
<tr>
<th>Anjuman’s Urdu-English Dictionary</th>
<th>Kitabistan’s Urdu-English Dictionary</th>
<th>Ferozsons’ Urdu-English Dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>سوؤی (H) soo’i (1) needle (2) pin (3) indicator (4) hand (of clock, etc.) (5) tongue (of scales) (6) pointer on compass.</td>
<td>سوؤی soo’i (1) needle (2) pin (3) indicator (4) hand (of clock, etc.) (5) tongue (of scales) (6) pointer on compass.</td>
<td>سوؤی (H) soo’i (1) needle (2) pin (3) indicator (4) hand (of clock, etc.) (5) tongue (of scales) (6) pointer on compass.</td>
</tr>
<tr>
<td>خرآب ‘kharab’ (A) adj. Bad; poor; broken down; ruined; spoiled; corrupt; unchaste; defiled; polluted; contaminated; unclean</td>
<td>خرآب ‘kharab’ (A) adj. Bad; (2) wretched; (3) miserable; (4) poor; (5) depraved; (6) corrupt; (7) obscene; (8) spoiled; (9) defiled; (10) ruined; (11) dilapidated; (12) deserted</td>
<td>خرآب ‘kharab’ (A) adj. Bad; depraved; ruined; depopulated; deserted; abandoned; miserable; spoiled; waste; poor; wretched; defiled; corrupt; obscene</td>
</tr>
<tr>
<td>خرآب کرنا v.: To spoil; to ruin; to waste; to demoralize; to pollute; to contaminate; to make clean; to violate (woman); to corrupt; to vitiate; to mar; to depopulate; to desolate</td>
<td>خرآب کرنا v.: To spoil; to pollute; to rape; to lay waste; to corrupt; to vitiate to mar to depopulate; to desolate</td>
<td></td>
</tr>
</tbody>
</table>

The greatest drawback of these dictionaries is lack of sense discrimination. In all the three instances given above, a string of translation equivalents has been offered. Shorn of context, they are most likely to confuse the dictionary user. Given the low proficiency in the target language on the part of the target users it seems improbable that they will
pick the appropriate translation equivalent of سوي sooi i.e. ‘indicator’. There is nothing to help her/him to choose the appropriate translation equivalent. If s/he happens to choose the right option, it is matter of chance only. Worse, none of the three dictionaries contains the appropriate translation equivalent of خراب kharab i.e. out of order. They fail the user miserably.

This is not an isolated example. None of the above-mentioned Urdu-English dictionaries contains verbal illustrations. The situation is simply deplorable. The reason is the ivory-tower approach adopted by the publishers. They do not bother to take into account the users’ needs and wants deeming them not worthy of it. They exploit the users’ trust who consider dictionaries to be infallible and the sole arbiters in matters of language use.

2.1.2 Collocations
Matters are complicated due to collocations. Even if the translation equivalent is given in the dictionary, the user needs to be guided regarding which word will occur with it. For example, the translation equivalent of the Urdu lexical item مہربان mihraan offered in the above dictionaries are as under:

<table>
<thead>
<tr>
<th>Anjuman’s Urdu-English Dictionary</th>
<th>Kitabistan’s Urdu-English Dictionary</th>
<th>Ferozsons’ Urdu-English Dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>مہربان mihraan, adj. kind; sympathetic</td>
<td>مہربان mihraan ADJ. (1)kind (2) considerate Loving; (3) affectionate; (4) friendly (5) loving</td>
<td>مہربان mihraan, adj. Loving; affectionate; kind.</td>
</tr>
</tbody>
</table>

As is clear from the above instances the dictionaries under consideration do not contain the collocations of the translation equivalents. Suppose the user wants to translate the following sentence into English

اﷲ ﮨﻢ ﭙﺭﻤﮩﺮﺑﺎﻦ ﮨﮯ Allah hum par mihrban hay

In all probability s/he will translate it as follows:

God is kind *on us. (Incorrect)

Instead of

God is kind to us. (Correct)

2.1.3 Count vs. Non-Count Noun
There are many other intricacies involved. For example, there is no such thing as count vs. non-count noun. In English the difference exists and has a significant bearing on the grammatical behaviour of the lexical item. To take an example, the translation equivalent of the Urdu lexical item ‘بال’ bal is hair. However, whereas the Urdu lexical item has plural, the English lexical item does not have. Pakistani learners carry over the feature into English. It is not uncommon to find such errors as:

(Urdu) Us kay bal ghungrialy hain.

*Her hairs are curly. Incorrect.Tr.

Her hair is curly. Correct Tr.

Again if one turns to the above-mentioned dictionaries, one will find the following information on the head word:
The information given is misleading, to say the least. The senses (Persian) ‘Wing’; ‘pinion’; ‘feather’; ‘strength’; ‘power’ and (Arabic) ‘Heart’; ‘mind’; ‘condition’; ‘state’; ‘power’ do not exist in Urdu. Further the senses (Hindi) ‘crack in glass or crockery /china’; ‘an infant/child’ exist only in compounds. The sense ‘ear of corn’ is the translation equivalent of ‘بالي’ bal. The sense ‘a dance’ is seldom used in Urdu.

Again, no care has been taken to disambiguate the meanings of the translation equivalents. The chance that the user would pick the correct option is very slim. More importantly, should s/he is fortunate enough to choose the right option, the entry does not give the much needed syntactic information which can help her/him use the word in the singular sense. Given the fact that the dictionary skills of the dictionary users have been found to be unsatisfactory (cf. Iqbal 1987), the effectiveness of these dictionaries is further reduced.

3. The Use of Illustrations
3.1. Verbal Illustrations

The most useful means to disambiguate various senses has been found to be the use of illustrations (Stein 2002). Verbal illustrations can be employed to indicate the syntactic behaviour of lexical items, e.g. the features ‘countable’ vs. ‘uncountable’ for nouns, ‘attributive’ vs. predicative’ vs. postpositive’ for adjectives, the use of demonstratives or possessive pronouns as determiners or pronouns, etc. The example sentences in a bilingual dictionary comprise the SL and TL halves. Their function is to show how the entry word can be employed in context.

Let me illustrate the point with a concrete example. The translation equivalents of Urdu lexical item پر ‘par’ ‘wing’, ‘quill’ and ‘feather’ and ‘quills’ may be illustrated in an Urdu-English dictionary with the help of verbal illustrations (cf. Ahmad 2008):

<table>
<thead>
<tr>
<th>پر</th>
<th>بال</th>
<th>بال</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘par’</td>
<td>‘bal’</td>
<td>‘bal’</td>
</tr>
<tr>
<td>Noun</td>
<td>N.M.</td>
<td>N.M.</td>
</tr>
<tr>
<td>wing:</td>
<td>(1) heart</td>
<td>(2) mind</td>
</tr>
<tr>
<td>(2) strength;</td>
<td>(3) condition;</td>
<td>state</td>
</tr>
<tr>
<td>(A) bal</td>
<td>N.M.</td>
<td>A wing;</td>
</tr>
<tr>
<td>n.m.</td>
<td>pinion.</td>
<td>(P)</td>
</tr>
<tr>
<td>bal</td>
<td>bal</td>
<td>bal</td>
</tr>
<tr>
<td>(H) n.m. &amp;f.</td>
<td>N.M.</td>
<td>&amp;f.</td>
</tr>
<tr>
<td>hair,</td>
<td>(1)</td>
<td>hair,</td>
</tr>
<tr>
<td>(2) crack in glass</td>
<td>(2)</td>
<td>a crack</td>
</tr>
<tr>
<td>or crockery;</td>
<td>(3)</td>
<td>in glass or china</td>
</tr>
<tr>
<td>a thin crack</td>
<td>(4)</td>
<td>N.M. ear (of corn)</td>
</tr>
<tr>
<td>(3) ear of corn.</td>
<td>(4)</td>
<td>(used as</td>
</tr>
<tr>
<td>(4) An</td>
<td>bal</td>
<td>bal</td>
</tr>
<tr>
<td>infant.</td>
<td>(E)</td>
<td>(E)</td>
</tr>
<tr>
<td>bal</td>
<td>N.M.</td>
<td>bal</td>
</tr>
<tr>
<td>(E) n.m.:</td>
<td>(1)</td>
<td>(1) heart</td>
</tr>
<tr>
<td>(2) A dance.</td>
<td>(2)</td>
<td>mind</td>
</tr>
</tbody>
</table>

Although the ostrich has wings, it can not fly. (ii) feather: پرندے کی چشمتی کی بھر: the feathers on the breast of the bird  پرندے کی چشمتی کی بھر: The pea-cock has a crest of plumes on its head. (iv) plume: مورکے سرپر گاڑے کا: The pea-cock has a crest of plumes on its head.

The ostrich has wings, it can not fly. (ii) feather: پرندے کی چشمتی کی بھر: the feathers on the breast of the bird  پرندے کی چشمتی کی بھر: The pea-cock has a crest of plumes on its head. (iv) plume: مورکے سرپر گاڑے کا: The pea-cock has a crest of plumes on its head.
(iii) Postposition:

at: ؛(vi) by: یور بینکرین؟؟؟

Who is at the door? (ii) in: لیو روکنی؟؟؟

There were birds in the sky. (iii) on: دیکی سرپینکه رهاتی؟

He put the books on the table. (iv) over: هیخارکی لیپوردیر انصار کرته؟

A bulb hung over her head. (v) upon: یور برینکرین؟؟؟

We depend upon plants for food. (vi) by: یور برینکرین؟؟؟

Do you go to school by bus? (vii) یور برینکرین؟؟؟

Let us discuss the matter.

3.2. Pictorial Illustrations

Pictorial illustrations are another means to enhance sense disambiguation. It is often said that ‘a picture is worth a thousand words’. In particular, they either supplement or substitute encyclopedic information in case of cultural items in dictionaries. To quote Stein (2002), “of the five senses sight plays a dominant role in the cognitive development of the human mind.” It is small wonder one of the important features of the modern lexicography is the use of pictures to further the users’ understanding.

The difference among the various senses of the translation equivalents of ‘par’ have been illustrated with the help of a picture (cf. Ahmad: 2008):

![Figure 2](par.png)
The following illustration can be used to show different kinds of سوئ ‘sooi’:

The translation of the sentence ميتر كي سوئ خراب بي ‘meeter ki sooi kharab hay’ will be: ‘The indicator of the metre is out of order.’

Illustrations can be effectively used to impart cultural information as well. For example, the difference between جهي ‘chapati’ or روتی ‘roTi’ and the English translation equivalent ‘bread’ has been made clear with the help of a pictorial illustration:

4. The Use of Corpus

Unlike the past when the lexicographers relied on their intuition and linguistic knowledge to select definitions/the translation equivalents and to construct example sentences, now it is possible to use corpus data to select genuine examples. The innovation was introduced in COBUILD, later followed by all other major publishers. The name COBUILD is the abbreviation of "Collins Birmingham University International Language Database". However, it is necessary to warn that corpus examples also have loop ends in that they may contain irrelevant information or may not illustrate the meaning, syntax and style in an effective manner and lead to oddities. Also, they are subject to the compilers’ bias. Last but not least, they may contain vocabulary that learners may not know. Hence corpus evidence must be supplemented by the editorial judgement in order to overcome the bias of the compilers of the corpus and make the examples appropriate as well as tidy so that they can meet the competence and needs of learners. In short, the motto of the lexicographers should be 'corpus based, but not corpus bound' (Summers; 1995).
5. Conclusion

The above discussion makes it clear that mere specification of different senses does not serve the purpose because listing a string of translation equivalents simply confuses the user in that it is not clear for her/him which sense ‘fits’ the particular case. Lack of sense discrimination has a very negative impact on their usability. In order to make dictionaries really effective for encoding purposes, illustrations – verbal as well as non-verbal- should be used. Last but not least, the illustrations should be taken from corpus. The EMILLE is the only corpus of Urdu available for research purposes. It represents the variety of used in India. The corpus of Pakistani is to be prepared.

References
Collins COBUILD Advanced Learner's Dictionary.
Feroze Sons Dictionary (1960). The name of the compiler not given. Lahore: Ferozsons Ltd.

Other Works
The EMILLE Project. Retrieved on Nov. 16, 2006 from http://www.emille.lancaster.ac.uk/
ASSOCIATING COLLOCATIONS WITH DICTIONARY SENSES

Abhilash Inumella
IIIT Hyderabad, INDIA
abhilashi@students.iiit.ac.in

Adam Kilgarriff
Lexical Computing Ltd., UK
adam@lexmasterclass.com

Vojtěch Kovář
Masaryk University, Brno, CZECH REPUBLIC
xkovar3@fi.muni.cz

Abstract: We describe a project to assign the collocations for a word, as automatically identified from a large corpus, to its distinct senses. The goal is, in the short term, a new English collocations dictionary (Macmillan Collocation Dictionary, MCD) and in the long term, a ‘disambiguating dictionary’ supporting a range of language tasks. The project is fully corpus-based, and the core infrastructure is the Sketch Engine corpus query system. We have explored both automatic methods using Word Sense Disambiguation algorithms and a computer-supported approach which integrates two new technologies: GDEX (Good Dictionary Example finding) and TBL (TickBox Lexicography). As at summer 2009, the lexicography for MCD is proceeding apace using both of these innovative methods.

1. THE DREAM OF THE DISAMBIGUATING DICTIONARY
It is now commonplace to link a dictionary to electronic texts (in word processors, web browsers or other tools) so that, by clicking or hovering over a word, the user can see the entry for the word in the dictionary. Many publishers offer their dictionaries in this form. The basic task is easy: it is one of matching the string in the text to a headword in the dictionary. Publishers have more, or less, successful solutions to the associated issues of correctly identifying dictionary headwords for the inflected forms found in texts and of matching multi-word expressions found in the dictionary.

One thing they do not do is take the user to the correct sense of a polysemous word. This is desirable. The user would no longer need to read the whole entry and work out which sense was relevant. For long entries this can be a forbidding task, particularly for learners who are struggling with the language in the first place. If the dictionary is bilingual, then the correct sense becomes the correct translation: the user could be directly given an appropriate translation.

If the dictionary publisher had the ability to disambiguate in this way for the user, then they would also have the ability to disambiguate offline, and that would be a great boon for automatic translation and a range of other language technology applications including question answering, information retrieval and information extraction. They would have a disambiguating dictionary, and they would have solved the problem of Word Sense Disambiguation (WSD).

WSD has been a challenge for language technology researchers since the earliest days of the field (see Agirre and Edmonds 2006) for a wide-ranging review of the field and description of the state of the art). It remains painfully intractable, with all systems in the
SENSEVAL and SEMEVAL competitions making errors in over a quarter of cases. So the disambiguating dictionary (which does not make many, many errors) remains a dream.

But what steps might be made in that direction? In Kilgarriff (2005) we make the case that collocations provide a productive framework for thinking about the issue. Yarowsky (1993) put forward the ‘one sense per collocation’ hypothesis: to the extent that it is true, collocates serve to disambiguate. As a step towards disambiguating occurrences of words in running text, we can associate a word’s collocations with one or other of its senses. This will probably be an easier task than full WSD because a collocation is only a collocation if it is a reasonably common pattern of usage for a word, so we will be able to find multiple examples of it in a sufficiently large corpus, so:

- We will not be aiming to disambiguate one-off and anomalous uses
- We will always have multiple contexts of a collocation to use as input to any disambiguation algorithm
- Collocations are often given (implicitly or explicitly) in dictionary entries
- It is a bounded task: whereas a word can appear in any number of contexts, its collocations will count in the tens or possibly in the hundreds.

Also, WSD systems generally work through collocations: they aim to find collocations (as well as grammatical patterns and domains) associated with each sense of the word, and then use them as clues to disambiguate new instances. So, if our goal is just to disambiguate collocations, we are ‘doing WSD’ but stopping before we attempt the most difficult part.

1.1 Manual, Semi-automatic, Automatic
How then might we associate collocations with senses? There are three options: by hand, by computer, or half-and-half. To do it by hand is a very large undertaking: there are perhaps 10,000 polysemous words to be covered (Moon, 2000) and perhaps an average of twenty or thirty collocations to be assigned per sense. Fully automatic methods are possible but are likely to make many errors. Semi-automatic methods look promising, and have been tried in the WASPS project (Kilgarriff and Tugwell, 2002).

2. THE PROJECT
The project arose because a leading dictionary publisher, Macmillan, wished to develop their resources, in particular to develop a collocations dictionary (hereafter MCD). They were also aware that a partial solution to the disambiguating dictionary could lead in many other interesting directions. They were active in English Language Teaching, and had a range of ELT dictionaries including an advanced learners’ dictionary (MEDAL 2002, 2007) and several smaller titles based on it. The project was devised in consultation between Macmillan and Lexical Computing Ltd, a company with expertise as the intersection of lexicography, corpus linguistics and language technology.

MCD is to start from MEDAL (2007), and will provide a full account of the collocations of the core senses of around 4,500 common and highly ‘collocational’ (Kilgarriff, 2006) words. As in other collocations dictionaries such as Oxford’s (OCD 2002, 2009) (and also as in word sketches, see below) collocations will be organised according to grammatical relations. Some collocations will be illustrated with examples in the paper book; all will have examples available by mouse-click in online and other electronic versions.
2.1 Infrastructure: The Sketch Engine
From the outset, the project was to be corpus-driven. Lexical Computing Ltd has a corpus tool, the Sketch Engine (Kilgarriff et al 2004, http://www.sketchengine.co.uk) which would provide sophisticated corpus-handling to support the process. In particular, the Sketch Engine creates ‘word sketches’ – one page, corpus-driven accounts of a word’s grammatical and collocational behaviour. Word sketches have been in use in lexicography for ten years now and have received a number of positive reviews; a formal evaluation is underway. The word sketches identify the collocates that we want to assign to senses.

The Sketch Engine was used via its web API: its analyses were available even though the project was proceeding as a separate development in a different continent.

2.2 The Corpus
The output of a corpus tool is only as good as the corpus. The project required a large corpus, so that, even for words and their collocations which are not so common, we would have plenty of evidence. The language was English. The corpus we used is UKWaC (Ferraresi et al., 2008), a corpus of 1.5 billion words for English, drawn from the web, and carefully ‘cleaned’ (to remove advertisements, navigation bars, copyright statements and other ‘boilerplate’ text) and de-duplicated, with all duplicate and near-duplicate documents removed (Pomikalek, 2008). The corpus had already been tokenized, lemmatised and part-of-speech-tagged using the leading tool TreeTagger (Schmid 1994, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/).

3. AUTOMATIC APPROACHES
3.1 Thesaurus method
The thesaurus consists of groups of semantically close word senses. All word senses are defined in the dictionary with each member having a unique ID; this ID is used to refer to them in the thesaurus. On average there are 28 members per class and a total of 10,2531 members are covered in 3,664 thesaurus classes. An example thesaurus class with six members is shown in Figure 1.
The method works on the premise that a sense shares its collocates with its thesaurus class members. The basic intuition has been used by a number of other WSD researchers: Yarowsky (1992) used Roget’s thesaurus, and Agirre and Martinez (2004) use WordNet and large datasets from the web for finding shared collocations. The method is:

**Algorithm 1: Thesaurus method**

1. Input: a sense \( s \)
2. Output: collocates of sense \( s \)
3. Get all the thesaurus classes containing \( s \). Let \( T \) be the set of all the members of all these classes
4. Get collocations of all the members of \( T \) using Sketch Engine
5. Assert a collocate if it occurs as collocate of \( s \) and at least 2 other members of \( T \).

The method combines two resources, the Sketch Engine and the manually crafted thesaurus. The former is used to find the collocations, the latter to assign them. In the Sketch Engine, collocates are lemmas (rather than word forms) standing in a particular grammatical relation such as subject, object, modifier to the headword. This is an advantage for the current method as the lemmatisation makes useful generalisations, and the grammatical constraints increase the specificity of the collocates, thereby reducing noise and eliminating the need for complex scoring mechanisms.
Let us present a sample run of the method. Consider sense 5 of the word *queen* (see Figure 1). In step 3 we find that the thesaurus class members are \( T = \{ \text{cricket, butterfly, leech, worm, bee} \} \). In step 4 the collocates of *queen* and all the members of \( T \) are obtained using the Sketch Engine. In step 5 the collocates common to *queen* and two other members of set \( T \) are identified. A few confirmed collocates are shown in Table 1.

Table 1: Associating collocations to the 5th sense of *queen*: sample output.

<table>
<thead>
<tr>
<th>Collocation</th>
<th>Supporting Thesaurus Class Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;queen, young, a_modifier&gt;</td>
<td>{queen, bee, cricket}</td>
</tr>
<tr>
<td>&lt;queen, fly, subject_of&gt;</td>
<td>{queen, bee, butterfly}</td>
</tr>
<tr>
<td>&lt;queen, lay, subject_of&gt;</td>
<td>{queen, butterfly, leech, worm}</td>
</tr>
<tr>
<td>&lt;queen, feed, object_of&gt;</td>
<td>{queen, worm, bee}</td>
</tr>
<tr>
<td>&lt;queen, breed object_of&gt;</td>
<td>{queen, worm, butterfly}</td>
</tr>
<tr>
<td>&lt;queen, become, object_of&gt;</td>
<td>{queen, worm, butterfly}</td>
</tr>
</tbody>
</table>

While the method produced some promising results, it also made many errors, for reasons including the following:

1. The thesaurus was not ideal for our purposes because many word senses were in the thesaurus class for several senses of a word. Such thesaurus class members are of limited use for disambiguation.
2. Word sense frequencies are typically highly skewed (Kilgarriff, 2004; McCarthy et al., 2004). Then, depending on the skew and the nature of the senses, most of the collocates often relate to the dominant sense. This is an issue for the ‘bee’ sense of *queen* as it is rare.
3. Thesaurus class members are often themselves ambiguous, and the relevant sense will often not be the dominant sense. Then there are unlikely to be many collocates for the relevant sense in the word sketch. For example, the dominant sense for *cricket* is the sport, and the collocates largely relate to the sport sense.
4. Thesaurus classes in MEDAL are not constrained to be words that are distributionally similar to each other. Thus, staying with the *cricket* (sense 1) example, the thesaurus class is \{cricket, ashes, bat, lbw, out, misfield, duck, Lord’s, googly, ... \}. Though the thesaurus class members are all related to the game of cricket they do not share collocates. They belong to different semantic classes, viz *cricket* is a game, *Lord’s* is a place, *ashes* is a tournament etc.
5. Collocates of thesaurus class members sometimes occurs with the head word where the headword is not in the relevant sense. For example, the collocation <queen, become, object_of> is associated with the 5th sense of *queen* since it is also a collocation of *worm* and *butterfly*. This is incorrect both linguistically and zoologically: linguistically because <queen, become, object_of> is usually the first ‘human royalty’ sense of *queen*, and zoologically because queen bees get special treatment from birth so non-queens never become queens!
6. Some collocates occur with most of the words: For example, the collocation <queen, see, object_of> is valid with all the senses of *queen* and it is also a collocation of all the thesaurus class members.

We would like to model the frequency distribution of senses. It is very often the highest-scoring WSD strategy to assign all instances to the commonest sense (McCarthy et al., 2004). In recent work we are modelling frequency with the indirect evidence that
(1) MEDAL senses tend to be in frequency order, commonest first, and
(2) The frequency of a sense tends to vary with entry length.
To use this model, we need the thesaurus algorithm to output a score instead of a Boolean. This is work in progress.

3.2 Dictionary-parsing
Each sense of a word in MEDAL is provided with definition and examples. There are in all 84,799 example sentences containing a very large number of collocates, particularly since example sentences are selected in order to show collocates (alongside other constraints). To extract collocates we parsed the example sentences.

We used the Stanford dependency parser (Klein et al., 2003). It outputs the dependency structure: a tree where each edge denotes a labelled dependency between two lexical items, ‘edge-parent’ and ‘edge-child’. In Figure 2 the leftmost edge is labelled $nsubj$ (edge-label) and it is between $analyse$ (edge-parent) and $scientists$ (edge-child).

Once the dependency structure is obtained, matching rules are employed to match the edges in the tree and then a collocation generation rule is employed to generate the collocation. For every matching rule there is an associated collocation generation rule. The matching and generation rules have the following structure:

Edge Matching rule: <edgeLabelName, 0/1>
Collocation Generation rule: <gramrel, 0/1>

An edge matching rule can be read as follows: Match an edge only if it has the edge-label edgeLabelName and the headword occurs as the edge-child or edge-parent corresponding to 0 or 1 respectively. Once the edge is matched the following generation rule is employed to generate the collocation.
A collocation generation rule can be read as follows: Generate a collocation with the grammatical relation gramrel and the edge-child/edge-parent as modifier corresponding to 0 or 1 respectively.

Example:
Edge matching rule: <subject_of, 1>
Collocation generation rule: <subject, 0>
The above rule can be read as follows,
If:
there is an edge label subject_of in dependency tree and
the senseHead occurs as edge-parent
then generate collocation with:
subject as grammatical relation and
depth-child as modifier

Separate rules are written for noun, verb and adjective senses. (Other word classes have not been addressed.) There are nineteen matching rules along with their corresponding generation rules. These are employed to extract 54,575 <headword, grammatical-relation, collocate> triples.

Below are some of problems encountered using the dependency parser.

1. There are a number of stop words listed as collocates. For example, have, through, he, be.
2. Sometimes the parser changed the spelling of the word, For example it changed grey to gray.
3. The parser failed to give the correct dependency structure of some sentences.
4. In some cases the parser did not detect the sentence boundaries. This happens because the parser recognises the sequence <single capitalized letter, period,
space, capitalized word> as a person name, for example John F. Kennedy. But it is not always always correct: it does not hold in the following sentence:

Write your full name in Box A. There were names ...

As a result the parser interpreted the above as a single sentence.

In addition to parsing dictionary examples, we:

- manually crafted rules to extract collocates from the dictionary’s ‘grammar patterns’
- used style labels for filtering out senses to which collocates would not be assigned. Senses marked humorous, impolite, offensive, literary, old-fashioned, very formal, very informal presented other concerns and were not to be included in the analysis
- excluded phrases that had their own entry in the dictionary, for example Academy Award, Monty Python, Nobel Prize and Rhodes Scholarship. There are 8,885 phrasal headwords in MEDAL.

3.3 Yarowsky

Yarowsky (1995)'s method performs minimally supervised word sense disambiguation. It uses the collocates that are reliable indicators of the sense of the word to perform the disambiguation. It begins with a small set of sense-specific collocates and incrementally augments them using a combination of two heuristics:

- ‘one sense per collocation’: collocations tend to occur only with one sense of a word
- ‘one sense per discourse’: multiple occurrences of a word in the same text, or discourse, tend to all be uses of the same sense of that word (at least at the coarse-grained level: see Krovetz (1998)).

The algorithm first collects a large number of concordances for the word. Then for each sense of the word a small number of examples representative of that sense are identified. These representative examples are called sense seed sets. Collocates specific to the sense are used to identify these seed sets: we used the high-precision collocates obtained from dictionary parsing. After this step a number of concordances still remain unassigned to a sense. It then proceeds in an iterative manner by updating the seed sets in each step. The seed sets are used in training a decision list algorithm. This in turn is used to label the unassigned samples and for adding the high probability samples to seed sets. In this process the sense specific collocations are identified by the decision list algorithm. Finally a stopping criterion is used to terminate the algorithm.

The algorithm is summarised as follows:

Algorithm 2: Yarowsky

1. Get 50000 concordances of the polysemous word using the Sketch Engine
2. Identify seed collocates obtained by dictionary-parsing.
3. Build initial seed sets of concordances: the concordance lines containing seed collocates
4. Training and labelling new samples:
   4a Train the supervised classification algorithm on the sense-specific seed sets using the decision list algorithm
   4b. Use the resulting classifier to label the entire sample set
   4c. Add the high scoring unassigned samples to seed sets
4d. Collect the high scoring collocates.
4e. Use one sense per discourse constraint to augment this addition
4f. Go to step 4a
5. Stopping criteria: the algorithm is run until convergence. (In our experiments we stopped the algorithm after 5 iterations.)

4. A REVIEW: AUTOMATIC vs COMPUTER-SUPPORTED
As at January 2009, the performance of the algorithm was still disappointing, with many non-standard features of specific word senses still not handled correctly. The lexicographic work on MCD was to start shortly, and this required a working environment for lexicographers in which the automatic work speeded up their work rather than causing them delay. Editing incorrect work tends to be slower and more awkward than doing it correctly in the first place. A decision was taken to provide computational support for manually assigning Sketch Engine collocates to MEDAL senses rather than attempting the task automatically. The development of the automatic system has recently restarted, in readiness for future projects by which time we expect to be able to assign collocates more accurately, in parallel with the lexicography for MCD.

5. TICKBOX LEXICOGRAPHY
In corpus lexicography we:
- identify the senses for the word and then, for each sense:
  - identify the key patterns, collocate and phrases
  - find example sentences.
This is then the base analysis of the language which will serve for the development of a range of dictionaries, monolingual and bilingual (where the language analysed is the source language, and the same analysis will form the basis whatever the target language) (Atkins, 1994; Atkins and Rundell, 2008: 97-101).

5.1 GDEX (Good Dictionary Example Finder)
Kilgarriff et al. (2008) presents GDEX, a system, integrated with the Sketch Engine, for finding good dictionary examples for a word or collocation. It works by taking the sentences in a concordance and sorting them according to a number of heuristics including sentence length, the proportion of high-frequency words, the number of obscure words (if any) and the number of capital letters, numbers and other non-letter characters. It then either sorts the concordance lines ‘best-first’ in the Sketch Engine, so typical users see only the best examples, or outputs the best examples to some other process.

GDEX opens the way to further supporting the lexicographer by letting them select collocates and examples by ticking boxes rather than re-typing or using standard cut-and paste. We call this “tickbox lexicography” (TBL).

The process is as follows:
- the lexicographer sees a version of the word sketch with tickboxes beside each collocate
- for each sense and each grammatical relation, they tick the collocates they want in the dictionary
- they click a 'next' button
- they then see, for each collocate they have ticked, a choice of six (by default) corpus example sentences, chosen by GDEX, each with a tickbox beside it
- they tick the ones they like
- they tick a "copy to clipboard" button.

The system then copies the collocates and examples, embedded in an XML structure, onto the clipboard. (Each target dictionary has its own TBL application, to meet the requirements of the user’s dictionary-editing system and target dictionary.) The lexicographer can then paste the structure into the dictionary editing system.

Thus, TBL models and streamlines the process of getting corpus data out of the corpus and into the dictionary.

5.2 Customising TBL for MCD

To set up TBL for MCD, we first developed customised word sketches in which the grammatical relations were those to be used in MCD. This required work on the underlying part-of-speech tagging and grammatical-relation-finding software. GDEX was also customised, with the incorporation of a long list of stop words, to minimise the chances that GDEX would select examples containing offensive material.

In the first trials, lexicographers selected all the example sentences (typically six per collocate) that were to be used in the electronic version of MCD, but this proved too slow. We changed to a strategy where only the examples which are to appear in the book are selected by lexicographers. For all others, GDEX will be trusted to deliver good examples. (The manually-selected items will be edited as necessary by lexicographers, whereas the others will be full and unedited corpus sentences.) These sentences will be selected in a batch process after the lexicography is done, as this will reduce the volume of data to be handled by the clipboard and the dictionary editing system, and will allow the GDEX process to be fine-tuned, using experience during the project as a guide, towards the end of the project.

MEDAL 2007 already contains 500 ‘collocation boxes’ for word senses of common words, with collocations classified according to grammatical relations, and further collocations in bold in regular entries. It was desirable to carry them across into MCD in a way which integrated with MCD lexicography. To this end we:

- analysed MEDAL to find all collocations, either in collocation boxes or shown in bold within regular entries
- identified the grammatical relation they stood in to the headword
- checked to see if they were already in the word sketch:
  - if they were (as they usually were), colour them red (in the word sketch) and pre-tick the tickbox, as they will almost always be wanted in MCD
  - if they were not, add them in (in red), with links to their corpus instances and pre-ticked tickboxes.

The dictionary editing software used for MCD accepts XML pasted from the clipboard so, once the lexicographer has
- called up the customised word sketch for the headword
- selected the grammatical relation
- selected collocates
- selected examples for the paper dictionary
they click a ‘copy to clipboard’ button, and then paste the material (using standard CTRL-V) into the dictionary entry.
5.3 Current Status
The computer-supported approach, with TBL and GDEX, has been in use in production mode, by a team of six lexicographers, since mid 2009. Following three months of intensive development to optimise the process for the specific situation of MCD and MEDAL, both the lexicographers and the Macmillan management team are clear that the approach is both speeding up the compilation and contributing to the accuracy and completeness of MCD. We believe TBL has great potential for both streamlining corpus lexicography and making its outputs more accountable to the corpus.

6. REFERENCES


---

i There are of course many qualifications which might be made here, as figures depend on how and what you count. Readers are referred to Agirre and Edmonds (2006) and Agirre et al (2009) for detailed discussions.

ii We use collocation to refer to the two-word unit, and collocate to refer to the word that collocates with the headword.

iii Our version of UKWaC is 20% smaller than the one described in Ferraresi et al. (2008). Both Ferraresi’s group and ours have undertaken further rounds of removing unwanted material, and both groups now use versions which have benefited from the other group’s work.

iv We are aware that become is a copula verb and does not take an object. In the simplified grammar of the word sketches, the distinction is not made.

v In each iteration the top 2% of collocates over all senses are collected.
Abstract: State-of-the-art lexicography requires corpora, but for many languages there are no large, general-language corpora available. Until recently, all but the richest publishing houses could do little but shake their heads in dismay as corpus-building was long, slow and expensive. But with the advent of the Web it can be highly automated and thereby fast and inexpensive. We have developed a 'corpus factory' where we build lexicographic corpora. In this paper we describe the method we use, and how it has worked, and how various problems were solved, for five languages: Dutch, Hindi, Telugu, Thai and Vietnamese. The corpora we have developed are available for use in the Sketch Engine corpus query tool.

1. INTRODUCTION
Lexicography needs corpora. Since the innovations of the COBUILD project in the 1980s, the benefits of large electronic corpora for improving accuracy have been evident, and now, any dictionary which aspires to take forward the description of a language needs to be corpus-based.

For the major world languages including Arabic, Chinese, English, German, Italian, Japanese, Portuguese and Spanish, large corpora are publicly available [1]. (By ‘large’, we mean at least 50m words.) But for most other languages, they are not.

In the early days of corpus linguistics, corpus collection was a long, slow and expensive process. Texts had to be identified and obtained, also permission of the copyright holder, and then they were usually not available in electronic form and had to be scanned or keyed in. Spoken material had to be transcribed. The costs were proportional to the size of the corpus and the projects generally took several years.

But then came the internet. On the internet, the texts were already in electronic form and could be obtained by mouse-click. The copyright issue took on a different complexion since what a corpus collector was doing was in outline the same as what Web search engines were doing, and no-one was challenging the legality of that (at least in straightforward cases). The prospects were first explored in the late 1990s (Ghani and Jones 2000; Resnik 1999). Grefenstette and Nioche (2000) showed just how much data was available, even for smaller languages, and a general-purpose, open source tool, BootCaT, was presented by
Baroni and Bernardini in 2004. Keller and Lapata (2003) established the validity of Web corpora by comparing models of human response times for collocations drawn from Web frequencies with models drawn from traditional-corpus frequencies, and showing that they compared well. So, at a theoretical level, the potential and validity of Web corpora for a wide range of languages has been shown. To what extent has the potential been actualised?

Sharoff has prepared Web corpora, typically of around 100 million words, for ten major world languages, primarily for use in teaching translation and similar at Leeds University, but publicly accessible for searching at http://corpus.leeds.ac.uk/internet.html (Sharoff 2006). Scannell has gathered corpora of, in most cases less than a million words for several hundred languages (see http://borel.slu.edu/crubadan/stadas.html).

Here we aim to systematically add to the list of languages for which corpora of around 100m words - large enough for general lexicography – are available.

1.1 Outline of Method and Structure of Paper
The method is:

- Gather a ‘seed word’ list of several hundred mid-frequency words of the language
- Repeat several thousand times (until the corpus is large enough):
  - Randomly select three (typically) of these words to create a query
  - Send the query to a commercial search engine (we have used Google and Yahoo)
  - Google or Yahoo returns a ‘search hits’ page. Retrieve all the pages identified by Google/Yahoo as the search hits. Store
- “Clean” the text, to remove navigation bars, advertisements and other recurring material
- Remove duplicates
- Tokenise, and, where tools are available, lemmatise and part-of-speech tag
- Load into a corpus query tool.

We have applied the method to Dutch, Hindi, Telugu, Thai and Vietnamese.[2]

The method is as used by Sharoff and is similar to that used by Scannell and the Bologna group (Baroni and Kilgarriff 2006, Ferraresi et al. 2008). Like BootCaT, it piggybacks on the work of the commercial search engines. They crawl and index the Web, identify text-rich pages and address character-encoding issues (though they do this with mixed success, as we see below). By using this work already done (and usually, very well done) by the search engines, we save ourselves many tasks.

In section 2 we describe each step in detail, comparing our experiences for the four languages and discussing any particular difficulties that arose. In section 3 we consider how the work might be evaluated, including comparisons with Wikipedia corpora and, for Dutch, a comparison with another large, general-language corpus.
2 METHOD
2.1 Seed Word Selection
For each language, we need seed words to start the process. Sharoff used 500 common words drawn from word lists from pre-existing corpora: the British National Corpus for English, Russian National Corpus for Russian, IDS corpus for German and Chinese Gigaword for Chinese. But for the languages we are most interested in, there are no large, general corpora (which is why we are building them).

Wikipedia (Wiki) is a huge knowledge resource built by collective effort. It has articles from many domains. The whole dataset can be downloaded. One possibility is to treat the Wiki for a language as a corpus. However it may not be large enough, or diverse enough in text type, for general lexicography (see also the evaluation section). It will be small compared to the Web for that language. So we do not use the Wiki as the finished corpus. However it can be used as an intermediate corpus to prepare frequency lists to supply seed words. These seeds can then be used to collect large Web corpora. Currently, Wikipedia hosts around 265 languages including all those for which we hope to build corpora. So we use Wikis as sources of seed terms. This has the advantages that we can apply the same method across many languages, and that the corpora so produced should be ‘comparable’ – or at least more similar to each other than if we had used a different method for gathering seed words in each case.

2.1.1 Extracting Wiki Corpora
A Wikipedia, or Wiki, Corpus is extracted from a Wiki dump for a language. A Wiki dump is a single large file containing all the articles of the Wikipedia. It includes Wiki markup, for example

```
==Sub Heading==
```

where the equals signs are Wiki markup telling the interpretation program to format “Sub Heading” as the title of the subsection of an article.

The steps involved in extracting plain text from an XML dump are:

- Download Wiki XML dump of the target language
- Extract XML pages (one per article, with embedded Wiki markup) from the dump
- Parse Wiki XML page to remove Wiki tags to get plain XML pages
- Extract plain text from the plain XML pages using the Wikipedia2text tool [3].

We used a slightly modified version of the Wikipedia2Text tool to extract plain text from the Wiki XML dump. Table 1 gives some statistics.

<table>
<thead>
<tr>
<th>Language</th>
<th>Wiki XML dump</th>
<th>Wiki XML pages</th>
<th>Plain XML pages</th>
<th>Plain text pages</th>
<th>After filtering files below 10KB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Size in MB</td>
<td>Size in words</td>
<td>Size in MB</td>
<td>Size in words</td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td>1.8 GB</td>
<td>4.1 GB</td>
<td>4.9 GB</td>
<td>2700 MB</td>
<td>83 MB</td>
</tr>
<tr>
<td>Hindi</td>
<td>149 MB</td>
<td>445 MB</td>
<td>485 MB</td>
<td>367 MB</td>
<td>35 MB</td>
</tr>
<tr>
<td>Telugu</td>
<td>108 MB</td>
<td>447 MB</td>
<td>469 MB</td>
<td>337 MB</td>
<td>12 MB</td>
</tr>
<tr>
<td>Thai</td>
<td>463 MB</td>
<td>1.1 GG</td>
<td>1.2 GB</td>
<td>698 MB</td>
<td>89 MB</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>426 MB</td>
<td>1.1 GB</td>
<td>1.3 GB</td>
<td>750 MB</td>
<td>57 MB</td>
</tr>
</tbody>
</table>

Table 1: Wiki Statistics
An alternative is to extract the text from a Wiki HTML dump. We found that the XML dump gave a cleaner corpus than the HTML one. Even though Wiki text may contain HTML tags, most of the text is written using Wiki tags which proved easier to parse than the HTML.

We found that most of the Wiki articles do not have connected text but are short definitions, sets of links, or ‘stubs’: articles which exist for purposes of being pointed to by other articles but which have not themselves been written yet. They need filtering out. Generally they are small. Ide et al. (2002) give an estimation of minimum 2000 words as an indicator of connected text. In line with that, we consider a Wiki file to have connected text if its size is above 10 KB. Our Wiki corpus then comprises all text in files larger than 10KB. Even in this case, we found that some files are still not connected but their effect on frequency lists is not significant. We use this Wiki corpus to build a frequency list for the language.

2.1.2 Words, Lemmas and Tokens
For most languages, most search engines do not index on lemmas but on word forms. They treat different forms of the word as different words. For example the Telugu word నురాంతంలో (“in location”) gave more Yahoo search hits than its lemma నురాంతం (“location”). Sharoff (2006) discusses similar findings for Russian. We used a frequency list for word forms rather than lemmas, and used word forms as seeds.

To get the frequency list of a language from its Wiki Corpus, the corpus needs to be tokenised. The tokenisation details of each language are specified below.
- Hindi, Dutch and Telugu tokenisation is straightforward. Words are separated by white space and punctuation marks.
- In Vietnamese, a word may contain more than one lexical item. We used a Vietnamese word list to identify words in the Wiki Corpus. The algorithm moves a pointer along the sentence and tokenise words such that the maximum number of lexical items fits in the current word. An example is given below
  Input: Vợ tôi, người cùng tôi chia sẻ với văn khó trong
  Output, with slashes to show word ends: Vợ/ tôi/, người/ cùng/ tôi/ chia sê/ với/ văn/ khó/ trong/
- In Thai, words are joined together without spaces to form a sentence, as here
  ปัญหาของประเทศพม่าในภูมิภาคคืออะไร
  We used the open-source SWATH tool for word segmentation which gives:
  ปัญหา/ของ/ประเทศ/พม่า/ใน/ภูมิภาค/คือ/อะไร

2.1.3 From frequency list to seed words
Considerations in collecting seed words are:
- they should be sufficiently general: they should not belong only to a specialist domain
- very-high-frequency function words do not work so well: they are not the focus of search engine companies’ efforts as they are present in most pages for a language so are not useful for discriminating between pages. Search engines may treat them as
stop words and not index them, or give otherwise unhelpful results. Also they are often very short and, in latin-alphabet languages, confusable with words from other languages.

- Capitalised words are normalised to lower case words for Dutch and Vietnamese.

Some studies (Grefenstette and Nioche 2000; Ghani et al. 2003) used only seed words that were unique to the target language, to avoid accidental hits for pages from other languages. Three of the five languages in our sample (Hindi, Telugu, Thai) use their own script so, if the character encoding is correctly identified, there is no risk of accidentally getting a page for the wrong language. For the two latin-script languages, Dutch and Vietnamese, we adopted different tactics.

- For Dutch, we used a word length constraint of at least 5 characters to filter out many words which are also words in other languages: it tends to be short words which are words in multiple languages of the same script. Many words from other languages are not filtered out. However:
  - We are only likely to get a page from another language if all seed terms in a query are also words form the same other language. This becomes less likely where there are multiple seeds and where many multi-language words have been filtered out
  - We have a further stage of filtering for language, as a by-product for filtering for running text, using the highest-frequency words of the language (see below).

- A Vietnamese word may have more than one lexical item and the size of these lexical items are found to be small. Word length is not a good constraint in this case. We used the constraint that a Vietnamese word should contain atleast one Unicode character which is not in ASCII range. [7] Chế biên, dụng, tạo are valid Vietnamese words.

Once the Wiki corpus is tokenised, term frequency and document frequency are calculated. Words are sorted in the frequency list based on document frequency.

When we generate a seed word list, items containing digits and other non-letter characters are excluded as are items not meeting the length and accents constraints for Dutch and Vietnamese. We then set aside the top 1000 words and use the 1000th to 6000th words (i.e. 5000 words) as seed words. The Wikipedias are in UTF-8 encoding and so are the seedwords.

2.2 Query Generation
Web queries are generated from the above seeds using BootCaT's query generation module. It generates tuples of length \( n \) by random selection (without replacement) of \( n \) words. The tuples will not be identical nor will they be permutations of each other. We needed to determine how to set \( n \). Our aim is to have longer queries so that the probability of results being in the target language is high. Also, more different queries can be generated from the same seed set if the queries are longer. At the same time, we have to make sure that the hit count is not small for most of the queries. As long as we get a hit count of more than ten for most queries (say, 90%), the query length is considered to be valid. We define the best query length as the maximum length of the query for which the hit count for most pages is more
than ten. We use the following algorithm to determine the best query length for each language.

**Algorithm 1: Best Query Length**
1: set n = 1
2: generate 100 queries using n seeds per query
3. Sort queries by the number of hits they get.
4: Find hit count for 90th query (min-hits-count)
5: if min-hits-count < 10 return n-1
6: n = n +1, go to step 2

Best query lengths for different languages obtained from Yahoo search hits are shown in *Table 2*. We used a minimum query length of two, so did not apply the algorithm fully for Telugu.

<table>
<thead>
<tr>
<th></th>
<th>length= 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch</td>
<td>1,300,000</td>
<td>3,580</td>
<td>74</td>
<td>5</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>Hindi</td>
<td>30,600</td>
<td>86</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>Telugu</td>
<td>668</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>Thai</td>
<td>724,000</td>
<td>1,800</td>
<td>193</td>
<td>5</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>1,100,000</td>
<td>15,400</td>
<td>422</td>
<td>39</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

Once query-length was established we generated around 30,000 queries for each language.

**2.3 URL Collection**
For each language, the top ten search hits are collected for 30,000 queries using Yahoo’s API. *Table 3* gives some statistics of URL collection.

<table>
<thead>
<tr>
<th></th>
<th>Unique URLs collected</th>
<th>After filtering</th>
<th>After de-duplication</th>
<th>Web corpus size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MB</td>
<td>Words</td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td>97,584</td>
<td>22,424</td>
<td>19,708</td>
<td>739 MB</td>
</tr>
<tr>
<td>Hindi</td>
<td>71,613</td>
<td>20,051</td>
<td>13,321</td>
<td>424 MB</td>
</tr>
<tr>
<td>Telugu</td>
<td>37,864</td>
<td>6,178</td>
<td>5,131</td>
<td>107 MB</td>
</tr>
<tr>
<td>Thai</td>
<td>120,314</td>
<td>23,320</td>
<td>20,998</td>
<td>1.2 GB</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>106,076</td>
<td>27,728</td>
<td>19,646</td>
<td>1.2 GB</td>
</tr>
</tbody>
</table>

We found that Google gave more hits than Yahoo, particularly for languages that have non-ASCII characters. The reason for this may not be the difference in index size. Google normalises many non-UTF8 encoding pages to UTF8 encoding and then indexes on them whereas Yahoo does less normalisation and more often indexes the words in the encoding of the page itself. We verified this for Telugu. [http://www.eenadu.net/](http://www.eenadu.net/) is a famous news site in Telugu which uses non-UTF8 encoding. We restricted the search hits to this news site and for the query చంద్ాబాబు (the name of a famous politician) we got 3170 Google search hits and 3 Yahoo hits. We also ran the query with the original encoding used by Eenadu. There were 0
Google hits and 4500 Yahoo hits. This shows that Yahoo indexed Eenadu but did not normalise the encoding. Since we use UTF8 queries, Google would serve our purposes better for Telugu. But because for licensing and usability reasons, we have used Yahoo to collect search hits to date. We plan to investigate this further, including exploring how other search engines (including yandex, Microsoft’s bing) handle a language and its most common encodings before choosing which search engine to use for a language.

We extended BooTCaT’s URL collection module to store the current query, page size and MIME type for each URL.

2.4 Filtering
The URLs are downloaded using unix wget. Since we already had MIME information for the URL, we downloaded only those pages whose MIME type was text/HTML. We also had page size, so downloaded only those files above 5 KB so that the probability of connected text was greater. Files larger than 2 MB were discarded to avoid any particular domain files dominating the composition of the corpus, and also because files of this length are very often log files and other non-connected text.

The downloaded pages contain html markup and ‘boilerplate’ text like navigation bars, advertisements and other recurring material like legal disclaimers. To remove such content and extract only the connected text, we used the Body Text Extraction algorithm (BTE, Finn et al. 2001). BTE starts from the observation that Web pages typically have material at the beginning and end which is rich in boilerplate and which tends to be heavily marked up, and material in the middle, the ‘body text’, which is linguistic and is the material we want, and is relatively light in markup. It calculates the ratio of text to markup for different parts of the page, divides the page into three sections on the basis of this ratio, and retains only the middle one. BTE was performed on all the downloaded pages to get plain text pages.

These pages are further filtered to check for connected text. Connected text in sentences reliably contains a high proportion of function words (Baroni, 2007). If a page does not meet this criterion we discard the page. We assume that the top 500 words in the frequency list (as prepared from the Wiki corpus) include most function words. To set a threshold for the proportion of tokens to be accounted for by the top-500 words, we sorted all Wiki files according to the proportion of top-500 words in the file. We found that most of the Wiki files at the bottom (below 75-80 %) of this sorted list did not contain connected text. This is either due to bad cleaning by the Wikipedia2Text tool or because the page really did not contain connected text. The Wiki file at 70th % of the sorted list is used to set the threshold: if, in the 70th-percentile file, words from the top-500 list accounted for 65% of all words, then the threshold for the language was set at 65% and any page where less than 65% of the words were from the top-500 list was discarded.

2.5 Near Duplicate Detection
We used perl's Text::DeDuper module for near duplicate detection. This module uses the resemblance measure as proposed by Broder et al (1997) to detect similar documents based on their text. This is a memory intensive task. N-grams (n=5) for each document are generated and similarity (sim_thresh=0.2) is measured between two documents based on the number of overlaps in their n-grams. Since main memory size is limited and can hold only a limited number of files, duplicate detection is done using a sliding window approach. At each iteration a fixed number of non-duplicate files, say 500, whose n-grams can fit in memory,
are identified using the DeDuper module. All other files are taken one file at a time and compared with the n-grams of these non-duplicate files to identify if they are duplicates or not. This process is repeated until all files are covered. A detailed algorithm is given below. After this step, we get the final Web corpus. Sizes are given in Table 3.

Algorithm 2 : Identify Near Duplicates

1: Sort the file names by their file sizes and store all the filenames in a list
2: Identify first 500 non duplicate documents (traversing linearly on filenames list) using DeDuper module
3: Compare rest of the files, a file at a time, with these 500 non-duplicate documents
4: Remove any duplicate files found and store the rest of the filenames in next_filenames list
5: filenames = next_filenames
6: Continue from step 2.

In future, we expect to explore and use the method proposed in Pomikálek and Rychlí (2008).

2.6 Part-of-speech Tagging and Lemmatisation
We part-of-speech-tagged and lemmatised the Dutch Web corpus using Tree Tagger [8]. The resulting corpus is about 2.3 GB in size. For the other four languages, we have not yet been able to find suitable POS-taggers or lemmatisers. We hope to either find them shortly, or possibly work with NLP groups with expertise in the language to co-develop them.

2.7 Loading into a Corpus Query Tool
The corpora were then loaded into the Sketch Engine, where they are accessible at http://www.sketchengine.co.uk. A screenshot of a concordance for the Thai pronoun ข้าพเจ้า (first person singular pronoun used in formal contexts) is given in the appendix.

3. EVALUATION
Corpus evaluation is a complex matter. What does it mean for a corpus to be good? Of course, it depends what we want to use the corpus for. The straightforward answer to the question is “if it supports us in doing what we want to do”.

We anticipate that our corpora will be evaluated in this way, by lexicographers and other language researchers, over time. As they use a corpus and get to know it they will come to realise what it is good for and what it is not. We have had this experience with large English corpora, particularly the Oxford English Corpus, which has now been in use for several years and where new phases of corpus-building have been designed to address the lexicographers’ criticisms of previous versions, which they had got to know very well.

But this kind of evaluation takes time: how might we do a first-pass evaluation of the corpora without waiting?

The only strategy we know of is by comparison: comparing one corpus with another, and, in particular, comparing frequency lists of the two corpora. The topic is explored in general in Kilgarriff (2001) and frequency-list-comparison methods are used for Web corpus evaluation
in Baroni and Kilgarriff (2006), Sharoff (2006), Ferraresi et al (2008). (There are also many studies using frequency list comparisons, also often called keywords analyses, to compare corpora of different text types or regional varieties, to explore the differences between the varieties. Usually word frequency lists are used, though sometimes frequencies related to word classes or grammatical constructions have been explored, notably in Biber (1989).)

For each of the languages, we have two corpora available: the Web corpus and the Wiki corpus. In the case of Dutch, we also have access to a carefully-designed lexicographic corpus. We have evaluated our corpora in relation to both of these points of comparison.

3.1 Comparing Web and Wiki corpora
The Wiki corpora were prepared as sources of seeds for the Web corpus building. But they are also corpora which may be of interest in their own right. How do they compare with the Web corpora? It is possible that they are better for some purposes: they may have a higher proportion of well-written material, as they do not include arbitrary texts in the way that the Web corpora do.

The first point to make is simply that they are far smaller, see Table 4.

<table>
<thead>
<tr>
<th></th>
<th>Wiki Corpora (millions of words)</th>
<th>Web Corpora (millions of words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch</td>
<td>11.3</td>
<td>108.6</td>
</tr>
<tr>
<td>Hindi</td>
<td>3.9</td>
<td>30.6</td>
</tr>
<tr>
<td>Telugu</td>
<td>0.47</td>
<td>3.4</td>
</tr>
<tr>
<td>Thai</td>
<td>6.5</td>
<td>81.8</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>6.8</td>
<td>149</td>
</tr>
</tbody>
</table>

Another hypothesis is that the Wiki corpora are more ‘informational’ and the Web ones more ‘interactional’. Biber (1988) shows how the dominant dimension of variation for English is ‘interactional vs informational’: some kinds of language use are principally concerned with interaction between participants whereas others are principally for conveying information, and this is the principal axis along which texts are best classified for register. Biber (1995) shows how this holds across a number of languages.

Informational language is typical written, and interactional, spoken. It is usually easier to gather large quantities of informational registers, for example newspapers, official reports, academic papers and Wikipedia articles, than interactional ones, including spontaneous conversation. In general, we might expect a Web corpus to be more interactional, and traditional and Wiki corpora more informational. The Web, particularly ‘Web 2.0’, supports interaction and informality. Ferraresi et al (2008) explore register variation in UKWaC, a large Web corpus, comparing it with the British National Corpus, and find UKWaC to be markedly more interactional. However in our case the Wiki corpus was used, via the seed words fed into Yahoo, to generate the Web corpus. One criticism of our method would be that since we use Wikipedia texts, we are very likely to have an imbalance of informational as opposed to interactional texts in the Web corpora.
We explored the question by noting that first and second person pronouns are strong indicators of interactional language. For each pair of corpora, for each of the five languages, we made a list of ten of the commonest first and second personal pronouns (for English the list would be *I* me my mine you your yours we us our) and counted their frequencies in the Web and Wiki corpora. We normalised figures to per-million and calculated the ratio, Web:Wiki, as in Table 5.

Table 5: 1st and 2nd person pronouns in Web and Wiki corpora. All figures in ‘Web’ and ‘Wiki’ columns are frequencies per million words. For Dutch and Vietnamese, counts are case-insensitive. The figure in the Ratio column is the Web:Wiki ratio.

<table>
<thead>
<tr>
<th>Dutch</th>
<th>Hindi</th>
<th>Telugu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Web</td>
<td>Wiki</td>
</tr>
<tr>
<td>ik</td>
<td>5786</td>
<td>2526</td>
</tr>
<tr>
<td>je</td>
<td>4802</td>
<td>975</td>
</tr>
<tr>
<td>jezelf</td>
<td>96</td>
<td>9</td>
</tr>
<tr>
<td>jij</td>
<td>188</td>
<td>37</td>
</tr>
<tr>
<td>jou</td>
<td>102</td>
<td>19</td>
</tr>
<tr>
<td>jouw</td>
<td>99</td>
<td>19</td>
</tr>
<tr>
<td>jullie</td>
<td>367</td>
<td>112</td>
</tr>
<tr>
<td>me</td>
<td>599</td>
<td>294</td>
</tr>
<tr>
<td>mezelf</td>
<td>41</td>
<td>5</td>
</tr>
<tr>
<td>mij</td>
<td>768</td>
<td>344</td>
</tr>
<tr>
<td>Total</td>
<td>14221</td>
<td>4771</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Thai</th>
<th>Vietnamese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Web</td>
</tr>
<tr>
<td>ผม</td>
<td>2935</td>
</tr>
<tr>
<td>สิ่งัน</td>
<td>133</td>
</tr>
<tr>
<td>ชืน</td>
<td>770</td>
</tr>
<tr>
<td>ศัตรู</td>
<td>1722</td>
</tr>
<tr>
<td>ท่าน</td>
<td>2390</td>
</tr>
<tr>
<td>กระผม</td>
<td>21</td>
</tr>
<tr>
<td>พวกเจ้า</td>
<td>434</td>
</tr>
<tr>
<td>ตัว</td>
<td>2108</td>
</tr>
<tr>
<td>กู</td>
<td>179</td>
</tr>
<tr>
<td>ช้าน</td>
<td>431</td>
</tr>
<tr>
<td>Total</td>
<td>11123</td>
</tr>
</tbody>
</table>
For forty-eight of the fifty pronouns, the ratio is greater than one, often many times greater.
The ratio across all ten pronouns varies between 2.4 times more common (Thai) to over five times (Hindi). The Web corpora are far more interactional than the Wiki corpora used to develop them.

### 3.2 Comparing NLWaC and ANW

The ANW corpus is a balanced corpus of just over 100 million words compiled at the Institute for Dutch Lexicology (INL) and completed in 2004 to support the lexicography for the ANW, a major new dictionary of Dutch currently in preparation. It comprises: present-day literary texts (20%), texts containing neologisms (5%), texts of various domains in the Netherlands and Flanders (32%) and newspaper texts (40%). The remainder is the ‘Pluscorpus’ which consists of texts, downloaded from the internet, with words that were present in an INL word list but absent in a first version of the corpus.

To compare the Dutch Web corpus (called NLWaC) with the ANW corpus, we prepared frequency lists for word forms for both corpora and found the ‘keywords’ of each corpus in contrast to the other using the formula

\[
\frac{\text{Freq-per-million-in-corpus1} + 100}{\text{Freq-per-million-in-corpus2} + 100}
\]

(For discussion of the formula and the parameter, see Kilgarriff 2009). We then look at the words with the highest and lowest scores.

The twenty highest-scoring (ANW) keywords and the twenty lowest-scoring (NLWaC) keywords, with English glosses and clustered by themes, are given in Table 6.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Word</th>
<th>English gloss</th>
<th>Theme</th>
<th>Word</th>
<th>English gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANW</td>
<td>Brussel</td>
<td>(city)</td>
<td>Religion</td>
<td>God</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Belgische</td>
<td>Belgian</td>
<td></td>
<td>Jezus</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vlaamse</td>
<td>Flemish</td>
<td></td>
<td>Christus</td>
<td></td>
</tr>
<tr>
<td>Fiction</td>
<td>keek</td>
<td>Looked/watched</td>
<td></td>
<td>Gods</td>
<td></td>
</tr>
<tr>
<td></td>
<td>vorig</td>
<td>previous</td>
<td></td>
<td>http</td>
<td></td>
</tr>
<tr>
<td></td>
<td>kreek</td>
<td>watched/looked</td>
<td></td>
<td>Geplaatst</td>
<td>posted</td>
</tr>
<tr>
<td></td>
<td>procent</td>
<td>Percent</td>
<td></td>
<td>Bewerk</td>
<td>edited</td>
</tr>
<tr>
<td></td>
<td>miljoen</td>
<td>million</td>
<td></td>
<td>Reacties</td>
<td>Replies</td>
</tr>
<tr>
<td></td>
<td>miljard</td>
<td>billion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>frank</td>
<td>(Belgian) Franc</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>zei</td>
<td>said</td>
<td></td>
<td>www</td>
<td></td>
</tr>
<tr>
<td></td>
<td>aldus</td>
<td>thus</td>
<td></td>
<td>And</td>
<td>In book/film/song</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>titles, names etc</td>
</tr>
<tr>
<td>Newspapers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Meppel</td>
<td>City with local</td>
<td></td>
<td>Arbeiders</td>
<td>workers</td>
</tr>
<tr>
<td></td>
<td>gisteren</td>
<td>yesterday</td>
<td></td>
<td>Dus</td>
<td>thus</td>
</tr>
<tr>
<td></td>
<td>Foto</td>
<td>Photo</td>
<td></td>
<td>Macht</td>
<td>power</td>
</tr>
<tr>
<td></td>
<td>Auteur</td>
<td>Author</td>
<td></td>
<td>Oorlog</td>
<td>war</td>
</tr>
<tr>
<td></td>
<td>Van</td>
<td>(in names)</td>
<td></td>
<td>Volk</td>
<td>people</td>
</tr>
<tr>
<td></td>
<td>hij</td>
<td>Him/he</td>
<td></td>
<td>We</td>
<td></td>
</tr>
<tr>
<td></td>
<td>haar</td>
<td>She/her(/hair)</td>
<td></td>
<td>Ons</td>
<td>us</td>
</tr>
</tbody>
</table>

Table 6: Keywords in ANW and NLWaC
The classification into themes was undertaken by checking where and how the words were being used, using the Sketch Engine. The analysis shows that these two large, general corpora of Dutch have different strengths and weaknesses, and different areas that might be interpreted as overrepresentation or under representation (depending on one’s perspective.) The ANW has a much stronger representation of Flemish (the variety of Dutch spoken in Belgium). It has 20% fiction: keek (looked, watched) is used almost exclusively in fiction. It is 40% newspaper and newspapers talk at length about money (which also interacts with time and place: franks were the Belgian currency until 1999; also the units were small so sums in franks were often in millions or even billions). There is a particularly large chunk from the Meppel local newspaper. Most occurrences of Foto were in “Photo by …” or “Photo from …” and of Auteur, in newspaper by-lines, which might ideally have been filtered out. Daily newspapers habitually talk about what happened the day before, hence gisteren. Vorig and aldus (previous, thus) are fairly formal words that get used more in newspapers than elsewhere.

NIWaC has a large contingent of religious texts. It is based on Web texts, some of which could have been more rigorously cleaned to remove non-continuous-text and other non-words like URL components www, http, nl. The English might appear to be because we had gathered mixed-language or English pages but when we investigated, we found most of the instances of and and the were in titles and names, for example “The Good, the Bad and the Ugly”, where the film was being discussed in Dutch but with the title left in English. Perhaps modern global culture, with its tendency to use English in film, book and song titles, institution names and catch phrases, is better-represented in NIWaC than in ANW. Political history is also well-represented.

Finally we note that pronouns occur in both lists: third-person ones in the ANW list, and first and second person ones in the ANW list. This confirms the hypothesis discussed above and the evidence from Ferraresi et al (2008): Web-based methods as described in this paper give us the opportunity to access more interactional language than was possible for large traditional corpora.

4 FUTURE WORK
As discussed above in relation to Telugu, the method is sensitive to character encoding issues; both our knowledge of the encodings most often used for a language, and how the main search engines handle them. We are planning an extensive survey of these questions, in relation to Hindi, Telugu, Thai, Tibetan and Vietnamese.

We would like to prepare corpora for further languages. High on our priority list are all the official languages of India; Korean; Tibetan; and all the official languages of the European Union.

We would like to not only extract corpora, but also estimate how large the Web is for each language. This will interact with the character-encoding research.
In a parallel stream of work focusing on English we have developed a high-accuracy, scaleable, de-duplication method (Pomikalek et al 2009). We shall explore applying this method in the Corpus Factory.

The paper has mainly discussed the preparation of plain-text corpora. To set up the corpora for lexicography and language research, they should be accurately segmented, lemmatised and part-of-speech (POS) tagged; loaded into a corpus tool such as the Sketch Engine; and supplemented with a ‘Sketch Grammar’. Then, lexicographers and others can see ‘word sketches’, one-page summaries of a word’s grammatical and collocational behaviour. Word sketches have widely been found to be a good starting point for dictionary-writing (see eg Kilgarriff and Rundell 2002). But for this to be realised we need the language-specific tools. For segmenters, lemmatisers and POS-taggers we have often used open-source tools, for example SWATH for segmenting Thai, but for some languages and tasks, they do not exist: for example there are no pos-taggers for Vietnamese, or word-level segmenters for Tibetan. In these cases we are looking out for partners with computational linguistics expertise in the language, to work together on creating the tools. We want to work with people with those skills to prepare sketch grammars.

5 SUMMARY
The ‘corpus factory’ presents a method for developing large general-language corpora which can be applied to many languages. In this paper we have described the method, and how it has worked when we have applied it to five languages from different language families, each presenting different issues in terms of character encoding and orthography. We have produced a set of five large corpora. We think they are high-quality resources, better for lexicography than any others currently in existence for four of the five languages. (This is mainly because, except for Dutch, to the best of our knowledge, there are no other large general-language corpora available.) We have evaluated the corpora, as far as we were able given the lack of other resources for comparison. The corpora are available for use in a leading corpus tool. We believe the ‘Corpus Factory’ has a great deal to offer Asian (and other) lexicography in the years to come.

Acknowledgements
We would like to thank Avinesh PVS for his help on Hindi, Carole Tiberius for her help on Dutch, Giao Chi Le Thi for hers on Vietnamese, and John Hartmann for his on Thai.

6 REFERENCES


Notes

[1] Corpora for a number of languages are available through the Linguistic Data Consortium and the European Language Resources Association (ELRA). Sizes, text types, cost and legal constraints vary widely from corpus to corpus. Very large web corpora for English, French, German and Italian are available from the University of Bologna at [http://wacky.sslmit.unibo.it/doku.php](http://wacky.sslmit.unibo.it/doku.php). Large corpora are available for online searching at, amongst others,
- Leeds Univ. ([http://www.comp.leeds.ac.uk/ssharoff/](http://www.comp.leeds.ac.uk/ssharoff/)) Large web corpora for nine languages
- Leipzig Univ. ([http://corpora.uni-leipzig.de/](http://corpora.uni-leipzig.de/)) Newspaper corpora for 56 languages
- The Sketch Engine ([http://www.sketchengine.co.uk](http://www.sketchengine.co.uk)) A range of corpora, 15 languages
- VISL ([http://visl.sdu.dk](http://visl.sdu.dk)) A range of corpora, seven European languages
- Brigham Young Univ. ([http://corpus.byu.edu/](http://corpus.byu.edu/)) Well-structured and diachronic corpora for English, Spanish, Portuguese.

[2] Colleagues have used similar methods for Greek, Persian, Swedish and Hebrew. All of the corpora are, or will soon be, available in the Sketch Engine.

[3] Description of the tool can be found at [http://evanjones.ca/software/wikipedia2text.html](http://evanjones.ca/software/wikipedia2text.html)

[4] To clarify our terminology with an English example: *invade, invades, invaded and invading* are all words, or word forms, of the lemma *invade*. Lemmas are what are usually used as dictionary headwords whereas words or word forms are found in corpus text.

[5] Vietnamese word list can be downloaded at [http://www.informatik.uni-leipzig.de/~duc/software/misc/wordlist.html](http://www.informatik.uni-leipzig.de/~duc/software/misc/wordlist.html)


[7] We are aware that Vietnamese will sometimes be written informally using ASCII, so without diacritics, and that search engines typically allow searches without diacritics to match texts with them. Our method will usually miss pages without diacritics (or with non-standard encodings). We shall review the issue as part of the wider review of character-encoding issues discussed under ‘Future Work.


---

**Appendix**

*Table: Top 50 words in Web Frequency lists of different languages*

<table>
<thead>
<tr>
<th>Dutch</th>
<th>Hindi</th>
<th>Telugu</th>
<th>Thai</th>
<th>Vietnamese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Een</td>
<td>की</td>
<td>ఎ</td>
<td>นี่</td>
<td>của</td>
</tr>
<tr>
<td>in</td>
<td>से</td>
<td>సందర్భ</td>
<td>ได้</td>
<td>và</td>
</tr>
<tr>
<td>het</td>
<td>का</td>
<td>က</td>
<td>在</td>
<td>trong</td>
</tr>
<tr>
<td>van</td>
<td>को</td>
<td>రా</td>
<td>ที่</td>
<td>được</td>
</tr>
<tr>
<td>de</td>
<td>के</td>
<td>ฉัน</td>
<td>ของ</td>
<td>là</td>
</tr>
<tr>
<td>te</td>
<td>में</td>
<td>ສາ</td>
<td>ให้</td>
<td>có</td>
</tr>
<tr>
<td>op</td>
<td>पर</td>
<td>ถ้า</td>
<td>เป็น</td>
<td>cho</td>
</tr>
<tr>
<td>met</td>
<td>और</td>
<td>อยู่</td>
<td>ไป</td>
<td>với</td>
</tr>
<tr>
<td>en</td>
<td>है</td>
<td>है</td>
<td>จะ</td>
<td>một</td>
</tr>
<tr>
<td>English</td>
<td>Marathi</td>
<td>Chinese</td>
<td>Vietnamese</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>voor</td>
<td>भी</td>
<td>มหา</td>
<td>dâ</td>
<td></td>
</tr>
<tr>
<td>zijn</td>
<td>एक</td>
<td>Atat</td>
<td>nhung</td>
<td></td>
</tr>
<tr>
<td>dat</td>
<td>कि</td>
<td>ว่า</td>
<td>dể</td>
<td></td>
</tr>
<tr>
<td>die</td>
<td>इस</td>
<td>ไม่</td>
<td>không</td>
<td></td>
</tr>
<tr>
<td>aan</td>
<td>ने</td>
<td>การ</td>
<td>nguôi</td>
<td></td>
</tr>
<tr>
<td>als</td>
<td>कर</td>
<td>กับ</td>
<td>vè</td>
<td></td>
</tr>
<tr>
<td>om</td>
<td>ही</td>
<td>จาก</td>
<td>các</td>
<td></td>
</tr>
<tr>
<td>is</td>
<td>हो</td>
<td>ความ</td>
<td>vào</td>
<td></td>
</tr>
<tr>
<td>niet</td>
<td>नहीं</td>
<td>อูย</td>
<td>khi</td>
<td></td>
</tr>
<tr>
<td>er</td>
<td>यह</td>
<td>ด้วย</td>
<td>ó</td>
<td></td>
</tr>
<tr>
<td>door</td>
<td>लिए</td>
<td>แต่</td>
<td>cùng</td>
<td></td>
</tr>
<tr>
<td>ook</td>
<td>हं</td>
<td>นี้</td>
<td>denn</td>
<td></td>
</tr>
<tr>
<td>dan</td>
<td>करने</td>
<td>ทำ</td>
<td>nhiều</td>
<td></td>
</tr>
<tr>
<td>bij</td>
<td>तो</td>
<td>ถนน</td>
<td>tù</td>
<td></td>
</tr>
<tr>
<td>worden</td>
<td>किया</td>
<td>แล้ว</td>
<td>dó</td>
<td></td>
</tr>
<tr>
<td>uit</td>
<td>अपने</td>
<td>กี</td>
<td>trên</td>
<td></td>
</tr>
<tr>
<td>of</td>
<td>साथ</td>
<td>ซื้น</td>
<td>ra</td>
<td></td>
</tr>
<tr>
<td>naar</td>
<td>तक</td>
<td>อย่าง</td>
<td>nay</td>
<td></td>
</tr>
<tr>
<td>maar</td>
<td>गया</td>
<td>ถึง</td>
<td>phải</td>
<td></td>
</tr>
<tr>
<td>deze</td>
<td>जो</td>
<td>ยัง</td>
<td>nhur</td>
<td></td>
</tr>
<tr>
<td>over</td>
<td>रहे</td>
<td>ค้น</td>
<td>lai</td>
<td></td>
</tr>
<tr>
<td>dit</td>
<td>कुछ</td>
<td>มาก</td>
<td>mà</td>
<td></td>
</tr>
<tr>
<td>tot</td>
<td>रहा</td>
<td>ต้อง</td>
<td>sè</td>
<td></td>
</tr>
<tr>
<td>meer</td>
<td>बाद</td>
<td>ลิ้น</td>
<td>nhung</td>
<td></td>
</tr>
<tr>
<td>hebben</td>
<td>जा</td>
<td>โดย</td>
<td>còn</td>
<td></td>
</tr>
<tr>
<td>heeft</td>
<td>अपनी</td>
<td>ลีด</td>
<td>chi</td>
<td></td>
</tr>
<tr>
<td>nog</td>
<td>कोई</td>
<td>หรือ</td>
<td>mới</td>
<td></td>
</tr>
<tr>
<td>andere</td>
<td>दिया</td>
<td>นัน</td>
<td>sür</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>Kannada</td>
<td>Telugu</td>
<td>Urdu</td>
<td>Hindi</td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>--------</td>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td>zich</td>
<td>हुए</td>
<td>ఎందుకు</td>
<td>खा</td>
<td>hon</td>
</tr>
<tr>
<td>kunnen</td>
<td>होने</td>
<td>ہو</td>
<td>में</td>
<td>hàm</td>
</tr>
<tr>
<td>geen</td>
<td>था</td>
<td>దుర్కూర్</td>
<td>झिं</td>
<td>bi</td>
</tr>
<tr>
<td>wordt</td>
<td>करते</td>
<td>మంగు</td>
<td>outlet</td>
<td>thi</td>
</tr>
<tr>
<td>al</td>
<td>रही</td>
<td>తో</td>
<td>ห่าง</td>
<td>rāt</td>
</tr>
<tr>
<td>wat</td>
<td>पहले</td>
<td>3</td>
<td>द्वार</td>
<td>nām</td>
</tr>
<tr>
<td>zo</td>
<td>जब</td>
<td>ఏండి</td>
<td>फें</td>
<td>dang</td>
</tr>
<tr>
<td>kan</td>
<td>यह</td>
<td>370</td>
<td>तो</td>
<td>có thể</td>
</tr>
<tr>
<td>hun</td>
<td>किसी</td>
<td>93</td>
<td>पहुँच</td>
<td>theo</td>
</tr>
<tr>
<td>veel</td>
<td>कहा</td>
<td>188</td>
<td>री</td>
<td>khác</td>
</tr>
<tr>
<td>alle</td>
<td>अब</td>
<td>18800</td>
<td>हो</td>
<td>vi</td>
</tr>
<tr>
<td>wel</td>
<td>समय</td>
<td>3668</td>
<td>क्वार</td>
<td>cà</td>
</tr>
<tr>
<td>waar</td>
<td>या</td>
<td>900</td>
<td>छिय</td>
<td>vân</td>
</tr>
</tbody>
</table>
Figure 1: Concordances of ข้าพเจ้า from Sketch Engine.

1 แผนก ั้งำณี นี ข้าพเจ้า ข่อ กล่าว ธรรมิก
1 <p>ใน ที่สุด นี ข้าพเจ้า มี ความ ปิดินัก
19 ป้องกัน ไม่ ให้ ข้าพเจ้า กำจัด ศาสนาอาศิสต์
16 ว่า : " ใน ที่สุด ข้าพเจ้า จะ ล้าง แค้น จนถึง
28 เดือน ปี ปี จอ ข้าพเจ้า ลูก ขี้้น มา เดือน
33 บัดนั้น เอง ขณะที่ ข้าพเจ้า ปฏิบัติ จริง นั้น
35 ปังเกิด ขึ้น กับ ข้าพเจ้า กังขา สดหยาย
53 วา บรม ดนตรี ของ ข้าพเจ้า มีได้ ผลอย่างเป็น
9 ลำบาก ณั้น ตกลง แก่ ข้าพเจ้า ณั้น คงอยู่ ไม่
A CORPUS BASED STUDY OF URDU ENGLISH LEXICOGRAPHY

Abrar Hussain Qureshi
English Govt.College Mianchuunu, PAKISTAN
abrarqureshi74f@hotmail.com

Abstract: In the 21st century, Urdu has been strategically important enough to attract the attention of foreign learners and bilingual lexicographers. The foremost aim of bilingual lexicography is to strive to give equivalents of the source language in the target language without unnecessary information. This is not an easy task. In the more distant past, Urdu-English bilingual lexicographers seem, by and large, to have been under the impact of the work of their predecessors, occasionally improving on it but often perpetuating their omissions. After analyzing the existing Urdu English dictionaries by the author, in Pakistan, the conviction has been strongly developed that Urdu to English dictionaries are not a perfect guide to real meanings for the learners. This led the author to survey available Urdu English dictionaries to find out what had fundamentally gone wrong with these. One of the main reasons to be found was that these dictionaries were not corpus based and were traditionally inspired by translation. The paper presents the current situation in traditional dictionaries, pointing out how a lexical item changes its semantic behavior, when analyzed in a larger collection of text and encourages the Urdu bilingual lexicographers to compile corpus-based Urdu-English dictionaries.

Key Words: Urdu, lexicography, corpus

1. Research Methodology

A recent perspective - corpus based statistical approach - has been adopted here to investigate how speakers and writers exploit the resources of their language rather than looking at what is theoretically possible in a language, EMILLE corpus of Asian languages, prepared by Lancaster University, United Kingdom has been used in this regard. Though EMILLE corpus is not a big collection of Urdu spoken and written text as The Bank of English, prepared by Birmingham University United Kingdom yet it has been used as a point of departure, hoping that a large corpus of Urdu language will be compiled for linguistic purposes. In order to process the data, Wordsmith, a software prepared by the Oxford University has been used for concordances, collocations and frequency list.

2. Literature Review

The tradition of bilingual lexicography in the subcontinent starts with Joan Josua Ketelaar who compiled his book of Hindustani grammar (in Dutch) in the late 17th Century when he was working with the Dutch East India Company in India. The next development was by Capt. George Hadley who wrote the first grammar book for the officers of the East India Company in 1772, which contained English and Hindustani words. This was the prototype of a bilingual dictionary. He was soon followed by Capt. J. Fergusson who produced his first Dictionary of the Hindoostan language in two parts in 1773. This work, in his own words,
"contained a great variety of phrases, idiom, and to facilitate the acquisition of the language" (Hadley 1772, 1801).

Then came the era of military grammar books. The vocabulary in these early military grammars relates to the kind of language that an officer of the East India Company might need to know. Therefore it contained words that they (officers) learned from their sepoys (men under their command) (Friedlander 2006).

The most elaborate work in this field was undertaken by Dr. Fallon (1879). For the first time he included, not only the colloquial terms and ordinary words of day-to-day speech but also the refined language of women of high social status.

Dr. Abdul Haq produced the Standard English Urdu Dictionary in 1937. Dr. Haq was truly a Samuel Johnson of Urdu. His work was related with his predecessors, especially Dr Faollon's, He tried to purify the language. He also worked on many Urdu neologisms. Many subsequent English Urdu dictionaries were based on Haq’s dictionary.

Kitabistan's English-English-Urdu dictionary was produced by Prof. B.A. Qureshi in 1957. This was compiled according to the needs of learners. There were many definitions and expressions in it but it was not a big success.

Ferozsons, a famous publisher, produced their own English into English and Urdu dictionary in 1960's. This dictionary was modeled on Chambers’ Dictionary yet it could not meet the entire needs of the learners; as it lacked many phrasal verbs and idiomatic expressions.

In the last two decades of the twentieth century, several indigenous English-Urdu dictionaries were produced, mainly in Pakistan. These were largely abridgements of the older works directed at school and college students.

The next important development in the field of Urdu bilingual lexicography was Dr. Jameel Jalibi,s Qaumi dictionary that was first published in 1992. It got inspiration from Webster’s encyclopedic dictionary 1986 and seemed to be an Americanized dictionary. In spite of being a Qaumi dictionary, it could not reflect the national mood, thoughts and social life.

Prof. Kalimuddin produced a six volume dictionary under his editorial guidance. It was named Jāmi) and completed in the late 1970's, though published between 1994 and 1998. There were translations for approximately 250,000 words and phrases.

OUP Karachi published the English Urdu dictionary in 2003. Its compiler Mr. Shanul Haq Haqee (d 2005) was a scholar and translator. This work (henceforth called Oxford Urdu) was partly based on the eighth and ninth editions of the famous Concise Oxford Dictionary (1990-1995 ed). It took him a long time to complete and covered nearly 125,000 word and idiomatic expressions.

It is a fact that the content and structure of a bilingual dictionary changes according to the needs of users. Thus English to Urdu dictionary for a native Urdu speaker is meant for a purposeful reading of the text in English, whereas English to Urdu dictionary for English speakers may be used for a better communication in the Urdu language. Therefore, the design
features in the two types of English to Urdu dictionaries will be different because of the different needs of the various users. In the author's opinion there are subtle differences between bilingual dictionary-making and pure translations. Although closely related, the two are not quite the same; literal translations may not encompass the different connotations that a word may carry, especially when the same word is used in different contexts. English Urdu lexicographers have sometimes fallen prey to this trap.

3. Data Analysis

Let us try to establish the fact that the traditional Urdu bilingual lexicography is at a loss to meet the needs of the learners at different levels. For this purpose “Urdu English Dictionary” (n.d.) published by Ferozsons has been chosen for analysis.
For example, the entry چاند chand has been defined as “the moon; a target, a white spot in the forehead of cattle; name of an ornament; n.f. the crown of the head.”
Now we will discuss some semantic issues related to this entry. Let us start with the first meaning that is the physical moon that shines in night. That is the strongest meaning of "moon" in the Urdu language. We endorse this viewpoint with the help of even EMILLE, a small corpus of Urdu language that shows that the strong collocates of "moon" are the words that talk about "moon" as a physical entity.

Now let us consider the rest of the meanings of the entry “moon”. The rest of the meanings are “a target, a white spot in the forehead of cattle; name of an ornament; n.f. the crown of the head”. It becomes quite clear after the evidence of EMILLE corpus that the rest of the meanings are not in use in the Urdu language.
Rather we find that some other connotative meanings of the entry چاند “moon” are used in the Urdu language for example, چاند. “moon” is often used metaphorically for a beautiful person or it is often used for beloveds of both sexes e.g. loving son, daughter, brother, sister, etc. The following concordance from EMILIE corpus shows how the word چاند “moon” is connotatively related with the above mentioned
4. Concluding Remarks

The value of corpus has established over the years due to the validity of electronic research in linguistics. This will create new opportunity to the Urdu linguistics in Pakistan as well. Utilization of Urdu corpus in language study especially in Urdu bilingual lexicography, will yield great results to discard old methods and theories of lexicography. Corpus based Urdu lexicography will modify old practices of lexicography and introduce an empirical model in this field. As a result, there is a dire need for the compilation of various representative corpora to use them as reliable databases.

Finally the author of this article notes with deep concern that as there is no formal research in Urdu linguistics, especially corpus linguistics with a few exceptions, the authorities in Pakistan and in India should collaborate their efforts in these neglected areas. Hence large scale efforts for Urdu corpus is the need of the hour and we hope that this research paper would encourage the efforts to build a big corpus of Urdu on the model of The Bank of English.
ACKNOWLEDGEMENT

I am indebted to Nick Groom of Birmingham University U.K. and Zafar Iqbal of BZ University Pakistan for providing unwavering direction, and enthusiastic inspiration with his knowledge of corpus and its software that led us to the final version of this research paper.

References


Contact Author:

Abrar Hussain Qureshi, Lecturer In English in Govt.College Mianchunnu (Pakistan). A Phd scholar in BZ university Multan (Pakistan)

abrarqureshi74f@hotmail.com
Currently in Japan, there are no “specific” instructions on teaching of dictionary use given in the official guidelines for school teaching laid down by the Ministry of Education, Culture, Sports, Science and Technology. However, there have been calls for more systematic teaching of dictionary skills at secondary school, and an increasing number of studies into dictionary use have been carried out. This paper, therefore, aims to review those studies and to establish the direction we need to take in order to enable introduction of systematic training of dictionary use into the secondary-school syllabi.

In this study, I will focus on empirical studies, which may fall into the six areas of research into dictionary use out of seven Tono (2002: 396) identifies: (1) the attitudes, needs, habits and preferences of dictionary users; (2) text or word comprehension; (3) text or word production; (4) vocabulary learning; (5) dictionary-related performance in testing; and (6) teaching dictionary skills. The past studies will be also analyzed in terms of their methods employed in the framework of Tono (2002).

I will then discuss in what areas and in what kinds of forms further research is to be conducted in order to get a clearer, comprehensive picture of the current situation of dictionary use and to identify specific skills to teach at each stage of learning at school.
Abstract: Kipfer (2007) claims “… dictionary companies report that sales of print dictionaries have not declined. Print dictionaries, …, are not going away anytime soon”. This does not correspond well to the Hong Kong situation, and probably to most major Asian cities. Most local students, if they use a dictionary at all, will be using either an electronic or internet dictionary. What is alarming, however, is that the kinds of internet dictionaries that are popular among students are those attached to search engines like Yahoo, and the information that students look up mainly is the Chinese equivalents of the English words in question. Computing technology might serve students well in terms of speed and convenience. It has not, however, been able to guide them well in learning vocabulary or solving linguistic problems with effective use of the dictionary information obtained, nor in choosing a dictionary which is relevant to their levels and to meet their needs. English teachers play a key role in helping students utilize and discriminate between the vast resources that computing technology offers them on the Internet. In this paper, I will examine if Hong Kong teachers of English possess the knowledge and skills to teach dictionary use to their students. I will also suggest assistance which lexicographers and publishers can offer to help English teachers gain a better understanding of their work partner (English dictionaries). This union will bring benefits to their mutual clients – learners of English.

Key Words: Dictionary use; Teacher training; EFL dictionaries

1. INTRODUCTION

Local Hong Kong students are required to purchase their own set of textbooks for classes in school. On every booklist, there is always a recommended English dictionary. For this reason, most students in Hong Kong would have bought an English learner’s dictionary during their study at school and their choice would probably be the one selected by the teachers of the school.

However, although most students own an English learner’s dictionary (either monolingual or bilingualised), it is generally observed that many are accessing the Internet for dictionary information. The most popular dictionary among students of Hong Kong is in fact the one attached to the search engine Yahoo, instead of an online English learner’s dictionary. Moreover, the dictionary information that students frequently look up is the Chinese (students’ mother-tongue) equivalents of the English words in question.

There are two interesting points here worth pondering: first, after seven years of school training and use of paper English learners’ dictionaries, many local students have turned to a
search engine for dictionary information; second, the kind of information that students are interested in looking up is restricted mainly to mother-tongue equivalents of the English words in question, which I will consider to be a sign of failure in the development of pedagogical lexicography, especially in the area of dictionary use teaching.

In this paper, I will examine if Hong Kong teachers of English possess the knowledge and skills to teach dictionary use to their students. I will also suggest assistance which lexicographers and publishers can offer to help English teachers gain a better understanding of their work partner (English dictionaries). There is a need to strengthen the bond between these two partners if learners are to derive benefit from them.

2. STUDENTS’ USE OF ENGLISH DICTIONARIES

Kipfer (2007) claims “… dictionary companies report that sales of print dictionaries have not declined. Print dictionaries …, are not going away anytime soon”. This does not correspond well to the situation in Hong Kong, and probably to situations in most major Asian cities. Most Hong Kong students, if they use an English dictionary at all, will be accessing the Internet for dictionary information. This view is supported by results obtained from a questionnaire given to 85 year one students studying at a university in Hong Kong in February 2009. The survey data reveal that 48% of the students would opt to access an Internet dictionary to search for meanings of the word they need, and 41% on how to use the word in question in writing, when they were given a range of choices of reference tools ranging from a paper dictionary to a mobile phone glossary. Only 12% of the students would choose a paper dictionary to look up word meaning, and 26% for information on how to use the word in writing.

The Internet dictionaries that most students claimed that they were using, however, were in fact dictionaries attached to search engines like Yahoo. And in the questionnaire, students expressed that they mainly looked up Chinese equivalents of the English words in question (60%). I believe many EFL learners do not know how to use an English dictionary to assist their learning of the language. With limited linguistic knowledge of the English language, and probably basic training in vocabulary learning, it is rather natural for these users to restrict their dictionary search to looking up the equivalents of the English words in their first language. For such a simple request, they are well served by free online dictionaries provided by search engines.

The general proliferation of websites that provide dictionary help, and the magnetism of a search engine dictionary for word look-up to students, as in the case presented here, should be of concern to both English teachers and pedagogic lexicographers. Computing technology might serve EFL students well in terms of speed and convenience while searching for English words in question. It is doubtful, however, that it can guide them well in learning vocabulary, or solving linguistic problems with effective use of the dictionary information obtained, nor in choosing a dictionary which is relevant to their levels and to meet their needs.

Moreover, if students’ habits of depending heavily on mother-tongue equivalents to learn about words, or, to learn new words in English are not challenged and changed, their English learning will surely be compromised. In Fan’s (2000) study which aimed at investigating the look-up behaviour of bilingualised dictionaries of over 1,000 Hong Kong tertiary students, she found that “the more students looked up the Chinese equivalents of the English words,
the more they would ignore other kinds of information in the dictionary, including English definitions” (131). Also, these students would miss out on the vast resources and assistance that learners’ dictionaries have prepared for them, which are in fact within their reach because most students in Hong Kong own at least one of such dictionaries (Chi, 2003).

English teachers play a key role to help students utilize and discriminate between the vast English learning resources that computing technology offers. They are also the perfect instructors to teach students how to use dictionaries to assist their learning of the foreign language since English teachers possess both the linguistic knowledge and teaching skills required for the task. How ready are our English teachers in Hong Kong to assume this role? In the following, I will examine if Hong Kong teachers of English possess the knowledge and skills to teach dictionary use to their students.

3. TEACHER TRAINING ON DICTIONARY USE

Findings of a survey conducted in 2001 (Chi, 2003) with 170 teachers of English at secondary (high) schools in Hong Kong showed that these teachers held strong convictions in the role dictionaries played in their students’ learning. All of them gave affirmative answers when they were asked whether they thought dictionaries could help students to learn English. To consolidate the validity of this view, the same question was given to 30 secondary school teachers in March 2008, and 93% of them gave the same positive answer.

However, almost half (45.5%) of the 170 teachers in the 2001 survey answered that they did not receive any training in teaching students how to use English dictionaries to assist their learning. Those who reported that they had received some form of training said they learned about the English learners’ dictionaries through attending dictionary publishers’ seminars or workshops, and that the training was in general incidental.

It is evident that dictionary publishers have been rendering assistance to English teachers on how to use their English dictionaries. There are dictionary workbooks with exercises explaining to users how to use the dictionaries. For teachers, worksheets with lesson plans with similar aims are made ready and accessible for free on many dictionary websites. Sometimes, these worksheets are sent to ELT teachers via emails free of charge. For example, Oxford University Press ELT marketing department sends emails with the subject line “ready-made activities from ELT dictionaries” to member teachers on a regular basis. Also, Oxford University Press (HK) has sent dictionary editors and English trainers to local secondary schools to give lessons to students on how to use their latest edition of the bilingualised Oxford Advanced Learner’s Dictionary.

However, the kind of help available is generally restricted to the introduction of a particular dictionary with exercises or activities offered that are rather specific or tied to some particular promoting features. I postulate that what teachers really need are the fundamentals of English learners’ dictionaries – an introduction of the characteristics of English learners’ dictionaries; the various features offered by different learner’s dictionaries available in the market; and why these features would help to meet the particular needs of EFL learners. With such knowledge, English teachers should then be offered training on how to teach their students with an aim to enable them to refer to a variety of learners’ dictionaries to meet different linguistic needs.
If English teachers want to read about dictionary use teaching or to receive systematic training on dictionary use, unfortunately it is not easy to find literature or a course which provides such training. First, most of the existing research studies on pedagogical lexicography have focussed on EFL learners, rather than on teacher training. Dictionary use research studies are sporadic and limited to individual teachers’ or scholars’ personal teaching experience (Kipfer, 1985; Tono, 1989; Nesi, 1994; Bogaards, 1998b; Fan, 2000; Chi, 2003) of a particular group of students. Due to variables like the diversity of learners’ linguistic and cultural backgrounds and their levels of English proficiency; the nature and objectives of the teaching conducted in each case; the typology of dictionaries being examined and so on, these research studies have yet to offer English teachers a clear framework that illuminates how dictionary use should be taught.

Second, formal courses on lexicography, particularly on dictionary use teaching, are rare. After a quick search on courses offered by universities in Hong Kong, I found three courses with the title related to dictionary study and they are “Computational Lexicography” (see note 1) and “Bilingual Lexicography” (see note 2). All three courses target university students who are, or who want to be trained as, translators or linguists. None of the courses includes dictionary use training or teaching in its course objectives online.

4. TEACHERS’ KNOWLEDGE OF ENGLISH LEARNERS’ DICTIONARIES

If EFL teachers have not learned about English learners dictionaries during their professional training, we could assume that their knowledge of the books is obtained from the following three sources:

1) Teachers’ own initiatives to learn about English dictionaries through self-study, attending seminars and visiting dictionary websites.
2) Teachers’ conventional wisdom acquired through years of teaching and exposure to teaching materials and reference books.
3) Teachers’ own experience in using dictionaries to learn English.

All these pathways of acquiring knowledge of the dictionary stretch over a long period of time. Any learning that might have taken place would have been laborious, non-systematic and unreliable, mostly depending on individuals’ intrinsic interest and perception in using dictionaries to learn English. If the English teachers are native speakers of English, their learning experiences with the kind of dictionaries that they consulted might not be pertinent to the needs of their EFL students.

Based on my knowledge and experience with local teachers, I will assume that most of them will use their own experience of a particular dictionary as the base of recommendation to their students. In the following, I will like to use the character Susan to illustrate how a local teacher would have acquired knowledge of a dictionary through years of using it. I will argue that the knowledge gained will be inadequate and unreliable for the teacher to use as the basis for advice on dictionary use to their students.

Let me introduce Susan, our English teacher from Hong Kong. She is forty-five years old and has been teaching English in a local secondary school for twenty-three years. She was 12 in 1975 when she entered secondary school and bought her first dictionary as required by the school. At that time, the only dictionary for learners of English available in the market was
Oxford Advanced Learner’s Dictionary of Current English 3rd edition, published in 1974 (OALD 3). Subsequently, Susan entered the University of Hong Kong studying for a bachelor degree majoring in history. It was in the year 1982 and Susan was 19 years old. To prepare for her study at the university, Susan wondered for a moment if she would need a new dictionary. The choices of dictionaries for advanced learners available in the market at that time included the OALD 3, and Longman Dictionary of Contemporary English first edition, 1978 (LDOCE 1). It would be another seven years before OALD 4 (1989) was released. Without much knowledge of the publisher Longman, nor the advantages of the new LDOCE 1 over OALD 3, Susan decided to stay with her OALD 3, and saved some money.

When Susan started teaching world history in a local secondary school three years later (1985), she was also asked to teach English language to complete her teaching load of 30 lessons per week. The dictionary recommended to students in this school was the OALD 3. Upon learning of the OALD 3 being the required dictionary Susan was happy, since she had been using the dictionary for her own study and had some knowledge of it.

Now we are in the year 2009 and Susan has worked in the same school for twenty-three years. Her OALD 3 is still on her shelf for a rather sentimental reason – the dog-eared book reminded her of the days when she was a school girl and a novice teacher, though she seldom uses it these days. In fact, she hardly uses a dictionary because schoolwork and lesson preparation will not pose any challenge requiring her to turn to one for help. Should she need a simple and straightforward explanation for her students, she could use the glossary attached to the teacher’s handbooks of most ELT English textbooks.

Yet, Susan still insists that her students check all the words they do not know with the help of the dictionary recommended to them before they start a new unit of the textbook. She shares the same conviction as those English teachers who were subjects of the teachers’ surveys previously mentioned in this paper. These teachers believed that their students would benefit from the use of English dictionaries to assist their learning of English. Susan will not check, however, if her students have done their homework or not, or if they have any question regarding the word search. The English classes have a tight teaching schedule and Susan never feels she has time for dictionary work. Students should do the work for their own benefit and they should know how to use a dictionary, Susan reassures herself. Nobody taught Susan how to use her dictionary at school and why would her students need any help from her. Susan also makes it compulsory for her students to have the dictionary on their desk when they have their writing class every other week.

When Susan flips through the OALD 7 on her student’s desk during one lesson, she can see that the dictionary has used a second (blue) colour to display headwords and some selected features, while the rest of the information is still printed in black ink. It was a good idea to give the dictionary some colours. It does help to make the book look more vibrant and user friendly to school children, Susan thinks to herself.

Yet, Susan would not know the reason why OALD 7 is using two different colours to display information and the impact of such changes on its users. OALD only started using two colours to display information in its seventh edition (2005) but for a reason not explained in the dictionary. Unknown to users, another significant change of the dictionary that was not explained was the adoption of a defining vocabulary from their fifth edition (1995) onward. When LDOCE 1 first appeared in 1978, its major distinct and much appraised feature from
the then existing OALD 3 was that it adopted a controlled set of vocabulary to write
definitions. Although OALD 4 (1989) was published after LDOCE 2 (1987), Cowie (chief
editor of OALD 4) chose not to follow in the footsteps of LDOCE and did not adopt the same
convention in his edition. Cowie (1999) argued that simplicity and comprehensibility are not
“the only properties that definitions written within a controlled vocabulary need to possess. It
is also important that they should be accurate, concise and written in natural English, and
LDOCE provides some evidence that, in striving to be simple and comprehensible, compilers
can sometimes lose sight of one or more of the other criteria” (111).

However, a defining vocabulary of 3,500 words was adopted by Crowther (chief editor of
OALD 5) to write definitions in OALD 5. There was no information given in OALD 5 to
explain or justify to users why such a major change was made, nor on how an addition of
1,500 words in the controlled vocabulary has helped OALD compilers to resolve problems
that Cowie had raised regarding LDOCE’s defining vocabulary of 2,000 words. The change
was simply announced in the dictionary Preface (1995):

Another new feature is the introduction of a 3500-word vocabulary within
which definitions are written. The words … were chosen principally
according to their frequency in the language, … but also as a ‘core’
vocabulary of real value to students of English. The size of the list was
determined by the minimum requirement for producing definitions in
natural English that are both accurate and easy to understand.

What if Susan caught the phrase “defining Vocabulary of 3,500 words” in the blurb on the
back cover of OALD 5 and turned to Appendix 10 of the dictionary to read for more
information? She would still be unclear as to the rationale behind this major deviation from
the earlier editions of the dictionary since this is what she would have read:

In order to make the dictionary definitions easy to understand, we have
written them using only the words in the following list. The words in the
list were carefully chosen according to their frequency in the language and
their value to students as a ‘core vocabulary’ of English. …

Changes made to a new edition of an existing dictionary are often underpinned by linguistic
and users’ research findings at the time, and these changes reflect major lexicographical
principals and ideologies of the compilers. It would thus be fair to suppose that when OALD
5 deviated from the previous editions of the dictionary and adopted a defining vocabulary to
write definitions, the compilers of this edition believed that they had made progress in
dictionary making. Presumably, with the support of corpus evidence, compilers must have
held the conviction that the new definitions would better serve their users. From users’
perspectives, this major change and its perceived positive impact on the users should have
been explicitly explained and publicised in the dictionary to help them understand the value
of the book in assisting them to learn English. With limited coverage on the change being
made in the dictionary, it is doubtful that general users of OALD editions prior to the 5th
would have noticed the new defining convention. Indeed, it is well acknowledged that not
many dictionary users have the habit of reading the Preface of a dictionary, nor comparing
the old edition of a dictionary with its latest version. Susan probably would have overlooked
this major shift in writing definitions in OALD 5, which continues on in OALD 7 (changed
to the Oxford 3000™), and her knowledge of the dictionary remains from the experience she obtained when she used OALD 3 as a student.

As a matter of fact, the OALD 3 that Susan owns contains features that were not found in the earlier two editions. For example, a number of photographs were used for illustrations and bold sanserif typeface for headwords and sub-headwords was used, which Cowie (1999) claimed to offer more legible reading and ease of reference compared to the two older editions. But Susan would not have known that these features were new to OALD 3 because she probably would not have had the previous 2 editions (1947, 1963) of the dictionary for comparison. What Susan, like any other average users of a dictionary would have assumed instead is that the features found in OALD 3 are standards for all English dictionaries.

5. CONCLUSION

Due to the length limitations of this paper, I have highlighted only several changes in two editions of a single sample of English learners’ dictionaries. Nonetheless, I would like to establish that Susan’s knowledge and experience of the various editions of the OALD characterize the general understanding of a learner’s dictionary among locally trained English teachers of Hong Kong. From the illustration of Susan, it is clear that her personal experience of using the OALD 3 as a student would be insufficient as a base for her to give advice to her students regarding what the current OALD 7 can offer to help them learn English.

At a glance, the basic information categories like definitions and examples have remained more or less the same in most English learners’ dictionaries since the first appearance of such dictionary typology almost 60 years ago. A careful examination of the current dictionaries, however, will reveal to users how profoundly computing technology and research findings on dictionary users have impacted the compilation of such dictionaries. For example, the use of a large corpus has significant bearing on the quality, selection criteria and presentation of dictionary definitions and examples alike. If English teachers like Susan want to understand and keep abreast of the development of learners’ dictionaries with respect to how these dictionaries can support their students’ learning of English, they need to be trained systematically. For such a course, I propose the following topics to be included:

<table>
<thead>
<tr>
<th>Topics</th>
<th>Knowledge of such topics will help teachers to:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dictionary typology</td>
<td>- advise their students on the kinds of dictionaries that will meet their linguistic needs</td>
</tr>
<tr>
<td>Macrostructure and microstructure of a dictionary</td>
<td>- train students to speed up the look-up process</td>
</tr>
<tr>
<td>Pedagogical rationale behind the use of a defining vocabulary, a convention used by most English learners’ dictionaries</td>
<td>- ease the concern of students who are reluctant to read English definitions because they believe that they will not understand the words used in writing definitions. As a result, many of such students opt to read only the translation equivalents</td>
</tr>
<tr>
<td>Collocation information</td>
<td>- teach students to use the information to produce idiomatic English phrases or sentences</td>
</tr>
</tbody>
</table>
The concept and use of word frequency - teach students how to build up their vocabulary profile using the information

Although English learners’ dictionaries are widely used by students and teachers of English, education and training on how to use them have not yet drawn much academic attention. It is unfortunate and perplexing that systematic teaching of pedagogical lexicography with regard to the teaching of English as a second or foreign language is not forthcoming from academia. In most language institutions and universities worldwide, pedagogic lexicography, or indeed, lexicography, is not treated as an academic subject. While we await the day when pedagogical lexicography assumes a formal academic status to be taught generally at tertiary institutions, publishers and lexicographers can help to resolve some of the problems identified. The training course proposed could be offered as a standalone enhancement course for veteran and novice teachers alike, or it could be offered as a component of an EFL/SSL teacher training course. To this end, dictionary publishers, pedagogical lexicographers, and various lexicographical associations would be the natural sources of support.

In this paper, I cautioned that Hong Kong students will require guidance in their selection of dictionaries and ways of using them to assist their English learning. However, local teachers of English may not possess the latest knowledge of English learners’ dictionaries in playing the advisory role. They may also require training on how to help their students use dictionary information to learn English. It is advised that teachers take courses on pedagogical lexicography, particularly on the teaching of dictionary use. Such courses could be supported by dictionary publishers, pedagogical lexicographers and various lexicographical associations. If English learners’ dictionaries want to maintain their relevance or strengthen their bond with users for years to come, teacher training is the way forward.

Notes:
(1) City University of Hong Kong, Department of Chinese, Translation and Linguistics (undergraduate level)
University of Hong Kong, Linguistics Department (graduate level)

(2) Chinese University of Hong Kong – Translation Department (graduate level)

REFERENCES


Dictionaries cited:

TOWARDS A BETTER PERSPECTIVE IN THE SELECTION OF ENTRIES FOR AN ENGLISH DICTIONARY OF FINANCE FOR INDONESIAN STUDENTS

Deny Arnos Kwary
Aarhus University, DENMARK
deak@asb.dk

Abstract: Dictionaries covering English and Indonesian have become increasingly important in the business sector and in theoretical lexicography over the last few decades. However, only few LSP (language for specific purposes) dictionaries exist and no lexicographical literature discusses the principles for making dictionaries that provide optimal help to Indonesian users. This paper describes the preparation of an English-Indonesian dictionary of finance. The dictionary is aimed at assisting Indonesian students who are preparing to take the CFA (Chartered Financial Analyst) exam in understanding financial texts written in the English language. The focus of the paper is on the selection of entries. The paper starts with a description on how to use computerized methods in selecting specific terms. It then continues with the implementation of the modern theory of lexicographical functions in selecting the entries. The discussion rounds up with a better perspective in the selection of entries for LSP dictionaries, in particular for an English-Indonesian dictionary of finance.

Key Words: LSP dictionary, Range, Concordance, TermoStat, lexicographical function, text reception, Chartered Financial Analyst, factual competence, linguistic competence.

1. INTRODUCTION

The significant role of dictionaries to assist learners when reading foreign language texts has been noted in several studies (cf. Baxter 1980, Cook 2001, Nation 2001, and Wingate 2002). In Indonesia, the needs for dictionaries are also apparent among learners, especially those at the higher education level. The needs are even more evident for LSP dictionaries since learners often have problems in comprehending terms which they commonly encounter when they are reading their textbooks. However, up to now, only few LSP dictionaries exist in Indonesia and there is no lexicographical literature which discusses the principles for making LSP dictionaries that provide optimal help to Indonesian users. This paper discusses one of the most important lexicographical considerations in making LSP dictionaries, i.e. the selection of entries.

In this paper, the terms LSP lexicography and terminography are not differentiated. This is in line with the opinion put forward in Bergenholtz and Nielsen (2006: 281-282) that there is no dividing line between LSP lexicography and terminography, because both have LSP dictionaries as the object. The discussion on the selection of entries in this paper focuses on a dictionary of finance, but it may also be applicable for other LSP dictionaries.

2. USING COMPUTERIZED METHODS IN THE SELECTION OF TERMS

Several studies have proposed the use of computerized or automatic methods in selecting terms or entries for LSP dictionaries. These methods are believed to lighten the burden of lexicographers in the selection of terms. In this section, the methods are grouped into...
three categories, i.e. Vocabulary Classifications, Key Word Analysis, and Term Extractions. The description and the evaluation of each of them are given in the following sub-sections.

2.1. Vocabulary Classifications
Since the proposed LSP dictionary is intended for second language learners, it is reasonable to take into consideration some of the findings from Applied Linguistics in selecting the entries for the dictionary. Nation (2001: 11-13) distinguishes between four kinds of vocabulary in a text: high frequency words, academic words, technical words, and low-frequency words. These classifications are the basis for the computer program, called Range, which can be used to create word lists based on frequency and range. There are three wordlists available in the Range program as stop lists. The first and the second stop lists—each consists of 1000 word families—are considered the lists of high frequency words. These word families, called the Base One Families and Base Two Families, are based on the General Service List of English Words from Michael West (1953). The third stop list contains 570 headwords found in the Academic Word List from Coxhead (2000). To see how the Range program works, it is necessary to choose a particular text to run. In this paper, the text is taken from the CFA (Chartered Financial Analyst) textbook, Study Session 8, Book 3, Level 1. The Range results for this text is shown in Table 1.

<table>
<thead>
<tr>
<th>WORD LIST</th>
<th>TOKENS/%</th>
<th>TYPES/%</th>
<th>FAMILIES</th>
</tr>
</thead>
<tbody>
<tr>
<td>one</td>
<td>15990/67.41</td>
<td>837/46.09</td>
<td>493</td>
</tr>
<tr>
<td>two</td>
<td>1404/ 5.92</td>
<td>170/ 9.36</td>
<td>104</td>
</tr>
<tr>
<td>three</td>
<td>2846/12.00</td>
<td>378/20.81</td>
<td>217</td>
</tr>
<tr>
<td>not in the lists</td>
<td>3479/14.67</td>
<td>431/23.73</td>
<td>?????</td>
</tr>
<tr>
<td>Total</td>
<td>23719</td>
<td>1816</td>
<td>814</td>
</tr>
</tbody>
</table>

Table 1: Range Results

Table 1 shows that 55.45% (46.09 + 9.36) of the word types in the text are high frequency words, 20.81% of them are academic words, and 23.73% are either technical words or low frequency words. The Range program seems useful to lessen the burden of going through too many words manually to determine the technical words. In this case, for example, it is more convenient to go through 431 words instead of 1,816 words, to select the technical words. The problem with this method, as also realized by Nation (2001: 18), is that some technical vocabulary actually also occurs in the high frequency words. Consider an extract of the result of the Base One Families, shown in Table 2.

<table>
<thead>
<tr>
<th>BASE ONE FAMILIES</th>
<th>RANGE</th>
<th>TYFREQ</th>
<th>FAFREQ</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>THE</td>
<td>1</td>
<td>1539</td>
<td>1539</td>
<td>1539</td>
</tr>
<tr>
<td>OF</td>
<td>1</td>
<td>862</td>
<td>862</td>
<td>862</td>
</tr>
<tr>
<td>AND</td>
<td>1</td>
<td>587</td>
<td>587</td>
<td>587</td>
</tr>
<tr>
<td>IN</td>
<td>1</td>
<td>507</td>
<td>507</td>
<td>507</td>
</tr>
<tr>
<td>TO</td>
<td>1</td>
<td>403</td>
<td>403</td>
<td>403</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>372</td>
<td>527</td>
<td>527</td>
</tr>
<tr>
<td>FLOW</td>
<td>1</td>
<td>266</td>
<td>305</td>
<td>305</td>
</tr>
<tr>
<td>FROM</td>
<td>1</td>
<td>241</td>
<td>241</td>
<td>241</td>
</tr>
<tr>
<td>FOR</td>
<td>1</td>
<td>223</td>
<td>223</td>
<td>223</td>
</tr>
<tr>
<td>AS</td>
<td>1</td>
<td>208</td>
<td>208</td>
<td>208</td>
</tr>
<tr>
<td>BE</td>
<td>1</td>
<td>202</td>
<td>1271</td>
<td>1271</td>
</tr>
</tbody>
</table>
The extract shown in Table 2 consists of the words considered to be high frequency words. The problem is that some of them may be technical vocabulary. For example, the word stock in the last row in Table 2, has different meanings in a general sense and in the financial domain. Therefore, such words should not be stopped or be included as general words. Instead, they should be put in the technical vocabulary word list. In this case, it is clear that the Range program is essentially ineffective in assisting LSP lexicographers in the selection of terms. To solve this problem, Nation (2001: 18) suggested comparing the frequency of words in a specialized text with their frequency in a general corpus. The efficacy of this suggestion is discussed in the next sub-section.

2.2. Key Word Analysis

Nation’s proposal for solving the problem mentioned in the previous sub-section, can be addressed by using Key Word Analysis programs, such as WordSmith and AntConc. Since WordSmith is not a freeware, the Key Word analysis program used in this paper is run by AntConc which can be downloaded for free.

Scott (1997: 236) defined a key word as “a word which occurs with unusual frequency in a given text.” Based on this definition, a key word does not always mean a word with high frequency. The actual calculation of keyness of a word is done using the chi-square statistic to determine its outstandingness. If the occurrence of a word is outstandingly frequent in a target corpus than in a reference corpus, it will be considered a positive key word. All words which fall into that category are ordered in terms of their relative keyness.

As mentioned previously, the target corpus is created from the CFA textbook, Study Session 8, Book 3, Level 1. This target corpus is to be compared with a reference corpus. In this paper, the reference corpus comes from the British Academic Written English (BAWE) corpus. This reference corpus contains 6,506,995 words. The corpus is available free of charge to researchers who agree to the conditions of use and who register with the Oxford Text Archive.

An extract of the result of the key word analysis of the CFA textbook using AntConc program is presented in Table 3.

<table>
<thead>
<tr>
<th>NO.</th>
<th>FREQUENCY</th>
<th>KEYNESS</th>
<th>WORD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>534</td>
<td>4445.302</td>
<td>cash</td>
</tr>
<tr>
<td>2</td>
<td>336</td>
<td>2191.156</td>
<td>income</td>
</tr>
<tr>
<td>3</td>
<td>268</td>
<td>1459.710</td>
<td>flow</td>
</tr>
<tr>
<td>4</td>
<td>174</td>
<td>1244.679</td>
<td>net</td>
</tr>
<tr>
<td>5</td>
<td>179</td>
<td>1233.703</td>
<td>assets</td>
</tr>
<tr>
<td>6</td>
<td>173</td>
<td>1036.930</td>
<td>statement</td>
</tr>
<tr>
<td>7</td>
<td>129</td>
<td>1019.977</td>
<td>expense</td>
</tr>
<tr>
<td>8</td>
<td>152</td>
<td>1008.284</td>
<td>operating</td>
</tr>
<tr>
<td>9</td>
<td>131</td>
<td>960.956</td>
<td>shares</td>
</tr>
</tbody>
</table>
As shown in Table 3, the Key Word analysis provides a better indication of the terms related to CFA context, than the Range analysis. It is because it does not stop the general words if they are outstandingly frequent in the target corpus. For example, the word expense is stopped in the Range program, but it is not stopped in the AntConc program, because it is an important term in finance. However, there is still a problem with the results shown in the Key Word list, i.e. the analysis only results in single-word terms. This is a problem in determining the entries for a dictionary of finance, because most financial terms are multi-word terms. Take for example, the first four words from the key word list. It is more common for CFA candidates to look for the definitions of multi-word terms, such as cash flow and net income, rather than the single-word terms, i.e. cash, income, flow, and net. Therefore, the use of key word analysis alone is evidently inadequate to capture the entries which should be selected for an LSP dictionary such as a dictionary of finance. The problem with multi-word terms has been tried to be solved by terminologists. The efficacy of one of the methods used by terminologists in dealing with multi-word terms is discussed in the next sub-section.

2.3. Term Extraction
Terminologists have developed several Term Extraction programs to enable the extraction of terms automatically. One of the programs which is available for free is TermoStat. The online version of TermoStat can extract terms in French, English, Spanish and Italian. The program does not only extract single-word terms but also multi-word terms which have high keyness in the text. The program can also run a Log Likelihood test, which gives a better estimate of keyness. Table 4 shows an extract of the result of the Log Likelihood test on the CFA text using TermoStat.

<table>
<thead>
<tr>
<th>Fréquence</th>
<th>Candidat lemmatisé</th>
<th>Variantes orthographiques</th>
<th>Poids</th>
</tr>
</thead>
<tbody>
<tr>
<td>161</td>
<td>cash flow</td>
<td>cash flow</td>
<td>1704.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cash flows</td>
<td></td>
</tr>
<tr>
<td>132</td>
<td>asset</td>
<td>assets</td>
<td>1343.77</td>
</tr>
<tr>
<td>171</td>
<td>firm</td>
<td>firm</td>
<td>1258.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>firms</td>
<td></td>
</tr>
<tr>
<td>82</td>
<td>income statement</td>
<td>income statement</td>
<td>917.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>income statements</td>
<td></td>
</tr>
<tr>
<td>92</td>
<td>net income</td>
<td>net income</td>
<td>868.16</td>
</tr>
<tr>
<td>76</td>
<td>liability</td>
<td>liabilities</td>
<td>755.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>liability</td>
<td></td>
</tr>
<tr>
<td>112</td>
<td>cash</td>
<td>cash</td>
<td>742.87</td>
</tr>
<tr>
<td>71</td>
<td>inventory</td>
<td>inventory</td>
<td>702.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>inventories</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>revenue</td>
<td>revenues</td>
<td>678.39</td>
</tr>
<tr>
<td>64</td>
<td>dividend</td>
<td>dividends</td>
<td>590.11</td>
</tr>
<tr>
<td>67</td>
<td>equity</td>
<td>equity</td>
<td>576.24</td>
</tr>
<tr>
<td>93</td>
<td>share</td>
<td>share</td>
<td>541.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>shares</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>expense</td>
<td>expenses</td>
<td>539.94</td>
</tr>
<tr>
<td>50</td>
<td>balance sheet</td>
<td>balance sheet</td>
<td>517.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>balance sheets</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>activity</td>
<td>activities</td>
<td>400.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>activity</td>
<td></td>
</tr>
</tbody>
</table>
The results shown in Table 4 is consistent with the fact that financial texts contain more multi-word terms than single-word terms. For example, the multi-word term ‘cash flow’ occurs 161 times, while the single word cash only occurs 112 times. It shows that CFA candidates are more likely to find the word cash in combination with the word flow, i.e. cash flow than the word cash by itself. This seemingly good solution, however, still has a problem because that it can only identify multi-word terms which consist of nouns. This is a problem because other word classes may be a part of multi-word terms, and are the terms found in financial texts. Take an example the multi-word term cost of goods. Based on the concordance, from 38 hits for cost of goods, 35 of them are cost of goods sold and the other three are cost of goods available. If the lexicographers of a dictionary of finance only relies on the results of TermoStat, they may decide to include cost of goods as one of the entries. This decision does not only create redundancy as it is unlikely to be searched, but also creates a problem since it cannot be defined well in financial sense. Consequently, it will exclude the relevant terms from being selected.

The emphasis on nouns is assumed reasonable because most terms are nouns. However, it should be realized that other word classes, such as verbs, may also be terms. Take an example, the verb recognize. In the Oxford Advanced Learner’s Dictionary (seventh edition), there are five definitions of recognize, i.e.

1. [vn] recognize sb/sth (by / from sth) to know who sb is or what sth is when you see or hear them, because you have seen or heard them or it before.
2. recognize sth (as sth) to admit or to be aware that sth exists or is true.
3. recognize sb/sth (as sth) to accept and approve of sb/sth officially.
4. [vn] be recognized (as sth) to be thought of as very good or important by people in general.
5. [vn] to give sb official thanks for sth that they have done or achieved.

None of those definitions are suitable for the verb recognize when it is used in financial texts because ‘to recognize an amount or an item in the financial statements is to include that amount or item as an asset or a liability in the balance sheet or as income or expense in the income statement.’ (www.regnskabsordbogen.dk/regn/gbgb/gbregn.aspx). The focal role of verbs in CFA exams is clearly indicated by the fact that a list of verbs, called command words, is placed in the website of CFA under the course of study link. In this case, it is clear that a verb may also have a particular meaning in a certain domain. Therefore, it is necessary to include, not only nouns, but also other word classes in the selection of entries for an LSP dictionary.
3. USING THE MODERN THEORY OF LEXICOGRAPHICAL FUNCTIONS IN THE SELECTION OF ENTRIES

The discussion in the previous section clearly demonstrates that LSP lexicographers should not rely heavily on computerized or automatic selection of terms. Placing the computer programs and texts as the starting point in the selection of terms is evidently inadequate. Such methods are not acceptable in the point of view of the modern theory of lexicographical functions. Based on this perspective, a dictionary should be considered as a utility tool which serves one or more function(s) and is created for one or more user group(s). Therefore, the considerations about the function and the users should be the starting points. The following sub-sections describe the implementation of the modern theory of lexicographical functions in the selection of entries for the LSP dictionary.

3.1 Dictionary Function and Intended User Group

According to Bergenholtz and Tarp (2003: 176), for communication oriented functions, a dictionary can be used to assist the users in solving problems related to:

- text reception in the native language
- text production in the native language
- text reception in a foreign language
- text production in a foreign language
- translation of texts from the native language into a foreign language
- translation of texts from a foreign language into the native language

From the six functions mentioned above, the proposed English-Indonesian dictionary of finance serves the third one, i.e. to assist users in solving problems related to text reception or understanding of texts in a foreign language. It is the fact that the users—CFA candidates—should read some materials—financial texts—to prepare for the CFA exams. The users are Indonesians, while the texts are in English, a foreign language. Therefore, the users may have some problems when reading the texts, and they need a utility tool—a dictionary—to assist them in comprehending the foreign language texts. Knowing that the function of this dictionary is to help with text reception, it is reasonable if only written texts are chosen for the corpus, and spoken language is excluded. Consequently, setting the lexicographical function before conducting further lexicographical work will make the work of the lexicographers more focused.

After the main function of the dictionary has been established, it is necessary to draw up the profile of the intended users. Nielsen (2002) mentions two main competences which need to be considered in creating the profile of the intended user group, i.e. factual competence and linguistic competence. Based on the factual competence, which is also called encyclopaedic competence, the users can be divided into experts, semi-experts (such as students and experts from related areas), and lay people (Tarp 1995: 21). Whereas, for the linguistic competence, the users can be classified into high, intermediate, and low.

For this dictionary of finance, the intended users are Indonesian students of CFA programs who have completed their undergraduate programs from a faculty of economics or school of business in Indonesia. These users may be categorized as semi-experts in terms of factual competence. However, since the factual competence is related to the language, it is necessary to classify the users into a narrower competence category. For the factual competence in Indonesian language, the users are between experts and semi-experts or in the upper-semi-expert category. This is because they have become familiar with financial topics in Indonesian language during their undergraduate programs as most of the textbooks are written in Indonesian. On the other hand, for the factual competence
in English language the users are between semi-experts and lay people or in the lower-
semi-expert category. This is because most of the textbooks are not in English and the
lecturers only occasionally use English terms in their lectures. The use of English
financial terms during the lectures at Indonesian universities is due to the fact that some
lecturers graduate from the United States and most of the textbooks are the Indonesian
translations of American textbooks.

For the linguistic competence, the classification is based on the English language
proficiency of the users. As graduates from universities, the users have learned English
for at least seven years, i.e. three years in junior high school, three years in senior high
school, and one year at the university. It is common in Indonesian universities that the
students are required to obtain a certain TOEFL (Test of English as a Foreign Language)
score before they graduate. Some universities require the students to obtain at least 450 in
the Paper-Based Test and other universities require 500. TOEFL scores are categorized
into three, i.e. high, intermediate and low. Scores 310 – 485 fall into Low category, and
scores 485 – 560 fall into Intermediate category. Based on the information from the
TOEFL website, test takers who receive a score at the Low level typically understand
some of the information presented in academic texts in English that require a wide range
of reading abilities, but their understanding is limited. On the other hand, test takers who
receive a score at the Intermediate level typically understand academic texts in English
that require a wide range of reading abilities, although their understanding of certain parts
of the texts is limited. Based on the linguistic competence, the users are between the Low
and Intermediate category or at the lower-intermediate level.

By knowing the factual competence and linguistic competence of the users,
lexicographers will be able to provide lexicographical solutions which are more suitable
for the intended user group. For instance, in dealing with some English abbreviations, the
lexicographers may simply provide the expansion of the abbreviated forms rather than
providing both the expansion and the equivalents. Consider the following text taken from
the CFA 1 Book 3, Study Session 8, page 61.

_The old standard, APB No. 20, provided for the cumulative effect of the accounting
change to be reported in the income statement, below the line, net of tax._

In dealing with the abbreviation, APB, the lexicographer can just provide the expansion,
i.e. Accounting Principles Board, rather than providing both the expansion and the
Indonesian equivalents. This is because the users are at the lower-intermediate level in
linguistic competence, so they know the meanings of accounting, principles, and board. In
addition, considering their factual competence in finance, they will also be able to infer
the type of organization or the board referred to in the text.

The considerations involving the function of the dictionary and its intended user group
should always be included in making other lexicographical decisions, including those
mentioned in the following sub-sections.

3.2. Classification of the Subject Field
The classification of the subject field has to be carried out to ensure that the dictionary
only contains the terms that are relevant to financial texts and to ensure that the needs of
the users to solve their text reception problems are met. One possible approach to classify
the subject field is by using the libraries’ universal decimal classification (UDC) and
existing special bibliographies (Pedersen 1995: 84). Another approach is by scrutinizing
special textbooks on the subject. In the case of CFA, a textbook which can be considered
is *The Analysis and Use of Financial Statements*, by White, Sondhy and Fried, third edition. 2003. It is the book which has obtained permission to reprint the CFA Examination questions. The following is the classification systemized based on the topics discussed in the textbook.

1. Principal Financial Statements
   1.1 Income Statement
   1.1.1 Net sales
   1.1.2 Alliance revenues
   1.1.3 Cost and operating expenses
   1.1.4 Other income
   1.2 Balance Sheet
   1.2.1 Assets
   1.2.2 Liabilities
   1.3 Cash Flow Statement
      1.3.1 Cash Flow from Operating Activities (CFO)
      1.3.2 Cash Flow from Investing Activities (CFI)
      1.3.3 Cash Flow from Financing Activities (CFF)

Starting from the classification of the subject field and its following activities, e.g. selection of texts and selection of entries, the lexicographer needs to work together with subject field specialists, i.e. financial experts in this case, to ensure that most important aspects in the subject field are included.

### 3.3 Selection of Texts

Based on the classification given in Section 3.2, the lexicographer can determine the texts which are relevant for the subject field, i.e. financial statements, which are to be included in the corpus. In this case, the texts that form the basis for the selection of the entries are from the following types of references and textbooks:

- Textbooks and learning materials recommended by the CFA Institute.
- Financial Accounting Standards (FAS) from the FASB (Financial Accounting Standard Board).
- Financial statements from American companies.
- American textbooks which discuss the income statements, balance sheets and cash flow statements.

Since the users are preparing to take the CFA exams, it is reasonable to select the textbooks and learning materials recommended by the CFA Institute to be the main sources for the corpus. Considering that the CFA Institute is an American organization, it is also reasonable to select FAS to be included in the corpus because FAS contain the principles which are applied in the financial statements of American companies. In relation to this, the financial statements from American companies can be included to enrich the corpus. Lastly, referring to the classification of the subject field and in order to obtain a larger corpus, American textbooks which discuss the income statements, balance sheets and cash flow statements should also be included.

On the other hand, the texts selected as the basis to determine the Indonesian equivalents and explanations of the English terms may be obtained from the following types of references and textbooks:

- *Pernyataan Standar Akuntansi Keuangan (PSAK ‘Indonesian financial accounting standard statements’)* from *Dewan Standar Akuntansi Keuangan (DSA K ‘Indonesian financial accounting standard board’).*
• Financial statements from Indonesian companies.
• Indonesian translation of American textbooks which discuss the income statements, balance sheets and cash flow statements.
• Indonesian textbooks which discuss the income statements, balance sheets and cash flow statements.
• Existing English-Indonesian dictionaries covering financial terms.
• Existing Indonesian dictionaries covering financial terms.

For the purpose of determining the Indonesian equivalents of the English terms, PSAK is a good source because it is the Indonesian equivalent of the FAS. Another good source is the financial statements from Indonesian companies which may contain similar terms as those found in the financial statements from American companies. In addition, some American textbooks which discuss the income statements, balance sheets and cash flow statements have been translated into Indonesian, and they are good sources for this purpose. However, since Indonesian translation qualities are still questionable, some Indonesian textbooks which discuss the income statements, balance sheets and cash flow statements should also be taken into account.

The existing English-Indonesian dictionaries covering financial terms can be used to select the equivalents. Nevertheless, those dictionaries should be used with great caution, because the equivalents are not always correct or may refer to another meaning in Indonesian context. Therefore, existing Indonesian dictionaries covering financial terms also need to be consulted when determining the Indonesian equivalents of the English terms. If the equivalent has an additional and different meaning in Indonesian context, the lexicographer may need to put a usage note in the dictionary article. Take for example the term fiscal which should not be simply provided with the Indonesian equivalent 'fiskal'. In the English language, fiscal ‘pertains to money, especially government taxation and spending policies’ (investorwords.com, accessed on 5 May 2009). In Indonesian context, fiskal has at least two meanings; the first is the same as that in the English language; and, the second refers to an amount of money paid by an Indonesian who is about to travel to another country but he/she cannot present his/her NPWP (taxpayer identification number). The English equivalent for this type of fiskal is ‘exit tax’, which is an obsolete term used to mean ‘a tax imposed on emigrating citizens by the former Soviet Union’ (encarta.msn.com, accessed on 5 May 2009).

3.4 Selection of Entries
After the relevant texts have been selected based on the subject field delimitation, a corpus that forms the basis for selecting the entries for the dictionary can be built. As explained in Section 2, lexicographers should not only rely on computer programs to select the entries. The selection should also take into account the special terms and other words relevant to the subject field and to the needs of the users. Both single-word terms and multi-word terms should be included in the entries if they are relevant. Nouns and other word classes should also be included if they are relevant.

During the selection of entries, it is essential to make an internal subject classification which further specifies the contents of the sub-fields so that more terms which are relevant can be selected (Pedersen 1995: 85-86). For example, item 1.2.1 Assets, under the subject field classification, can be further divided into smaller parts, so that all terms related to Assets are also selected. From the examination of the relevant texts selected previously, it is possible to know that Assets can be divided into two, i.e. Fixed Assets and Current Assets. In addition, each of them can be split up into smaller components.
The following is an example of an internal subject classification of assets commonly found in the balance sheets of American companies.

1. Assets
  1.1 Fixed Assets
     1.1.1 Intangible Fixed Assets
     1.1.2 Tangible Fixed Assets
     1.1.3 Other Fixed Assets
  1.2 Current Assets
     1.2.1 Inventories
     1.2.2 Debtors
     1.2.3 Other Current Assets
     1.2.4 Cash & Cash Equivalents

An even more detailed classification, called the terminological classification (Pedersen 1995: 89) can also be carried out to achieve an optimal selection of the entries. Nielsen (2002) states that the terminological classification will ensure an in-depth coverage of the structure and terminology of a subject field. The following is an example of a terminological classification for 1.2.1 Stocks, under the internal subject classification.

1.2.1 Inventories
  1.2.1.1 Raw Materials
  1.2.1.2 Work in Process
  1.2.1.3 Finished Goods
  1.2.1.4 Inventory Prepayments

All of the classifications mentioned above are important in selecting the entries for a dictionary, especially for an LSP dictionary. The lexicographers are often not the specialists of the subject field, so they may not realize that some terms are missing from the entries put in the dictionary. In some cases, even if the lexicographer is a specialist in the subject field, there is still a risk of excluding a term which is related to other terms put in the dictionary. Take for example the Key Terms in Linguistics written by Howard Jackson (2007). Although Jackson has been teaching language and linguistics for over thirty-five years, he still misses some entries which are actually related to other entries. Kwary (2009) shows that in the Key Terms in Linguistics there are entries for future perfect and past perfect in the dictionary, but there is no entry for present perfect. Another example mentioned in the review is in the section on Phonetics and Phonology, where Jackson mentions Acoustic Phonetics, but omits the other two branches, i.e. Articulatory Phonetics and Auditory Phonetics.

The risk of excluding entries which are relevant and related to other entries is even higher for lexicographers who are not experts in the subject field of the LSP dictionary. Therefore, all of the classifications explained above should be carried out meticulously by the lexicographers, preferably together with subject-field specialists, in order to ensure that all relevant terms are included in the entries of the dictionary.

4. CONCLUSION
The selection of entries for LSP dictionaries is one of the most important aspects in creating the dictionaries. Careful considerations is necessary throughout the process. Using the computerized methods or programs such as Range, AntConc, and TermoStat may help in lightening the burden of the lexicographers during the selection process.
because they help to reduce the number of words to be scrutinized. However, as demonstrated in this paper, simply relying on those methods may lead to defective decisions in the selection of the entries for the LSP dictionary. The implementation of the modern theory of lexicographical functions has demonstrated a better result in the selection process, as shown in the case for the English Dictionary of Finance for Indonesian students who are preparing for the CFA exams. Therefore, the perspective described in this paper should always be taken into account by lexicographers who want to create an LSP dictionary, especially an English Dictionary of Finance for Indonesian students.

REFERENCES

(15) www.antlab.sci.waseda.ac.jp/software/antconc3.2.1w.exe
(16) www.cfainstitute.org/cfaprog/courseofstudy/commandwords.html
(17) www.coventry.ac.uk/researchnet/d/911 “British Academic Written English (BAWE) corpus was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).”
“TermoStat was developed by Patrick Drouin (2003) at the Observatoire de linguistique Sens-Texte, Department of linguistics and translation, University of Montreal, Canada.”
A LARGE PRONUNCIATION DICTIONARY FOR THAI SPEECH PROCESSING

Patcharika Chootrakool, Chai Wuttiwiwatchai, Krit Kosawat
Human Language Technology Laboratory, NECTEC, THAILAND
patcharika.cho@nectec.or.th, chai.wut@nectec.or.th, krit.kos@nectec.or.th

Abstract: This paper reports the design and development of a machine-readable pronunciation dictionary for the Thai language. LEXiTRON-Pro is useful for Thai speech processing as well as other related language processing areas. It is a list of words along with their associated pronunciations in the form of phoneme-symbol sequences. Derived from the International Phonetic Alphabet (IPA), a phoneme set, which was particularly defined in NECTEC, contains 21 phonemes of single initial consonants, 17 phonemes of cluster consonants, 24 vowels and 5 tones. Phoneme symbols are a computerized version of IPA symbols. We also created rules for deriving the phoneme transcription separated by syllables in order to avoid ambiguous pronunciation. Distinct words were selected from various sources of texts based on their frequently use. At present, there are over 130,000 word entries, composed of common words and named entities of person, place, and organization. The current usage of this dictionary is for training a letter-to-sound conversion tool, which has been used for building speech recognition corpora and used as a text processing part of text-to-speech synthesizer. The design, development process, application review, statistical analysis of dictionary, problem discussion, and future work are described in detail.

Key Words: pronunciation dictionary, Thai speech processing

1. INTRODUCTION

Pronunciation dictionary in common sense would be dictionary with audio sound of the word. However, the pronunciation dictionary in our sense means a collection of words associated with their pronunciations in the form of phone sequences. Phones could be derived from standard sound representatives such as International Phonetic Alphabets (IPA) or Speech Assessment Methods Phonetic Alphabets (SAMPA) (1). Figure 1 illustrates a sample part of CMUDICT (2), an open-source pronunciation dictionary of US English developed by Carnegie Mellon University. The left-most column in the sample dictionary denotes word entries, whereas the rest columns, separated by the space, are their corresponding pronunciations. In the CMUDICT, phones used to represent word pronunciations are defined with computerized symbols. There are 39 distinct phones excluding their variation to the level of stress, which can be 0 for non-stress, 1 for primary stress, and 2 for secondary stress.

This kind of dictionary is very important for many current spoken language processing tasks. A letter-to-sound conversion (LTS) module takes the pronunciation dictionary as a primary source of knowledge to convert any textual word to its corresponding phone sequence. The LTS is one major component in Text-to-speech synthesis (TTS), which facilitates people to listen to synthesized voice instead of reading. In Automatic speech recognition (ASR), the LTS module plays an important role in building the phone transcription of speech corpus given speech orthographies. The LTS system itself can be more useful in other related areas such as improving the search engine to be able to search by sounds, i.e. Soundex search. It is also very helpful for foreigners to understand how to
read text written in another language if the defined phones are international standard. In this manner, the output of LTS is a good source to convert further to transcriptions in another language, i.e. transliteration.

| ABBREVIATE     | AHR BRI YV I Y EY TAH D |
| ABBREVIATED    | AHR BRI YV I Y EY TAH D |
| ABBREVIATED(2) | AHR BRI YV I YEY TAH D |
| ABBREVIATES    | AHR BRI YV I Y EY TAH D |
| ABBREVIATING   | AHR BRI YV I Y EY TIH N |
| ABBREVIATION   | AHR BRI YV I Y EY SH AH N |
| ABBREVIATIONS  | AHR BRI YV I Y EY SH AH N Z |
| ABRUZZESE      | AHR BRI YV I Y EY SH AH N Z |
| ABSS           | AE B Z |
| ABBY           | AE B I Y |
| ABCO           | AE B KOW |
| ABCOTEK        | AE B KOW TEH K |
| ABDALLA        | AE B D AE LAH |
| ABDALLAH       | AE B D AE LAH |

Figure 1: An example of CMUDICT US English pronunciation dictionary

Although the Thai speech processing technology has been researched for a long time (Wutiwiwatchai and Furui, 2007), its advancement is limited by the difficulty of Thai natures including none of syllable, word, or sentence boundary marker in the written script, and a number of exceptions in letter and sound relationship. To overcome these problems, a list of possible Thai words associated with their pronunciations is very crucial. This kind of dictionary might already be built and integrated in some existing speech processing engines but none of them has been mentioned separately and in details. Moreover, by a decade of our experience in speech processing research, there have been continuously requirements in building pronunciation dictionaries in various domains such as traffic information, telephone conversation, newspaper, etc. These dictionaries have then been developed separately and thus there exist a lot of overlapping words among different dictionaries. It is hence advantageous to gather all words and their pronunciations from these existing dictionaries, forming a single dictionary containing all unique words. New words found later can then be compared with this dictionary to see whether they are already included. Only words not existing in the dictionary are then passed to the process of transcribing and appending to the dictionary. This process can not only avoid repeatedly transcribing existing words, but also can help enlarging the dictionary systematically.

This paper aims to report the development of LEXiTRON-Pro, a large-scale Thai pronunciation dictionary built by the National Electronics and Computer Technology Center (NECTEC) as an open-source infrastructure for future Thai speech and language processing areas. The Thai language will be briefly described in Section 2. The building process and the specification of LEXiTRON-Pro are shown in Section 3. Interesting statistics and problem discussion are in Section 4. Section 5 presents some example of using the LEXiTRON-Pro in our research. Finally, Section 6 will discuss about our future work and conclusion.
2. THE THAI LANGUAGE

The Thai script has 44 consonants and 24 vowels including short, long vowels and diphthongs. The Thai sound system can be derived in the format /Ci-V-(Cf)-T/, where Ci denotes an initial consonant, V a vowel, Cf a final consonant which is optional, and T a tone. The design of phone set is shown in the Tables below. Table 1 shows the detail of Thai consonants mapping to phonetic symbols, both initial and final consonants. Table 2 shows Thai double consonant and their mapped phones. Table 3 shows phones that are used in foreign words mixed in Thai scripts. Table 4 shows Thai vowels and mapped phones. More details can be found in Kasuriya et al. (2003).

Table 1: Thai consonant phoneme mapping

<table>
<thead>
<tr>
<th>Consonant</th>
<th>Phoneme</th>
<th>Consonant</th>
<th>Phoneme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial (Ci)</td>
<td>Final (Cf)</td>
<td>Initial (Ci)</td>
</tr>
<tr>
<td>ก,ข,ฃ,ค,ฅ,ฆ</td>
<td>k</td>
<td>k^</td>
<td>บ</td>
</tr>
<tr>
<td>ธ,น,บ,ป,พ,ภ,ผ</td>
<td>ng</td>
<td>ng^</td>
<td>ผ,พ,ก</td>
</tr>
<tr>
<td>จ,ฉ,ช,ช,ช,ศ,ษ,ส</td>
<td>c</td>
<td>t^</td>
<td>ร</td>
</tr>
<tr>
<td>ฉ,ฌ,จ,จ,จ</td>
<td>ch</td>
<td>t^</td>
<td>ส,ฟ,ส</td>
</tr>
<tr>
<td>ช,ศ,ษ,ส</td>
<td>s</td>
<td>t^</td>
<td>ว</td>
</tr>
<tr>
<td>ญ,ย,ญ,ย</td>
<td>j</td>
<td>j^</td>
<td>ฮ,ย,ฮ</td>
</tr>
<tr>
<td>ร,ร,ร</td>
<td>d</td>
<td>t^</td>
<td>ผ,พ,ผ</td>
</tr>
<tr>
<td>ต,ท,ต,ท,ท,ธ,ธ</td>
<td>t</td>
<td>t^</td>
<td>น,พ,น</td>
</tr>
<tr>
<td>ธ,น,บ,ป,พ,ภ,ผ,พ,พ</td>
<td>n</td>
<td>n^</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Thai Double Consonant Phoneme Mapping (12 phonemes)

<table>
<thead>
<tr>
<th>Double Consonant</th>
<th>Phoneme Symbol</th>
<th>Double Consonant</th>
<th>Phoneme Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>ป,ร</td>
<td>pr</td>
<td>กร</td>
<td>kr</td>
</tr>
<tr>
<td>ป,ล</td>
<td>pl</td>
<td>กล</td>
<td>kl</td>
</tr>
<tr>
<td>พร</td>
<td>phr</td>
<td>กว</td>
<td>kw</td>
</tr>
<tr>
<td>พอ</td>
<td>phl</td>
<td>คร</td>
<td>khr</td>
</tr>
<tr>
<td>ดร</td>
<td>tr</td>
<td>คล</td>
<td>khl</td>
</tr>
<tr>
<td>ทร</td>
<td>thr</td>
<td>ขว</td>
<td>kwh</td>
</tr>
</tbody>
</table>

Table 3: Foreign Language Consonant Phoneme Mapping (9 phonemes)

<table>
<thead>
<tr>
<th>Final Consonant (Cf)</th>
<th>Phoneme Symbol</th>
<th>Double Initial Consonant (Ci)</th>
<th>Phoneme Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>ฟ</td>
<td>f^</td>
<td>บร</td>
<td>br</td>
</tr>
<tr>
<td>ส</td>
<td>s^</td>
<td>บล</td>
<td>bl</td>
</tr>
<tr>
<td>ช</td>
<td>ch^</td>
<td>พร</td>
<td>fr</td>
</tr>
<tr>
<td>ล</td>
<td>l^</td>
<td>ผล</td>
<td>fl</td>
</tr>
<tr>
<td></td>
<td></td>
<td>คร</td>
<td>dr</td>
</tr>
</tbody>
</table>
Table 4: Thai Vowels Phoneme Mapping (24 phonemes)

<table>
<thead>
<tr>
<th>Tongue Height</th>
<th>Front (short/long)</th>
<th>Central (short/long)</th>
<th>Back (short/long)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Close</td>
<td>i, ii (อิ, อี)</td>
<td>v, vv (อิ, อี)</td>
<td>u, uu (อ, อ)</td>
</tr>
<tr>
<td>Mid</td>
<td>e, ee (เอ, อเอ)</td>
<td>q, qv (เอก, อเอก)</td>
<td>o, oo (โอ, อโอ)</td>
</tr>
<tr>
<td>Open</td>
<td>x, xx (แอะ, แอะ)</td>
<td>a, aa (อะ, ออา)</td>
<td>@, @@ (เอาะ, อัว)</td>
</tr>
<tr>
<td>Diphthongs</td>
<td>ia, iia (เอียะ, อีย)</td>
<td>va, vva (เอียะ, อัว)</td>
<td>ua, uua (เอียะ, อัว)</td>
</tr>
</tbody>
</table>

![Diagram](image)

Figure 2: An overall development process

3. BUILDING PROCESS
The overall development procedure is illustrated in Figure 2. Details are as follows.

- **Text and word sources**
  We collected words from existing word lists. An original word list contains Thai named entities including road names, point of interests, and person names. A requirement to construct a ASR engine that can recognize such names led to the development of this large pronunciation dictionary. The second source was from existing dictionaries; mostly from the LEXITRON dictionary (3) which has been developed for public use and research on speech and language processing in NECTEC. The last source of words was from a variety of text articles such as newspaper, novel, magazine, etc.

- **Word extraction**
  As mentioned, word lists that are selected to add into the dictionary are collected from variety of sources. In the case of text articles, we have to extract a word list from them. The process starts by parsing the text by automatic word segmentation tools, in our case the SWATH (Meknavin et al., 1997) and a Conditional random field (CRF) based tool (Haruechaiyasak et al., 2008). These automatic tools are still far from perfect, said none of them can produce more than 95% word segmentation accuracy. Therefore, a further process of human checking is required. In checking, we have some regulations to determine which text segment should or should not be a word unit to be included in our dictionary. Brief regulations are as follows.

  - Words are segmented into small unit but some words that frequently used together as compound words. For example, although words “ป้อม” (the mouth of spring),
“น้ำ” (water), “ร้อน” (hot) can be a word unit by themselves, their compounding “ป้องน้ำร้อน” (the mouth of hot spring) is defined as another word unit as the they are often occurred together.

- The named entity including person names, location names, and organization names is a word unit since this dictionary has been constructed for the purpose of ASR and these named entities are often required in developing the ASR.

- The abbreviation is always entered as one word unit in the dictionary and is transcribed directly as spelling, not its original word.

It is noted that recently, NECTEC has launched a word segmentation guideline through the Benchmarks for enhancing the standard of Thai language processing (BEST) (4). The guideline has been designed based on minimalism that allows segmenting text into the smallest meaningful word. A compounded word might not be segmented if the meanings of its word components are largely distorted from its original. Word definition in this work is however not aligned with the one provided by BEST for some reasons. First, we are more relaxing in defining the word as we would like to contain almost all words Thai people often use even those words are actually a combination of smaller words. This criterion could be considered similar to that defined in constructing the CMUDICT. Second, named entities mostly combining several words inside are defined as a word unit in our dictionary for sake of convenience in constructing ASR.

Extracted words from any source are stored in a word list, which is then compared with words in the LEXiTRO-Pro dictionary. Words existing in the dictionary will be ignored, whereas the rest words are further processed by letter-to-sound conversion and human checking.

- Letter-to-sound conversion (LTS) and human checking
Words not existing in the LEXiTRO-Pro are considered new words to be entered in the dictionary. Each new word is passed to one of automatic LTS tools built in NECTEC (Tarsaku et al., 2003; Thangthai et al., 2006). The tool produces the pronunciation representative of the word in the form of a phone sequence with syllable segmentation and syllabic tone marked. It also marks syllable boundaries in the textual word.

Figure 3 demonstrates the output from the LTS tool. The first column is the word list tagged by “|” for syllable boundaries and the same symbol is used to separate syllables in the word’s pronunciation. It is noted that the Thai script contains Thai alphabets arranging by spelling rules. Arrangement is not straightforward as /C-V-Cf-T/ in the sound system. Some vowels can be placed in front of, above or below the initial consonant, som even surrounding the initial consonant.

A special case is the drop of the alphabet of a short vowel “a” and hence there exists only an initial consonant alphabet of the first syllable connecting in front of or behind the second syllables e.g. “ชmetic” /kh-a-z^1/m-oo-j^0/ or “บงฑา” /b-e-n^0|c-a-z^1/. In more special case, the vowel character of the second syllable can even come before the initial consonant of the first syllable. An example is the word “เจริญ” which actually contains 2 connected syllables /c-a-z^1/ and /r-qq-n^0/. In this word, a part of the vowel of the second syllable “i” is placed in front of the initial consonant “จ”. Another special case is sharing between the final consonant of the first syllable and the initial consonant of the second syllable. An example is the word “อัฐิ” /z-a-t^1|th-i-z^1/, which spends the consonant “ฐ” together. In these cases, it is impossible to clearly separate these two syllables in their textual form. Therefore, we design to combine these multiple syllables as one special syllable unit. The
pronunciation of this kind of syllable will contain multiple syllables separated by a space as an example shown in the third line of the Figure 3. This concept of textual syllable boundary is actually derived from the work by Aroonmanakun (2005). Since the automatic LTS tool can be error prone, linguists are involved in this step to re-check and correct the resulting pronunciation.

![Figure 3: LTS outputs and the LEXiTRON-Pro format](image)

- Adding into the LEXiTRON-Pro dictionary

The LEXiTRON-Pro dictionary is simply an ASCII plain text file containing the word list with textual syllable boundary tagged and the corresponding pronunciations of words in the list in the form of phone sequences with syllable segmented accordingly. Currently, the format of the dictionary is indeed as simple as the output of LTS tools illustrated in the Figure 3. New words passed from the LTS and human checking process is appended in this dictionary.

4. LEXI TRON-PRO

4.1 Statistics of the current LEXiTRON-Pro

Table 5 summarizes interesting statistics found from the current LEXiTRON-Pro dictionary. Figure 4 illustrates the distribution of word length in terms of the number of syllables.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of word entries</td>
<td>133,627</td>
</tr>
<tr>
<td>Min. word length (syllables)</td>
<td>1</td>
</tr>
<tr>
<td>Max. word length (syllables)</td>
<td>16</td>
</tr>
<tr>
<td>Average word length (syllables)</td>
<td>2.8</td>
</tr>
</tbody>
</table>

![Figure 4: The distribution of word length in terms of the number of syllables](image)
4.2 Problem discussion

Major problems are found in the process of LTS as discussed below.

- Problems of Thai compound words

As described in the Section 3, some compound word in Thai must be pronounced jointly as two syllables even its written form seems to be one syllable. For example, the word “ภรรยา” is transcribed as /kr-i-t^3-1 s-a-z^3-1\n-aa-z^3-4/. This kind of words is often error prone when transcribed automatically by the LTS tools.

- Problems of homographs and allophones

Although there appears not often, the homograph is an important problem in Thai. An example is such as the word “พระ” which can be either of two meanings with different pronunciations, /pl-a-w^3-0/ (leggs) or /ph ee-z^3-0|-aa-z^3-0/ (time). Another case related to the homograph is the same word having two alternative pronunciations. For example, the word “ภรรญา” (wife) can be pronounced as /p-a-n^3-0|-aa-z^3-0/ or /p-a-n^3-0 l-a-z^3-3|-aa-z^3-0/. In this case, we insert multiple entries of the word into our dictionary and mark the word with its alternative labels e.g. ภรรญา^1 and ภรรญา^2. The allophone is another problem when some phones in the word can be more than one sound, such as the initial consonant “ทร” can sound either /thr/ and /sr/.

- Problems of foreign words

The tone level in transliterated words trends to be distorted from their written forms. It is still hard to predict the correct tone level of this kind of words (Thangthai et al., 2007; Aroonmanakun, 2004).

5. APPLICATIONS

During gradually enlargement of the LEXiTRON-Pro dictionary, it has been used in our research in speech and language processing. The first application is to use the dictionary as training and evaluation sets of the LTS module itself. The algorithm of our recent LTS tool (Thangthai et al., 2006) is based on automatic syllabification using a set of handcrafted context-free grammars (CFG). The CFG produces a number of hypotheses in the form of syllable sequences. Applying syllable n-gram results in the best syllable sequence having the highest n-gram probability. Each syllable form in the best sequence is then transcribed into corresponding phone sequences using a letter-to-phone table. This algorithm requires two knowledge sources from the LEXiTRON-Pro dictionary. First, by analysis, we can enhance the CFG rule set to cover almost all possible syllables in Thai. Second, the syllable segmented text and pronunciations are used to train the syllable n-gram model. The larger the training set is, the better the model accuracy achieves. Our preliminary test showed that the current LEXiTRON-Pro dictionary could improve the LTS accuracy by 2% by only adapting the n-gram model. With a huge computation power and storage today, the LEXiTRON-Pro dictionary itself could be installed in the LTS tool to alleviate the problem of letter-to-sound exception cases in which only the dictionary can resolve.

Another application of the LEXiTRON-Pro dictionary is its original motivation on the construction of specific ASR tasks. We have developed one ASR engine for an automatic call center in the traffic information domain. The task is to provide a voice-enabled interface for users to access to real-time traffic information inbound/outbound to/from Bangkok and real-time traffic information within Bangkok. Named entities including road names and point of interests are included in the ASR engine. Another ASR engine aimed for automatic transcription of broadcast news reports. In the current stage, a number of person names existing in the LEXiTRON-Pro dictionary are included in the ASR engine.
6. CONCLUSION AND FUTURE WORK

This paper presented a large-scale Thai pronunciation dictionary which can serve as an infrastructural resource for Thai speech and language processing. The LEXiTRON-Pro dictionary will be freely open-source for research and development communities. Although the word definition applied in this dictionary is quite weak as a text chunk is considered a word unit if it appears often in real text, the very large amount of word entries, over 130,000 currently, may include almost every Thai word except ones newly constructed such as person names.

In the future, besides enlarging the dictionary, we are expecting to improve the usability of the dictionary by categorizing words into several semantic groups such as place, person or organization names, titles, abbreviations, etc. Clustering words into these groups will facilitate users who need particular lists of Thai words in their applications. Moreover, we realize the problems of letter-to-sound conversion expressed in the Section 4. We have then a plan to analyze in depth the phonology problems and apply the analysis results to enhance the accuracy of out automatic LTS tool.

7. ACKNOWLEDGEMENT
Thank to Mr. Niran Angkawattanawit for calculating dictionary statistics.

8. NOTE
(1) SAMPA, http://www.phon.ucl.ac.uk/home/sampa/
(2) CMUDICT, http://www.speech.cs.cmu.edu/cgi-bin/cmudict
(3) LEXiTRON dictionary, http://lexitron.nectec.or.th/
(4) BEST project, http://www.hlt.nectec.or.th/best

9. REFERENCE
SOUTH AFRICA’S NEW AFRICAN LANGUAGE DICTIONARIES - USEFUL EDUCATIONAL TOOLS?

Juliane Klein
Universität Leipzig, GERMANY
klein.juliane@freenet.de

Abstract: Dictionaries are valuable means to empower people, as they are useful tools to enhance literacy. In addition, they provide people with the necessary words to talk about a broad range of topics, this is especially important in a multilingual setting as South Africa, where people need to be multilingual, if they want to be able to participate in all domains of life. As the majority of South Africans still live in a pre-dictionary-culture environment, the new African language dictionaries could be introduced as useful educational tools to improve literacy, not only in English but also in the mother tongue. The National Curriculum of South Africa contains six learning outcomes for the language sector, which all can be reached with the help of dictionaries. Dictionaries must be introduced as useful, modern tools in the language classroom that help students with decoding and encoding all kinds of texts. As some of the newer dictionaries have language activities included in their back matter or come with additional workbooks, teachers can either use those exercises or develop similar ones according to those examples. In addition to the language classroom, dictionaries could play an important role in other subjects as well, especially if the medium of instruction is not the mother tongue of the students. In such situations, dictionaries can help students to understand what the teacher is saying or what the question on the exam paper means.

Key Words: bilingual dictionaries; general dictionaries; learning outcomes; language structure and use; listening; Outcomes-based education; reading and viewing; South Africa; speaking; terminology lists; thinking and reasoning; writing

1. SOUTH AFRICAN BACKGROUND

South Africa lies at the southern tip of Africa and has a population of approximately 48 million, of which 79% are classified as Black African, 8.9 % are classified as Coloureds, 2.5% as Indian/Asian and 9.6% as White.
It is a multilingual country that has eleven official languages, i.e. two Germanic languages (Afrikaans and English) and nine Bantu languages. The Bantu languages can be split into three groups, i.e. the Nguni languages (Ndebele, Zulu, Xhosa), the Sotho languages (Sesotho, Sesotho sa Leboa, Setswana, Siswati) and Venda and Xitsonga (which are not related but are often grouped together).

Table 1: Home languages according to race

<table>
<thead>
<tr>
<th>Home language</th>
<th>Black African</th>
<th>Coloured</th>
<th>Indian or Asian</th>
<th>White</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afrikaans</td>
<td>0.7</td>
<td>79.5</td>
<td>1.7</td>
<td>59.1</td>
<td>13.3</td>
</tr>
<tr>
<td>English</td>
<td>0.5</td>
<td>18.9</td>
<td>93.8</td>
<td>39.3</td>
<td>8.2</td>
</tr>
<tr>
<td>IsiNdebele</td>
<td>2.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.1</td>
<td>1.6</td>
</tr>
<tr>
<td>IsiXhosa</td>
<td>22.3</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>17.6</td>
</tr>
<tr>
<td>IsiZulu</td>
<td>30.1</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
<td>23.8</td>
</tr>
<tr>
<td>Sepedi</td>
<td>11.9</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>9.4</td>
</tr>
<tr>
<td>Sesotho</td>
<td>10.0</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>7.9</td>
</tr>
<tr>
<td>Setswana</td>
<td>10.3</td>
<td>0.4</td>
<td>0.0</td>
<td>0.1</td>
<td>8.2</td>
</tr>
<tr>
<td>SiSwati</td>
<td>3.4</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>2.7</td>
</tr>
<tr>
<td>Tshivenda</td>
<td>2.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.3</td>
</tr>
<tr>
<td>Xitsonga</td>
<td>5.6</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>4.4</td>
</tr>
<tr>
<td>Other</td>
<td>0.3</td>
<td>0.2</td>
<td>3.8</td>
<td>1.1</td>
<td>0.5</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Figure 2: Home languages of all South Africans

The above graph shows that most South Africans speak an African language as their home language, only 21% speak Afrikaans or English as their home language.

1.1. Status of the official languages

According to the constitution (ZA government 1996), all eleven languages have the same status and the government must not only promote these languages but must also create conditions for their use and development. Nevertheless, the African languages still lack the status Afrikaans and especially English have. English is still seen as the language of success, whereas the African languages are often considered to be less worth than English or Afrikaans. This is because of the South African history, in which language always has been used to exclude the majority of people from education and from a “better” life. African languages are still considered to be underdeveloped and not “good enough” for formal domains of life. Linguistically, this is not true, as all languages have at least a standardised orthography and dictionaries. The terminology that is needed to use those languages in domains such as medicine or law exists but is not yet documented and standardised, which makes the languages look underdeveloped. The Bill of Rights (Section 29.2), which is part of the constitution, states that: “Everyone has the right to receive education in the official language or languages of their choice in public educational institutions where that education is reasonably practicable”. Nevertheless, most of the education system still uses only Afrikaans and/or English as medium of instruction. The African languages are only used in primary education as medium of instruction up to Grade four (Kamwangamalu 2000) and are offered as subjects up to matric level and as university subjects. According to the current National Curriculum, pupils must study their home language and one additional language, which is in most cases English. Western Cape planned to introduce all three official languages of the province (Afrikaans, English, Xhosa) as compulsory subjects for all pupils but still struggles with the implementation.

1.2. Education system

South Africa’s education system is divided into three bands: 1. general education (grades R to 9); 2. further education (grades 10-12); and 3. higher education. The current education system is outcomes-based, i.e. the “outcomes describe the knowledge,
and values learners should acquire and demonstrate during the learning experience” (The revised National curriculum statement 2005).
Outcome based education has identified the following major learning outcomes for the language modules:
1. Listening; 2. Speaking; 3. Reading and Viewing; 4. Writing; 5. Thinking and Reasoning; and 6. Language Structure and Use. The content of those learning outcomes varies, depending on the Grade and learning area, i.e. home language, first or second additional language. Van der Horst and McDonald (2003) summarize the core learning outcomes as follows: “1. Listening: The learner is able to listen for information and enjoyment, and respond appropriately and critically in a wide range of situations. 2. Speaking: The learner is able to communicate confidentially and effectively in a spoken language in a wide range of situations. 3. Reading and Viewing: The learner is able to read and view for information and enjoyment, and responds critically to the aesthetic, cultural and emotional values in texts. 4. Writing: The learner is able to write different kinds of actual and imaginative texts for a wide range of purposes. 5. Thinking and Reasoning: The learner is able to use language to think and reason, and access, process and use information for learning. 6. Language Structure and Use: The learner knows and is able to use the sounds, words and the grammar of a language to create and interpret texts”.

In May 2009, Jacob Zuma, the new South African President, drastically changed the cabinet and split the Department of Education into two new departments, one for basic (primary and secondary) education and one for tertiary education. What happens to the current language in education policy and outcomes based education teaching methods is not clear yet, although there might be changes and the outcomes based education teaching model might be abandoned. It is also not clear, if the teaching of the African languages will continue as it is, will increase or will be reduced. A change here would probably have a change of prescribed textbooks and dictionaries as a result.

1.3. Lexicographic scene in South Africa

The lexicographic scene in South Africa is quite diverse and can roughly be separated into four groups: 1. commercial publishers; 2. National lexicographic Units; 3. government related organisations; 4. individuals that compile terminologies or small technical dictionaries. The most successful of those groups are the commercial publishers as their main aim in publishing dictionaries is earning money. This means that they are interested in compiling dictionaries that have a certain quality and are interesting in making the product available. The second group, the National lexicographic units (NLUS), consist of a group of eleven units, one for each official language. Although they are funded and monitored by PanSALB (Pan South African Language Board), their registration as independent non-profit Section 21 companies allows them to raise additional funds. PanSALB is responsible for language matters in South Africa and is a government related organisation. Their main tasks are the promotion of multilingualism and “[c]reating the conditions for the development of and the equal use of all official languages” (PanSALB online 2009). The NLUS are based at tertiary education institutions throughout South Africa and their main task is to compile explanatory monolingual dictionaries for their respective language, as well as other dictionaries. In 2009, bilingual dictionaries of different sizes and qualities exist for all official languages. Several multilingual general and technical dictionaries were compiled by OUP and Macmillan Publishers. A Math and Science terminology list for all official languages, compiled by the National language service of the Department of Arts and Culture, as well as some other terminologies exist. In addition, some bilingual online and cell phone dictionaries are available.
2. THEORETICAL BACKGROUND: DICTIONARIES IN OUTCOMES-BASED EDUCATION

In this education system, language officially plays an important role, as it “is used for personal, communication, educational, aesthetic, creative, cultural, political and critical purposes, plus the rest of the curriculum is learnt through language” (National Revised Curriculum Statement 2005). Dictionaries are valuable educational tools in language classrooms. They can help pupils in the reception and production of spoken and written texts.

2.1. Dictionaries in language classrooms

Although the use of dictionaries is not specified in all six learning outcomes, all can benefit from dictionary use. The National Curriculum specifically demands that dictionaries are to be used in some of the learning outcomes. Examples are: Grade Two home language, demands that dictionaries should be used in writing to check the spelling and meaning of words and in language structure and use to confirm the spelling of words. Grade Three, first additional language, thinking and reasoning outcome, demands that students should use a children’s dictionary and keep a personal dictionary. The writing outcome expects students to be able to enter words in a personal dictionary. Grade Five second additional language, demands that students use a bilingual/monolingual dictionary for language learning to achieve one of the reading and viewing outcomes. Grade Seven, additional language, demands that students use and understand a dictionary entry as one of the reading and viewing requirements.

2.2. Dictionaries in a multilingual setting where the mother tongue is not the Medium of instruction

Dictionaries can also play a vital role when used in an environment where the medium of instruction is not the mother tongue of the pupils (and the teachers). In such cases, bilingual dictionaries and especially bilingual technical dictionaries are important tools that enable and improve the communication between pupils and teacher. Technical dictionaries will provide students with the necessary terms to understand the textbooks and the teacher. Even if the teacher has the same mother tongue as the pupils, and can use some code-switching in order to explain the concepts and tasks, the exams usually have to be written in the official medium of instruction. This implies that a pupil who is not functionally proficient in the medium of instruction as well as in his mother tongue, will get lower marks just because of the language barrier. Using a dictionary in a non-mother tongue environment can help students to understand the questions and answer them correctly because the dictionary provides them with the necessary words.

3. DICTIONARIES IN SOUTH AFRICAN CLASSROOMS

African language dictionaries can now be found on textbook lists supplied by the Gauteng Provincial Department of Education (online resource). They include dictionaries compiled by the National Lexicographic Units as well as by commercial publishers. In addition, the Department of Education’s language policy specifically encourages bilingualism - dictionaries are considered to be valuable tools in encouraging bilingualism. As most South Africans still live in a pre-dictionary culture, dictionaries are still not taken for granted in most South African language classrooms. The main reason
for this is that dictionaries for the African languages have been neglected during the last
decades and therefore, only a few old, euro-centric dictionaries written by missionaries
existed before the 1996 constitution paved the way for new African language dictionaries.

Three different types of dictionaries and their use in outcomes-based education language
classrooms will be discussed in this section. The first type consists of beginners
dictionaries which are compiled for very young children who have just started to learn
how to read and write. The second type is a bilingual school and learner’s dictionary for
older pupils and the last one are technical dictionaries and bilingual terminology lists that
can be used for all subjects if the medium of instruction is not the mother tongue of the
learner.

3.1. First Bilingual Dictionaries in language teaching

As the National Curriculum explicitly demands the use of dictionaries for home
languages from Grade Two onwards, the First Bilingual Dictionary series compiled by
Oxford University Press South Africa are good examples for dictionaries for young
children. This series consists of bilingual dictionaries that contain English and one of the
official African languages. They are compiled for the use from Grade Two onwards. The
dictionaries are ordered according to semantic fields, for example *inside our house* or *a
sports day*. Each of the different semantic fields has a double page showing a picture that
includes all the words which are translated on these pages. The English words are
displayed in alphabetic order together with a translation in the target language and a little
picture showing the word. In addition, a few example sentences in English as well as in
the target language are provided. Little tasks or exercises for students can be found on
these pages, too. As those dictionaries also include some language activities at the back,
which meet the curriculum requirements, these dictionaries can be considered useful
educational tools, if the teachers use the dictionaries and the exercises. According to the
curriculum, a grade three student (first additional language) should be able to talk about a
picture as one of the speaking outcomes. This can easily be practised with a dictionary
that contains pictures. In the writing area, students are expected to enter words in a
personal dictionary – an activity that is included in the back matter of those dictionaries.
Learning outcome five (thinking and reasoning), explicitly demands that learners should
keep a personal dictionary and use a children’s dictionary. Therefore, those dictionaries
do not only fulfil the curriculum requirements, they also help students to grasp important
lexicographic concepts, as for example the alphabetical order of a dictionary. In addition,
they also help to familiarize students with dictionaries, a first step towards an active
dictionary culture. The latter is very important in South Africa, as most students do not
know what a dictionary is or how to use it. Having a user-friendly dictionary that is
appropriate for the children’s age helps them to build a positive relationship with
dictionaries and takes away the fear of the sometimes intimidating, larger dictionaries that
are used later on in education.

3.2. Multilingual Primary Dictionaries in language teaching

Examples of this dictionary type include the Multilingual Primary Dictionary series
published by Oxford University Press South Africa. They are multilingual dictionaries
with English as the source language and either Afrikaans, IsiXhosa, IsiZulu, SiSwati or
Afrikaans, Sepedi, Sesotho, Setswana as target languages. According to the back matter
of the dictionary, it is designed for Grades four to seven and contains approximately 1500
English words. It is compiled for learning English but it is not specified if it is intended for English as a first additional language or English as a second additional language. The dictionary has a short user guide that explains how the dictionary is to be used. Dictionary entries are kept simple, i.e. they consist of the English lemma, its part of speech, an English example sentence and the translations of the lemma into the target languages and in some cases pictures are added. Those dictionaries and their workbooks are recommended on the Gauteng Department of Education textbook catalogue for usage in Grades four to seven. The exercises in the workbook cover the most important concepts, which are needed to use dictionaries successfully. Students can practise the alphabet or learn to distinguish between consonants and vowels. Such tasks cover some of the language structure and use learning outcomes. Another activity for students, which is specifically mentioned in the thinking and reasoning outcome, is compiling their own dictionary. Although this dictionary is compiled for Grades Four to Seven, it might be more useful for younger learners in Grades Four to Five than for older ones.

3.3. Bilingual dictionaries for grades 8-12 in language classrooms

This group consists of a variety of bilingual dictionaries that could be used in language classrooms. I have chosen two different sub-types, one published by a National Lexicographic Unit and one published by OUP South Africa.

3.3.1. Xitsonga-English Dictionary (Xitsonga NLU)

This bilingual Xitsonga – English dictionary was compiled by the Xitsonga National Lexicographic Unit and is on the textbook catalogue of the Gauteng Department of Education for Grade Eight onwards. According to the cover, “[t]his Xitsonga – English bilingual and explanatory dictionary has been compiled to meet the needs of Xitsonga Home Language learners, First Additional learners, Xitsonga students as well as speakers of other languages who need a working knowledge and understanding of Xitsonga words” (Xitsonga NLU 2005). It consists of 7790 entries, i.e. approximately 3900 entries per language. The entries consist of the lemma, Part of Speech information, in some case a very short explanation of the lemma or some synonyms and the translation of the lemma. At the beginning, it contains a short user guide in Xitsonga and English and some information on the speech sounds and pronunciations of Xitsonga but no pronunciation information for each lemma. This dictionary can be used to reach the learning outcome of reading and viewing, as students can use it to find translations for unknown words. Using it successfully for writing activities depends on the language competence of the user, as this dictionary does not provide usage information and does not provide example sentences and does not provide enough grammatical information. In other words, this dictionary does not provide any information on irregular verb forms, the verb to be is only listed as auxiliary but without providing the different verb forms, i.e. a user who does not know that the verb form is belongs to the entry be has no chance to find the correct translation as is is neither entered under to be nor has it its own entry. In contrast to the other dictionaries discussed in sections 3.1- 3.3., this dictionary does not come with an extra workbook and does not include language activities. This implies that teachers need a lot initiative on their own, to transform it into a useful, easy to use educational tool.

3.3.2. Sepedi-English dictionary

This is a corpus-based bilingual Northern Sotho (Sepedi) – English dictionary that includes 5000 headwords for each language. The target user groups are students learning
Northern Sotho or English. It has a detailed but easy to understand user-guide in Sepedi and English, which explains all parts of the dictionary entries and explains how the headwords were selected and what is included in the dictionary. In the middle of the dictionary the user finds some dictionary activities, a mini Northern Sotho grammar, a list of English irregular verbs, as well as information on pronunciation and spelling. This dictionary focuses on the needs of learners at school, as it includes not only 300 key words from the curriculum but also example sentences in Northern Sotho and English for each entry. In addition, various teachers gave their feedback on early draft entries of the dictionary which helped not only to improve the dictionary but also helped to make it suitable for use in (language) classrooms. It is accompanied by a dictionary workbook that provides curriculum-based language activities which can be solved with the help of the dictionary. All six learning outcomes can be reached with the help of such a dictionary. It helps students in listening as it covers a large variety of vocabulary and has pronunciation and spelling guide included. The speaking outcome is supported through pronunciation information of the headwords, a piece of information that is missing in many of the African language dictionaries. It helps students when reading, as it provides clear and frequent translations. In addition, it full fills the assessment standards for understanding how reference books work, i.e. understanding the different parts of a dictionary entry. As pronunciation is explicitly mentioned as part of understanding a dictionary entry in the National curriculum, this dictionary is the only one that reaches this requirement completely. The writing outcome is supported through activities in the work book and through examples of letters, e-mails etc in the middle of the dictionary. Using language and literacy through the curriculum, one of the assessment standards of the thinking and reasoning outcome, is supported, as key words from the curriculum are included in the dictionary. The last learning outcome, language structure and use, is supported through language activities in the work book that include using the correct tense etc.

**3.4. (Technical) dictionaries in non-mother tongue classrooms in South Africa**

Mahlalela and Heugh (2002) argue that there is still a lack of available teaching materials for the African languages because of the low status of the African languages and the resistance to the mother-tongue policy since the 1970s. Although bilingual dictionaries for all official languages are now available, there is still an alarming lack of technical dictionaries for the African languages. Nevertheless, the Gauteng Department of Education lists a pocket bilingual maths and a pocket bilingual science in their text book catalogue. In addition, the National language service has compiled a multilingual mathematics dictionary for grade R to Six and a Multilingual Science and Technology Dictionary for Grade Four to Six. The Mathematics dictionary is a unidirectional word list that contains approximately 1000 terms. The source language is English and the target languages are the other ten official languages. The Natural Science and Technology dictionary is similar to the mathematics dictionary, i.e. the source language is English and the other languages are the target languages. Both dictionaries are not dictionaries in the traditional sense, they are word lists. Nevertheless, they can be taken as a starting point to compile proper technical dictionaries that could be valuable educational tools in a multilingual environment. As technical dictionaries and even glossaries help students to better understand the subject and what the teacher is saying, they can be considered useful educational tools as long as the entries are correctly translated. If the medium of instruction is not the mother tongue of the learner, being able to understand the teacher, textbooks and exam papers is a prerequisite to successfully pass the education system. Having bilingual general and technical dictionaries and knowing how to use them helps students to succeed in education.
4. CONCLUSION

In a multilingual society where most students are not taught in their mother tongue, being functionally literate and fluent in the language that is used as medium of instruction for most part of the education is extremely important. Dictionaries can help students to become functionally literate not only in their home language but also in their additional language(s), as they are useful tools to improve their language skills. Although the curriculum does not mention dictionary use in all six learning outcomes of the curriculum, dictionaries are useful in all six learning outcomes. The listing of some of the dictionaries in textbook catalogues, as for example from the Gauteng Department of Education, gives them a legal background to be used as educational tools but does not automatically make them useful tools. In order to be useful, they must be user-friendly, correct and designed for the specific user groups or as Gouws and Prinsloo 2005 argue, “[g]ood dictionaries are products that can be used as linguistic instruments by their respective target user groups”. Most of the new African languages are useful educational tools but they must be used in order to help students to improve their language skills.

5. REFERENCES

5.1. Literature

(7) Map of South Africa. www4.southafrica.net (accessed 15.05.2009)

5.2. Dictionaries

SEMANTIC IDENTIFICATION OF PHRASE VARIANTS IN THE CASE OF ‘AND YET’ AND ‘BUT YET’ BASED ON A PHRASEOLOGICAL APPROACH

Ai Inoue
National Defense Academy, Yokosuka, JAPAN
aiinoue@nda.ac.jp

Abstract: The set phrases ‘and yet’ and ‘but yet’, together with just ‘yet’ are usually treated in dictionaries as simple variants with no significant semantic differences. This study suggests that this is actually not the case. Dictionaries tend to simply list phrase variants without carefully distinguishing the semantic and functional differences between them. As an alternative and worthwhile approach is to study these variants by using data obtained from the corpora. I will use examples from the British National Corpus (BNC) and WordBanksOnline for the analysing the abovementioned phrase variants. A simple comparison of the numbers of occurrences of each variant indicates a significant contrast between them: ‘and yet’ is used more frequently than ‘but yet’ (the latter occurs only 44 times in BNC and 51 times in WordBanksOnline). Viewed in terms of their syntactic features displayed in various contexts, ‘and yet’ has two functions: first, it works as an emphatic form of the conjunction ‘and’ by means of which the preceding statement and the following statement are copulatively connected. Secondly, ‘and yet’ also behaves in a manner similar to ‘yet’, when used as an adversary conjunction. ‘But yet’ has only one function and behaves like ‘but’ and ‘yet’ used alone. In other words, ‘but yet’ strongly emphasizes the difference between the previous statement and the following statement, which is evidenced by the fact that it always follows emphatic words like still; for example: I’ve been to so many job interviews and application sent off application forms but yet I’m still unemployed (WordBanksOnline).

Key Words: phraseology, semantic identification, emphatic conjunction, adversary conjunction

1. INTRODUCTION

This paper is an empirical research study that clarifies the functions, differences, and relationship of the expressions ‘and yet’ and ‘but yet’ observed in present-day English, based on a phraseological approach. In addition, it hypothesizes about the underlying principles that form these two phraseological units\(^{(1)}\).

Let us examine the following examples (underlined by the author).
In contrast, the restaurant manager is far less easy to define. I would argue that they are as charismatic as chefs and yet their skills are less tangible and harder to gauge. [BNC]

I am thunderstruck by this verse on a number of counts: first, that I can read it and understand it at all; second, by the sheer lyric of it. The rhyme of ‘know’ and ‘grow’ is so obvious and yet so surprising. [WB]

The pores in the skin are a classic example: they ‘cannot become perceptible to us by themselves, but yet their presence in the skin can be deduced from sweat’. [BNC]

… True, I know the grammar and the words, but yet I know not how to speak them. [Sinclair 2008:XV]

It is evident from (1) and (2) that it is very difficult to describe how ‘and yet’ and ‘but yet’ behave differently in their respective contexts. When we look them up in the latest dictionaries published in Europe, US, and Japan (CALD³, LAAD², LDCE⁵, LONGMAN, OALD⁷, G¹, WISDOM², Youth), although these contain many phraseological units which are referred to as ‘spoken idioms’ or ‘spoken phrases’, we cannot obtain a satisfactory explanation regarding their meanings and functions. Moreover, some of the dictionaries do not even define or describe ‘and yet’ and ‘but yet’ at all.

Similarly, grammar books such as Quirk et al. (1985) just cite the example of ‘and yet’ as indicated by the following example:

She tried hard and (yet) she failed. (Quirk et al. 1985: 931)

Quirk et al. (1985:931) provide a brief explanation with regard to (3): ‘The second clause is felt to be surprising in view of the first, so that the first clause as a CONCESSIONAL force’. No full explanation regarding the use of ‘and yet’ and ‘but yet’ has been given in any grammar book.

However, examples of ‘and yet’ and ‘but yet’ in the BNC and WB are too many to enumerate. The following table shows their frequency.
Table 1: The number of occurrence of ‘and yet’ and ‘but yet’ in BNC and WB

<table>
<thead>
<tr>
<th></th>
<th>Number of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>and yet</td>
<td>5027</td>
</tr>
<tr>
<td>but yet</td>
<td>96</td>
</tr>
</tbody>
</table>

As remarked above, the functions of ‘and yet’ and ‘but yet’ have not been fully discussed thus far; moreover, the differences between them have been discussed to a much lesser extent. Consequently, the present research aims to explore the function, difference, and relationship of ‘and yet’ and ‘but yet’ including ‘yet’ by responding to the following research questions in the course of the study.

(4) a. What are the functions of ‘and yet’, ‘but yet’, and ‘yet’?
   b. What is the relationship between ‘and yet’, ‘but yet’, and ‘yet’?
   c. What grammatical structures include the use of ‘and yet’, ‘but yet’?
   d. What is the underlying principle of the phraseological units ‘and yet’ and ‘but yet’?

2. (4a) WHAT ARE THE FUNCTIONS OF ‘AND YET’, ‘BUT YET’, AND ‘YET’?
2.1 The Functions Of ‘And Yet’
Examples retrieved from BNC, WB, and other materials. This research used the double underlining to clarify the proposition and the dotted lines to make an adversary part clear in each context.

(5) In contrast, the restaurant manager is far less easy to define. I would argue that they are as charismatic as chefs and yet their skills are less tangible and harder to gauge.
[BNC]

(6) This person claims to be a great detective and yet he has detected nothing, nothing at all. He has asked questions all over the village, but none of the answers have solved this dreadful crime.
[BNC]

(7) E. Neville Isdell: … Um, and we’ve come through a number of very interesting and largely successful phases to be at a time of tremendous challenge and yet at a time when the global economy has actually never been healthier.
[Tsuruta ・ Shibata 2007: 82ff.]

(8) And they also claim, for example in Britain and Denmark, to be the happiest countries in the world, with very, very high family satisfaction - much higher than, for example, countries like Italy or Germany. And yet, when we look at the actual divorce rate - in Britain it’s 37 per cent and in Denmark it’s fairly close – we’ve got the highest divorce rates in Europe.
[WB]
If you are saying that something is really grim, and yet you’re making jokes about it, and you’re having people wink at the camera, does that sort of imply that everything’s OK? [WB]

The phrase ‘and yet’ in the above examples has two functions: it behaves as an adversary conjunction in examples (5), (6), (7), (8), and (9). It also works as an emphatic form of the conjunction ‘and’ when it connects the preceding statement and the following statement copulatively: this is exemplified in (10) and (11).

When ‘and yet’ works as an adversary conjunction we can see two syntactic patterns: in one, a negative such as no, not, less (as described by Quirk et al. (1985)) appears in new information which follows ‘and yet’: this is seen in (5), (6), and (7). In a different pattern, there is no single negative in the new information. Observe examples (8) and (9) again. The new information following ‘and yet’ as seen in both examples consists of adversative statements which are not predictable from the proposition but yet are copulatively connected to the proposition, without use of a negative. ‘And yet’ as an adversary conjunction may be used at the beginning or the middle of the sentence. The dotted line regions in the above examples are difficult to predict from either the double-lined parts or the proposition.

Now, let us look at ‘and yet’ when it functions as an emphatic form of the conjunction ‘and’.

(10) I am thunderstruck by this verse on a number of counts: first, that I can read it and understand it at all; second, by the sheer lyric of it. The rhyme of ‘know’ and ‘grow’ is so obvious and yet so surprising. [WB]

(11) The novelist was very distressed. And yet he took no steps to reintroduce it in later editions. I believe he came to see it would not do. [BNC]

We can see from (10) and (11) that similar details to the proposition, where a dotted line is drawn, are mentioned in the new information following ‘and yet’. In the case of (10), the dotted line part (surprising) is not the opposite content to the double line part (obvious) but it adds further information to the proposition the rhyme of ‘know’ and ‘grow’. Similarly, the dotted line section he took no steps is deduced from the proposition the novelist was very distressed in the case of (11). Consequently ‘and yet’ functions to connect the preceding statement and the following one. In addition, some emphatic words like so are part of the function. Hence, it is clear by now that ‘and yet’ is functioning as an emphatic form of the conjunction ‘and’.
2.2 The Function Of ‘But Yet’

‘But yet’ has only one function: it behaves like ‘but’ and ‘yet’. Here are some examples.

(12) The pores in the skin are a classic example: they ‘cannot’ become perceptible to us by themselves, but yet their presence in the skin can be deduced from sweat.

[BNC]

(13) There had been loss of shipping and sales of overseas assets, but there was no great disruption and dislocation. There was some social and political unrest, it is true, but nothing on the scale seen in Poland, Austria, Germany, Belgium and France. But yet Britain had a poor record of economic achievement in the 1920s, before the world depression marked the years 1929–32.

[BNC]

(14) .... True, I know the grammar and the words, but yet I know not how to speak them.

[Sinclair 2008:XV]

(15) The initial recital was given by Mr. D. I. Steele, who later joined the Music staff of the School. An instrument of smaller but yet great importance was the ‘talkie cinema’ acquired by the School at about the same time.

[BNC]

(16) Supposing then that there was water with bloodstain in it, when somebody went to the toilet, urine, pass their urine, where, where would the damage be then do you think? Bladder Bladder or the Kidneys kidneys, so it’d come out in there, okay, so those are the areas where damage int internal organs, but yet you still have some sign eventually of blood loss, but there are still some organs in the body where there is no sn nowhere at all an outlet….

[BNC]

(17) I mean I’ve been unemployed a y+ almost a year over a year now and I’ve been to so many job interviews and application sent off application forms but yet I’m still unemployed.

[WB]

Some examples above, like (12) and (13), are accompanied by a negative (the double underlining parts). All examples illustrate the fact that it is very difficult to conjecture the new information (the dotted line parts), which shows unexpected statements far separated from the proposition (the double-lined parts).

Considering the above, in terms of function ‘but yet’ strongly emphasizes the difference between the previous statement and the following statement, which is evidenced by the fact that it always follows emphatic words like still in (17).

2.3 The Functions Of “Yet”

‘Yet’ functions as an adversary conjunction and as a copulative conjunction similar to ‘and yet’. Let us look at the following examples, where ‘yet’ works as an adversary
Apart from three months as Nigeria’s Foreign Minister, in 1983, Chief Anyaoku has worked at the Commonwealth Secretariat since it was set up in 1966. Yet he wasn’t everyone’s first choice. There was an unprecedented contest for the new Secretary General at the Kuala Lumpur Summit last year, in which Chief Anyaoku beat the ebullient former Australian Prime Minister Malcolm Fraser. [WB]

Dostoevsky no-home takes with him, likewise the transpersonal motif first voiced by Marmeladov in this novel, that a man must have somewhere to go. He is absent yet meticulous. Paying for a missing drink-shop teaspoon which has nothing to do with him, and spending a long time in the “interesting occupation” of trying to catch a fly. [BNC]

In particular the essay on Beerbohm is startlingly good: temperate, appreciative, sympathetic, yet in the end unsparing. [BNC]

The new statements following ‘yet’ (the double-lined parts) are difficult to be predicted from the previous sentence, but they are connected with the previous statements. ‘Yet’ as an adversary conjunction is mainly positioned at the beginning or the middle of a sentence.

For instance, let us take a look at example (10). Chief Anyaoku had been working hard at the Commonwealth Secretariat for a long time, (indicated by bold type): as an inevitable consequence, he would be the first candidate for new Secretary General. However, he was not chosen as the first candidate. This fact, indicated by double underlining, cannot be guessed from the part indicated by bold type. Similar to (18), the double-underlined sections in (19) and (20) cannot be predicted from the parts indicated by bold type.

Next, here are some instances of ‘yet’ as a copulative conjunction.

The article was by Percy Hoskins, the greatest Commissioners of Police and thieves and villains, who drank champagne at the Caprice, had an apartment in park lane, was an intimate of Lord Beaverbrook, yet kept the common touch. [BNC]

Surprise birthday parties come closest, but rarely do you have to stand in front of the party and smile gallantly while each guest tells a warm yet funny anecdote of which you are the subject. At least at a surprise party you can tell an anecdote that
'Yet' working as a copulative conjunction is used here in the middle of a sentence. In contrast to ‘yet’ as an adversary conjunction, the new statements following ‘yet’ have a similar content to the previous ones, although they cannot be predicted from the previous ones: this means that they preserve the continuity of the proposition. In addition, no emphatic words like so are not used in the case of ‘yet’ working as an adversary conjunction (unlike ‘and yet’). In the case of (21), ‘yet’ is used to add that Percy Hoskins commonly kept in touch with Lord Beaverbrook as well as privately. In a similar way ‘yet’ in (21) adds extra information that guests participating in the party tell a warm, as well as funny, story.

3. (4b) WHAT IS THE RELATIONSHIP BETWEEN ‘AND YET’, ‘BUT YET’, AND ‘YET’?

The relationship between ‘and yet’, ‘but yet’, and ‘yet’ may be expressed schematically thus:

![Diagram](image)

**Figure 1:** The correlation between ‘and yet’, ‘but yet’, and ‘yet’

Figure 1 illustrates two points: first, ‘yet’ is a phraseological unit operating somewhere between ‘and yet’ and ‘but yet’, which means that ‘yet’ is an unmarked phraseological units between them. In other words, ‘and yet’ and ‘but yet’ acquire their own functions centered on ‘yet’. Second, the core function of ‘and yet’, ‘but yet’, and ‘yet’ is that of adversary. This leads to the conclusion that the root function of ‘yet’ as an adversary still exists where it is functioning as a conjunctive adverb according to *OED*².
4. (4c) WHAT GRAMMATICAL STRUCTURES INCLUDE THE USE OF ‘AND YET’, ‘BUT YET’?

In the section above, it has been observed that the functions of ‘and yet’ and ‘but yet’ partially overlap with those of ‘yet’. I will go on to examine how ‘and yet’ and ‘but yet’ each come to be composed from ‘yet’.

4.1 The Structure Of ‘And Yet’

First, let us consider how the function of ‘and yet’ working as an emphatic copulative developed. As discussed above, ‘and yet’ retains two roles: an adversary and an emphatic copulative similar to ‘yet’. When ‘yet’ is used in a sentence, it is difficult to distinguish which of the two roles is operating. In order to avoid misunderstanding the two functions of ‘yet’ and to explain its copulative role, it can be considered that ‘and’, which can clearly help show a copulative, occurs with ‘yet’. Accordingly, ‘and yet’ comes to have the functions of both ‘and’ and ‘yet’, i.e. an adversary; then ‘and yet’ works as an emphatic copulative and new statements following ‘and yet’ have content that cannot be predicted from the previous sentence.

In the case of ‘and yet’ functioning as an adversary, the core function of phraseological units including ‘yet’ or adversary mainly works although the function of ‘and’ remains.

This phenomenon can be accounted for by ‘concept categorization’(2). ‘And’ and ‘yet’ originally had their own functions, but the unit ‘and yet’ gradually came into usage forming a phraseological unit ‘and yet’ in an example of “concept categorization”.

4.2 The Structure Of “But Yet”

The same sorts of explanations can be made about ‘but yet’. As mentioned above, it is problematic to judge which function of ‘yet’ is operating when it is used by itself. It can be assumed that ‘but’ implying an adversary role occurs with ‘yet’ to make the adversary function of ‘yet’ more striking. As a result, through ‘concept categorization’, ‘but yet’ comes to be recognized as a phraseological units in a similar way to ‘and yet’.

5. WHAT IS THE UNDERLYING PRINCIPLE OF THE PHRASEOLOGICAL UNIT ‘AND YET’ AND ‘BUT YET’?

This section discusses what kind of principle is working to form the unit ‘and yet’ and ‘but yet’. My opinion is that the basic principle, can be explained in terms of ‘redundancy’, a feature of ‘linguistic economy’(3).
A main principle of effective communication is that we should avoid using words which tend to cause misunderstanding. This principle consists of ‘economy’, where words tend to be excluded if they fall within a scope which might give a false impression, and ‘redundancy’, where similar words are used unnecessarily such as until to next Friday. ‘Redundancy’ is operating in the formation of ‘and yet’ and ‘but yet’. Whenever ‘yet’ is used, it introduces a possible misunderstanding about which of its functions is being used. It is natural to assume that ‘redundancy’ is deliberate in order to avoid miscomprehension. As a result, ‘and’ and ‘but’ are each added to ‘yet’. ‘And yet’ and ‘but yet’ therefore become established as phraseological units because of ‘concept categorization’ and come to have their own function by means of ‘redundancy’.

6. DICTIONARIES DEFINITION OF ‘AND YET’, ‘BUT YET’, AND ‘YET’
The previous sections have identified the functions of and relationship between the three phraseological units using the inductive method. In addition, it has been noted that how ‘and yet’ and ‘but yet’ are formed and what kind of basic principle is used to form them. I will propose my personal definition of ‘and yet’, ‘but yet’, and ‘yet’ based on the above observations.

It is quite difficult to predict new information following the three phraseological units from the proposition, as has been repeatedly pointed out. This is the same as saying that the central role of the three phraseological units is to introduce new information. Each phraseological unit has its own function on the basis of the role of giving new information. The following text introduces my own draft definitions of ‘and yet’, ‘but yet’, and ‘yet’.

(23) PHRASE:

and yet Used as an emphatic way of adding a new similar statement and relating it to the previous statement: She is kind to everyone and yet so congenial to me.

Used as a way of adding a new opposite statement and connecting it to the preceding statement: a simple and yet effective of losing weight

but yet Used as an emphatic of adding a contrasting statement to the previous statement: I thought I could manage to speak French but yet I couldn’t make myself understood at all in France last year.

yet Used as a way of adding a new similar statement and relating it to the previous statement: His story makes me happy yet comfortable.

Used as a way of adding a new opposite statement and connecting it to
the preceding statement: My boss and I have always hit it off well together, yet we haven’t met in our personal lives.

7. CONCLUSION
I should like to draw three conclusions from this analysis. First, ‘and yet’, ‘but yet’, and ‘yet’ have the function of giving new information. Little mention has been made of this function in research and dictionaries so far. Secondly, ‘and yet’, ‘but yet’, and ‘yet’ are closely related in their functions, but ‘yet’ is a phraseological unit intermediate between ‘and yet’ and ‘but yet’. Lastly, redundancy (as a principle of ‘linguistic economy’) is engaged in the formation of ‘and yet’ and ‘but yet’.

8. NOTES
(1) Phraseological units are defined here as units that have various meanings to suit each context irrespective of the number of words. They subsume idioms, collocations, lexical phrases, discourse particles, and the relation between phrase (word) and grammatical category. (Inoue 2007:104)

(2) According to Yagi (1999:105), ‘concept categorization’ is a process in which various concepts expressed in various syntactic units are interpreted as one syntactic unit. (i) is an example of such usage quoted from Yagi (1999:105).

(i) a. John is far from the destination.
   b. John is far from being honest.
   c. John is far from honest.

The far from in (ia) talks about the physical long distance, while the far from in (ib) and (ic) metaphorically refers to the situation where the real fact is that John is almost completely opposite to being honest. The process of deriving the use of far from from the basic use of the far from in its physical sense is as follows: the relevant part of (ia) is to be syntactically analysed as [far [from the destination]], far is an adverb followed by a prepositional phrase from the destination, which in turn may be reanalysed (in Kajita (1977)’s term) as [[far from] the destination], far from, as a lexical unit, holding an idiomatic status expressing the meaning of ‘physically a long way away from’. Once this idiomatic status is given to far from, then the next step to the semantic expansion of far from from its physical sense to its metaphorical sense easily occurs as can be seen in (ib) and (ic).

Yagi (1999) used the term ‘concept categorization’ because he wants to cover idiomatizations and functional shifts as is seen ‘none were as savvy and good-humored
and as *of-the-moment* as the Beatles’, where, *of the moment*, a prepositional phrase usually functioning as an adverbial, is turned into an ad hoc one-word idiom-like predicative adjective. The same thing seems to be happening in ‘You are in a *dog-eat-dog* newspaper business’, where ‘dog-eat-dog’ itself is a sentence-turned noun-phrase, meaning ‘a situation where a lot of competition is going on,’ is functioning as one complex attributive adjective. (Inoue 2007:125) 

(3) The basic role of languages is to deliver ‘effective communication’. Two elements i.e. ‘economy’ and ‘redundancy’ operate in all languages striking the right balance to carry out effective communication.

9. REFERENCES

Papers and Books


Dictionaries


SECOND EDITION OF BUDDHISM DICTIONARY COMPILED BY PRAPROMKUNAPORN CD-ROM VERSION

Prapat Uthayopas, Prayooth Bhandularb
Siam University, THAILAND
prapat1932@hotmail.com, prayooth@yahoo.com

Abstract: Originally, the first edition of Buddhism Dictionary with numerical Dhammas of Theravada Buddhism conducted by Prapromkunaporn* was presented in booklet format. It has been transformed into CD-ROM format afterward by Prapat Uthayopas and Prayooth Bundhularb in order to provide the convenience in searching for interested people. Since Prapromkunaporn has recently updated his dictionary with the definition of more than 1,100 words as well as adding some new Buddhist terms, therefore the CD-ROM format of this dictionary has to be revised to reflect those changes as well. This paper was conducted as a continual research to fulfill Prapromkunaporn’s objective in improving his Buddhism Dictionary.

*The Venerable Bhikkhu P.A. Payutto—the UNESCO’s Prize for Peace Education 1994

1. Introduction
The Buddhism Dictionary, the dictionary for the Buddhism technical terms compiled by Prapromkunaporn (a celebrated Buddhist monk) is a very useful academic reference book for everyone especially teachers, instructors and students from any fields of study. The CD-ROM containing dictionary software and the data developed by the researcher could be used easily and conveniently to search for the meaning of words via computer. This project has been successfully implemented in many educational institutes. After utilizing this Buddhism dictionary for some time, the writer of this dictionary has added and alternated some 1,100 new words into it. The researcher found it is very useful to further this research by adding the new words and their meanings into the dictionary as well as correcting some previous errors in order to make the Buddhism Dictionary as complete as possible.

2. Objectives
To add new words and their meanings into the dictionary and correct previous errors by using computer software.
To write the new system into CD-ROM.
To submit the CD-ROM to Prapromkunaporn in order for him to utilize the dictionary as needed.
To publicize in various places for other people to make use of this dictionary.

3. Goals and Benefits
General people including teachers, instructors and students can use the Buddhism Dictionary to search for the meaning of words conveniently and quickly.
To adopt the university’s policy in terms of providing academic services to the community.
Siam University gains good reputation in using collaborative technology for knowledge creation.

4. **Project Methodology**

The software program was written in Visual Basic and the data gathered from Prapromkunaporn was in rich text format. Therefore, the meaning of word can be searched from the Buddhism Dictionary Book, first edition - Computer version.

Since the outcome of this project is in digital format (CD-ROM), so the cost is considerably low compared to the dictionary in book format. Therefore, it can be produced in a large amount of copies to be able to sufficiently provide to the community.

The steps of development process are as followed:

4.1 Obtain permission from Prapromkunaporn for the modification of the Buddhism Dictionary.

4.2 Request from the writer for the data which are the collection of words and their meanings.

4.3 Design and develop the system.

4.4 Test and modify the program.

4.5 Run the software and trial usage.

4.6 Produce CD-ROMs.

4.7 Submit a number of finished CD-ROMs to Prapromkunaporn.

4.8 Publicize in various proper places.

5. **Instruction for the use of the Buddhism Dictionary.**

When the CD is inserted into the CD driver of a computer, the program will automatically download the data into the memory. An icon named Buddict will appear on the desktop. Double click on this icon to start using the Dictionary.

Double click on the Buddict icon, the first page of the Dictionary will appear as in Fig. 1.

![Fig. 1](image)

Finding the meaning of the words can be done in three ways:

1. Scroll the mouse to find the required word and double click on it directly, the meaning of the word will appear at the adjacent column, as shown in Fig. 2.
2. Select and double click on the alphabet groups at the top of the window.

Group of the selected alphabet will appear at word’s windows. This makes it easy to scroll and select the word within this alphabet’s group as in Fig. 3.

3. Type the required word in the empty space above the word window. The content in the word window will narrow down your selected word. You might get the meaning of the word right away before finish typing as shown in Fig. 4.
You can “save” or “print” the word and its meaning by clicking at the above shown buttons.

7. Conclusion
The Buddism Dictionary was tested and is now completed. A number of CD ROMs were produced and distributed to various places. A positive response is received.

8. Reference
NEW TRENDS IN ENGLISH-CHINESE LEXICOGRAPHY

Gao Yongwei
Fudan University, CHINA
ywgao@fudan.edu.cn

Abstract: With the ever-increasing use of computers and corpora in the making of dictionaries, bilingual lexicographers in China have also jumped on the digital bandwagon, making full use of corpora and the Internet in the compilation and revision of English-Chinese dictionaries. The influx of monolingual pedagogical dictionaries into the Chinese market and the growing popularity of their bilingualized versions among the Chinese students have diminished the market share of those existing English-Chinese dictionaries. All these have become important factors in the emergence of new trends in English-Chinese lexicography. Such trends include the conversion of paper dictionaries into electronic ones, the incorporation of features typical of learner’s dictionaries, the increasing use of corpora, and the heavy reliance on the Internet. This paper attempts to explore these new trends, and special emphasis will be put on the appraisal of the forthcoming A New English-Chinese Dictionary (4th Edition).

Key Words: English-Chinese dictionaries; lexicography; corpus; the Internet

1. INTRODUCTION
Constant change has been a hallmark of the dictionary-making scene in China since the 1990s. Spurred by the phenomenal economic growth and an ever-increasing trend in learning English among Chinese students at all levels, numerous English dictionaries, large or small, general or specialized, have been published by publishers nationwide. Originally there were only a handful of publishers (e.g. The Commercial Press, Shanghai Yiwen Translation Publishing House, Foreign Language Teaching and Research Press, etc.) which were allowed to publish bilingual dictionaries, but now scores of them have received the authorization to do so. As a result, the number of dictionaries published each year has been on the rise. A byproduct of such increase, however, is that English learners have been overwhelmed by such a plethora of dictionaries and are usually at a loss to separate the wheat from the chaff. Although there were several traditional household names among English-Chinese dictionaries in the 1980s, some of such dictionaries, due to various reasons, have either fallen by the wayside or gathered dust. Lack of regular revision can be said to be the prime reason for the decline. Meanwhile, a large proportion of dictionaries published in China are imported or bilingualized versions of British learner’s dictionaries which have so far dislodged English-Chinese dictionaries compiled by Chinese lexicographers and have been gaining favor from Chinese students. Nevertheless, there are still some that have stood the test of time and stood out as the cream of the crop. A case in point is A New English-Chinese Dictionary (hereinafter abbreviated to NECD) which has so far undergone three major revisions since 1975 and boasts cumulative sales of 10 million copies. This paper attempts to discuss the trends
evident among in the making and revising of English-Chinese dictionaries, and NECD and ECD (short for *The English-Chinese Dictionary* which has been regarded as the most comprehensive dictionary of its kind in China) will be used as examples in identifying such trends.

All in all, there are four major trends that can be easily found in English-Chinese lexicography. They are: the conversion of paper dictionaries into electronic ones, the incorporation of features typical of learner’s dictionaries, the increasing use of corpora, and the heavy reliance on the Internet.

**2. ELECTRONIC DICTIONAIRES**

Since the late 1990s, hand-held electronic dictionaries (as opposed to CD-ROM dictionaries which have never been a big thing due to fear of piracy on the part of publishers) have been all the rage in China. According to surveys on dictionary use now available, electronic dictionaries have gradually replaced paper dictionaries as the main reference tool among high-school and college students. Brands such as Noah, Meijin, Besta, Bubugao, Wenquxing, and Casio have already become household names, and their sales far exceed those of their dead-tree counterparts. These electronic dictionaries are all based on one or several or even dozens of existing dictionaries. When it comes to English-Chinese dictionaries, the fixtures in these electronic dictionaries include the bilingualized versions of OALD (*Oxford Advanced Learner’s Dictionary*) and LDCE (*Longman Dictionary of Contemporary English*) and NECD. Let’s take NECD for example. Since the year 2001, Yiwen, its publisher, has signed contracts with at least ten electronic-dictionary publishers, granting them rights of use of NECD. So far NECD has been incorporated into best-selling electronic dictionaries such as Noah, Meijin, and Casio. The revenue brought in by the sales of such electronic dictionaries far surpasses that coming from the sales of the paper dictionary. The prospect of renewing its contract with publishers of electronic dictionaries through every revision and then getting lump-sum royalties has been irresistible for its publisher whose urge and insistence has made it all possible to publish the new edition of NECD within three years (2007-2009). Teaming up with electronic-dictionary publishers does not mean the demise of paper dictionaries. Most dictionary publishers are still cherishing the hope that, through the sales of electronic dictionaries, their brands will gain wider recognition among the users and the sales of their paper dictionaries will ultimately get a boost. Such willful thinking is by no means groundless. The mild success the second edition of ECD has been enjoying since its publication in April, 2007 can be partly attributed to its incorporation into two models of Casio electronic dictionaries (e.g. EW-V2800H) as Casio, a well recognized brand in China, usually carries out extensive advertising campaigns for its products.

In short, the conversion of paper dictionaries into electronic ones presents win-win situations for publishers of either paper dictionaries or electronic ones since most electronic-dictionary publishers are not content providers themselves. This practice has
shortened the cycle of revision for paper dictionaries, and NECD in particular. Otherwise NECD could face the same fate as its English namesake *A New English-Chinese Dictionary* (hereinafter abbreviated to EC) which is published by The Commercial Press and whose sales have all but stagnated due to the fact that few major revisions have been made to it since 1990s. Moreover, such a trend has so far influenced the way dictionaries are compiled or revised, which more or less arises from the unstated principle in the use of electronic dictionaries, that is, more is better as far as the content is concerned (more specifically the number of entries included in a dictionary). This may have accounted for the reason why NECD, a dictionary with a total of about 100,000 entries, has been wooed by so many publishers of electronic dictionaries. The revisionary policy of adding as many neologisms as possible is formulated out of this consideration. This results in NECD’s inclusion of thousands of new words or new senses, some of which have even not been included by general dictionaries in English-speaking countries. Neologisms of this kind include *buzzkill* (something or someone that spoils an otherwise enjoyable event), *kettle* (to contain demonstrators in a confined area), *netbook* (an economical notebook computer), *plus-one* (a friend or date whom one brings along to an event), and *putback* (a move in which a player scores a basket immediately after securing an offensive rebound), etc.

### 3. THE INCORPORATION OF LD FEATURES

It is no exaggeration to say that bilingualized dictionaries such as OALD and LDOCE have dominated the current Chinese market of dictionaries for users with intermediate or advanced English proficiency, and these users account for the majority of English users in China. According to industry insiders, the sales ratio between OALD and NECD is 6 to 1. In order to better equip itself in competition with bilingualized rivals, NECD has been sparing no effort to improve its quality and increase its competitiveness through incorporating features typical of monolingual learner’s dictionaries. Such features include the clear presentation and ordering of senses, the furnishing of sufficient collocational information, the use of authentic examples, the provision of abundant multiword lexemes, usage notes, etc. Let’s take a closer look at how NECD tries to match up with its rivals in these aspects.

Traditionally, when English-Chinese dictionaries were compiled, they were more or less based on American dictionaries. For example: In the preface to *A Comprehensive English-Chinese Dictionary* (1928), a pioneering work in the history of English-Chinese lexicography, its compilers listed Webster’s Collegiate Dictionary and Funk & Wagnall’s Practical Standard Dictionary as its main references; as a result, two or more subsenses may sometimes be found within a definition, such as “a device composed of bristles” and “an act of brushing” for the first definition of the entry *brush*. Later dictionaries such as EC, NECD, and ECD somewhat followed in its footsteps in the arrangement of senses. The problem is most acute in EC and NECD since the previous revisions failed to take it into consideration. A browse of the entries in the two dictionaries can find that the same policy was implemented in their compilation:
Through the current revision of NECD, the clustering of subsenses within one definition has been eradicated, thus giving a clearer presentation of the senses. A similar problem can be found in the ordering of senses in English-Chinese dictionaries. For most English-Chinese dictionaries, the ordering of senses seems to be a non-dict as almost no principle of all kinds has ever been advocated or consistently implemented. This is true of ECD, and this is also true of NECD. Sometimes the senses in NECD are arranged historically, most probably in imitation of dictionaries such as Merriam Webster’s Collegiate Dictionary as they were used as the main references. However, this unstated historical principle was never followed to the letter, and the ordering of senses in many entries borders on haphazardness, as can be seen from the entry below:

freak 1. a seemingly capricious action or event; a sudden and odd or seemingly pointless idea or turn of the mind 2. archaic a whimsical quality or disposition 3. a person or animal having a physical oddity 4. slang drug addict; hippie; an ardent enthusiast...

In contrast, the definitions in learner’s dictionaries are mostly arranged according to their frequency which, as dictionaries such as LDOCE claim, is determined through the analysis of corpora. Obviously, to reorder the senses in NECD becomes one of the top priorities of the revisers. However, a total reordering seems to be a Herculean task as some entries may include as many as 20 senses and there are no statistics available that can show how frequently each sense is being used. Besides, extensive editing within the microstructure may entail numerous future errors. Nevertheless, the revisers have gone all out to improve the ordering of the senses. As the result, in most entries, the commonest senses have been put first, thus facilitating the lookup process of dictionary users.

The importance of collocations in a dictionary can never be over-stressed. A.S. Hornby’s “prolonged labour of gathering collocations” (Cowie, 1999) is evidence enough of the importance of collocations in a learner’s dictionary. Subsequent effort in including as much collocational information in a learner’s dictionary has made collocation the strong suit of this type of dictionaries. However, information on collocations English-Chinese dictionaries used to provide is far from satisfactory. Compared to that of learner’s dictionaries, the amount of collocational information in NECD seems to be meager. In order to achieve the ultimate goal of transforming NECD into a dictionary that can meet both the decoding and encoding needs of English learners, the revisers have made full use of reference books now available on collocations such as A Practical Guide to the BBI Dictionary of English Word Combinations and Oxford Collocations Dictionary for Students of English. So far hundreds of collocations have been added to entries throughout the dictionary, such as be open to abuse, step on the accelerator, work his magic on sb., a carefully orchestrated demonstration, submit one’s resume, etc.
Illustrative examples have been an indispensable part of a dictionary entry. In the past, their number in an English-Chinese dictionary was almost negligible. But as Chinese lexicographers have more access to corpora, this kind of shortage has been alleviated. Nevertheless, the problem with English-Chinese dictionaries is that many of their examples are far from typical, with some even bordering on absurdity (e.g. *Ask the barren land for grain* in the entry *ask*, *Of the two contradictory aspects, one must be principal and the other secondary* in the entry *other*, and *treasure manpower and materials resources* in the entry *resource*). Obviously, the untypicality of the above-cited examples in NECED lies in their failure in illustrating the typical uses of the headwords. The revisers of NECED have made great effort in addressing this problem through the addition of typical illustrative examples such as *They shouldered their burdens in burden in burden, an electoral landslide in landslide, The economy is in deep recession in recession, and an economically viable project in viable*, etc.

When it comes to the provision of multiword lexemes such as idioms and phrases, learner’s dictionaries are known for their inclusion of the most commonly used MWLs. English-Chinese dictionaries such as NECED are not paled by comparison in the number of MWLs included. This is because NECED, from the very outset, has adopted an all-inclusive policy in its inclusion of phrases, idioms, and phrasal verbs. However, the problem with it lies in the inclusion of many expressions which have already gone out of date and therefore should be destined for the chop during the revision. Consequently, hundreds of obsolete expressions (e.g. *bow down in the house of Rimmon, fairy money, merry men of May*, etc.) have been deleted. Meanwhile, NECED tries to maintain the balance through the addition of new idioms, phrases or even proverbs such as *be firing blanks, on the bounce, The pen is mightier than the sword*, etc.

Howard Jackson says that “there is a recognition that learners need a range of information about words—grammatical, semantic and pragmatic—in order to be able to construct accurate and appropriate sentences in the target language” (Jackson, 2002). This recognition calls for the provision of usage notes in dictionaries. Although a common feature in learner’s dictionaries such as OALD, usage notes are definitely not a fixture in medium- or large-sized English-Chinese dictionaries, which might result from the assumption of their compilers that users of this kind of dictionaries are able to make the distinction themselves and notes of any kinds are superfluous. But now there is almost a consensus regarding the provision of usage notes in English-Chinese dictionaries. Chinese lexicographers have come to realized that usage notes are a plus and will prove to be extremely useful to Chinese learners of English. In the current edition of NECED, about 600 boxes containing usage notes (sometimes warning notes) which include information on the differentiation of synonyms, restricted use, etc. For example:

damp / moist / humid / wet

这些词均可形容湿：
■ 程度最重且可用于修饰人的词是 wet:  The leaves were wet with rain. / They are all soaking wet.
Another feature NECD tries to borrow from learner’s dictionaries is the separate listing of run-on entries that have obtained meanings that can not be deduced from the base words, e.g. derivative adverbs such as *basically* and *regretfully*, both of which deserve a full-entry status. Such derivatives were most treated as run-on entries in NECD. Learner’s dictionaries, however, treat them as full entries, and devote considerable space to the explanation of their uses. In the new edition, NECD has combed through all the run-on entries and has so far listed scores of such derivatives such separate entries.

4. THE USE OF CORPORA

The use of corpora has become a norm in the compilation of learner’s dictionaries. Since the first edition of *Collins Cobuild Dictionary of the English Language* was published in 1987, other British dictionary publishers such as OUP and Longman have jumped on the corpus bandwagon and created corpora of their own, some of which have surpassed 300 million words (e.g. Longman Corpus Network [300 million words], Bank of English [524 million words], Oxford English Corpus [2 billion words], etc.). Although the idea of using corpora in dictionary-making is gradually catching on in China, truly corpus-based dictionaries are absent in the Chinese dictionary market. *New Age English-Chinese Dictionary*, published by The Commercial Press in 2004, claimed to be based on a corpus, but its corpus only amounts to several million words and therefore it is insufficient to make any sound lexical analyses. Its stagnant sales are another indication of the unsuccessfulness of this corpus-based approach. Then how should Chinese lexicographers make use of corpora? To create one that is specifically designed to make dictionaries should be regarded as the holy grail, but it is out of the question due to lack of funds and the absence of concerted efforts among Chinese publishers and lexicographers. Then the only option left to Chinese lexicographers is how to take advantage of the existing corpora. Since few publishers nowadays are willing to fork over millions of yuan to start a new medium- or large-sized dictionary, Chinese lexicographers are engaged mostly in the revision of existing dictionaries. As a matter of fact, the use of corpora does not feature prominently in the revision of dictionaries. Perhaps the only thing revisers can derive from corpora is illustrative examples. The astonishing scarcity of illustrative examples Chinese lexicographers faced in the past resulted in many a head-scratching session. This may have accounted for the use of many unidiomatic examples that were characteristic of Chinese English. The access to corpora containing real English has definitely facilitated the search for good or suitable examples. The revisers of NECD have found a gold mine for illustrative examples in Sketch Engine (http://www.sketchengine.com) which contains more than a dozen corpora in different
languages. Some of the newly-added examples in NECD come directly from ukWaC, the largest English corpus with Sketch Engine, while others were created on the pattern of sentences from the corpus. Examples of this kind include: *a major hindrance to peace*, *They decided on a whim to get married*, *Months of rejections after job interview sapped him of his confidence*, *Her career had been sidelined for several years by motherhood*, etc.

Meanwhile, corpora may sometimes be used by Chinese lexicographers to check the currency of existing collocations or find out the frequency of collocations that were not included in the previous editions. Besides the use of corpora, Chinese lexicographers also resort to the prototype of corpora, namely citation files. Since 1996, the lexicographic team in charge of the revision for both ECD and NECD has been culling newspaper and magazine citations from various sources and put them in an ever-increasing computerized citation file. Although the file has so far accumulated thousands of cites totaling half a million words, a meager number in comparison with existing corpora, its usefulness cannot be underestimated. This is because such a citation file has not only become the source from which many new-word entries are obtained but also been used to help identify new meanings or provide suitable examples for new entries. A typical cite from the citation file can be found below:

A more hard-knuckled executive would have gone to the counterparties of those derivatives contracts and suggested that it would be a real shame if AIG were forced to file for bankruptcy, and offered some sort of workout. [2009/3/18 Washington Post]

5. THE HEAVY RELIANCE ON THE INTERNET

Chinese lexicographers of English dictionaries began to use the Internet as early as 1996. Back then, ECD’s revisers created a web page soliciting cites for new words when they were compiling *The Supplement to The English-Chinese Dictionary* (1996-1999). Rather modest as the appeal for new words was, it indicated Chinese lexicographers’ willingness to embrace the new technology. Since then, the use of the Internet has become an inseparable part of the work of a Chinese lexicographer. Word-watching through online reading has become a routine for some lexicographers as they believe that “alien sensitivity”, as put forward by Prof. Lu Gusun (Lu, 1998), ECD’s editor-in-chief, has been extremely helpful in the detection of new words. This practice has been consistently adopted in the revision of ECD and NECD as many neologisms, screened through this process, were included in English-Chinese dictionaries way ahead of their monolingual counterparts. Nowadays, Chinese lexicographers rely a lot on the Internet mainly for four things, namely to look up words in online dictionaries, to search for Chinese equivalents, to check the currency of new words, and to garner new cites or possible illustrative examples.

Lexicographers in China have more opportunities than before to make good use of existing online resources which may even prove to be a greater boon than existing paper-edition dictionaries. Dictionary sites such as Wiktionary ([http://www.wiktionary.com](http://www.wiktionary.com)) and Dictionary.com have already become a bonanza of
information on new words and phrases. Moreover, Urban Dictionary (http://www.urbandictionary.com) has also been proved extremely useful in providing edifying (and sometimes more confusing) information on colloquial and slang expressions. As a matter of fact, the definitions for dozens of new words recently included in NECD come from the above-mentioned sites. For example, the Chinese equivalent for the idiomatic phrase *connect the dots* is more or less based on the definition from thefreedictionary.com--“to understand the relationship between different ideas or experiences”.

The provision of Chinese equivalents for English-Chinese dictionaries has never been an easy task as the Chinese language, since the 1980s, has become more receptive to foreign languages and translations for English words often evolve with the passage of time. This sometimes means a complete overhaul of the translations when a dictionary is in revision. In most cases, the best place to find variations or better versions of such translations for English headwords is the Internet. As a keyword search through Google may come up with widely different translations, the revisers are sometimes spoilt for choice. On such occasions, experience counts. For example, the recalling of tainted *mozzarella* made big news worldwide and the various translations of the word frequently appeared in Chinese media; a Google search for the term came up with half a dozen translations such as 莫兹瑞拉, 马苏里拉, 美滋来乐, 莫扎里拉, and 莫泽雷勒 (which has been regarded as the proper equivalent according to the transliteration standards in Chinese), but the revisers chose 莫扎里拉 as the final equivalent on the basis of its frequency in mainstream media.

Once Chinese lexicographers have found a new word in their readings or come across one online, the ideal place to check whether the new word is widely used or not is undoubtedly the Internet. One can check the currency of new words either through Google or through various news archives (e.g. Guardian’s archive). Usually if one is able to find more than five quotations which span a period of over one year or more and which come from different sources, it is most likely that the lexicographers will include such words. Of course, there are many exceptions to this rule. Exceptions are made usually of words for which scores or even hundreds of instances can be found. For example, *cenbank* (short for central bank) is one of the latest additions in NECD and it is included on the basis of dozens of quotations like “Mexican Congress ratifies cenbank deputy governor [Reuters 2009-5-27]” and “Dutch cenbank says too early to speak about recovery [Forbes 2009-5-19]”.

The Internet has also been used by Chinese lexicographers as the main source of news cites. As the major US and UK media outlets have established their online presence, their websites have been the haunt of word-watchers. Daily culling of suitable news cites that might be used as the basis of illustrative examples for future editions has more or less become a routine. As the citation file grows, so does the ability of Chinese lexicographers in providing a better and up-to-date picture of the English vocabulary. *The Supplement to The English-Chinese Dictionary* is a precursor of this new approach to dictionary-making. About 85% of the 8,000 sourced illustrative examples included in the dictionary came
from online newspapers and magazines. During the revision of NECD, the revisers sometimes visit news sites such as Google News (http://news.google.com) to find suitable illustrative examples. As a result, dozens of such examples have been included in NECD, such as the economy could be tilting into a recession, a deliberate attempt to sabotage the peace talks, the governor resigned after his implication in a sex scandal, etc.

The four trends discussed above are by no means applicable to all the makers of English-Chinese dictionaries, but they are truly reflective of the work the lexicographers involving in the compilation and revision of the major English-Chinese dictionaries are doing. Coincidentally, three out of the four trends are made possible with technological advances. It can be safely predicted that with the development of technology the coming years will witness more changes in the dictionary-making scene in China, and some changes will inevitably involve the conversion of paper dictionaries into online ones.

REFERENCES

A COMPARATIVE STUDY OF PATTERN DESCRIPTIONS OF THAT-TAKING VERBS BETWEEN ENGLISH-JAPANESE LEARNERS’ DICTIONARIES AND MONOLINGUAL LEARNERS’ DICTIONARIES

Sumiyoshi Makoto
Setsunan University, JAPAN
sumiyosi@ilc.setsunan.ac.jp

Abstract: This paper presents a case study on the differences in verb pattern description between monolingual learners’ dictionaries (MLDs) and English-Japanese learners’ dictionaries (EJLDs), focusing particularly on that-taking verbs. While recent corpus-based MLDs give the user comparatively detailed information about verb patterns, EJLDs are often loaded with more meticulous information about the patterns that a particular verb takes. That-clause patterns, of which no mention is made in MLDs, are often given in the verb entries of EJLDs. In some cases, EJLDs inform the user of verb patterns not included in even the vast range of information from the Oxford English Dictionary. This can be attributed to the characteristics of dictionary making in Japan and the essential roles that EJLDs have to play in English teaching. In the traditional Japanese method of dictionary making, lexicographers have clarified interesting syntactic patterns of English words, gleaning data from a wider variety of data sources besides corpora. The finer descriptions of verb patterns based on the facts of English are essential because EJLDs have to be tuned towards learners’ receptive as well as productive purposes. Although MLDs avoid mentioning such patterns in the entries due to them only rarely occurring in corpora, the patterns illustrated in the EJLDs are extremely important for Japanese learners of English. Moreover, identifying as many verb patterns as possible is a significant contribution that EJLDs can make to theoretical linguistics, correcting the misjudgments that theoretical linguists have made about the behavior of verbs.

Key Words: MLDs; English-Japanese learners’ dictionaries; patterns; that-taking verbs

1. INTRODUCTION

A. S. Hornby’s *Idiomatic and Syntactic English Dictionary*, published in Japan in 1942, is remembered as the first MLD that provided detailed information on verb patterns that could be used by EFL students for their encoding and decoding requirements. MLDs published since then have followed this path with their own innovations regarding how such information should be codified or described in the dictionaries. Since large-scale corpora started to be used for dictionary-making, MLDs have been giving fuller
descriptions of verb patterns. Detailed verb pattern descriptions have become one of the distinctive features of MLDs.

EJLDs soon adopted the idea of verb patterns initiated by Hornby. Japanese lexicographers have since devoted themselves to refining the descriptions of verb patterns that are to be included in EJLDs. It can be observed that only EJLDs contain information regarding patterns of a particular verb. EJLDs are often loaded with verb pattern descriptions, based on authentic data Japanese lexicographers have collected themselves, leading to the uniqueness of EJLDs. It is no exaggeration to say that Japanese lexicographers, being partly influenced by MLDs, theoretical linguistics, applied linguistics and corpus linguistics, have successfully reached the stage where EJLDs are something more than just a translation of MLDs (Yagi, 2006).

The purpose of this study is to make a comparative study of verb pattern descriptions between MLDs and EJLDs, especially focusing on that-taking verbs. That-taking verbs are defined as verbs that can occur with that-clauses, as in He knows that she is right. The comparison will make clear that, in some cases, EJLDs provide more meticulous information about that-taking verbs, which is valuable for Japanese learners of English for their receptive as well as productive purposes. In addition, what EJLDs have revealed can be a great contribution to theoretical linguistics, in that the findings can be used to correct misjudgments of theoretical linguists’ about verb behavior.

2. WHY VERB PATTERNS MATTER?
Acquiring greater familiarity with verb patterns is significant and vital in two points for EFL students.

2.1 For Sentence-building
Hornby (1975) argued that advanced EFL students need to become more acquainted with the patterns of English structures. While teaching English to Japanese students, he realized that they could put words together in the right order if they were more familiar with the patterns that each word takes. Emphasis was placed on verb patterns, in particular, and his verb patterns were developed to help students to use English specifically for their encoding purposes. Since then, lexicographers have devoted themselves to developing the ways of presenting verb patterns in an easy-to-understand and user-friendly manner in MLDs. Cowie (1999) and (2009) make detailed comparisons of verb pattern descriptions in representative MLDs such as OALDs, LDCEs, and Cobuild.
Moreover, a large number of lists, books and dictionaries of verb patterns have been published so far regardless of linguistic persuasion; Alexander & Kunz (1964), Levin (1993), Francis et al. (1995), Herbst et al. (2004), to name a few. While Levin (1993) is based largely on the insights intuitively brought to light by theoretical linguistics, most of those published so far are corpus-based and give as many verb patterns with higher frequencies as possible. Interestingly enough, it is often the case that different verb patterns are given in the same verb entry, or a certain syntactic pattern that a certain verb takes is listed in one dictionary but not in another. However, the aim shared by these publications is clear: to give learners or researchers of English as detailed information as possible about verb behavior. These books are intended to guide EFL students through sentence-building, telling them in which pattern a particular verb can or has to be used.

Behind this proliferation of the materials of verb patterns is the recognition of “the close interrelationship between grammar and lexis” and “the central role of verbs…in the overall structure of sentences”. (Herbst and Klotz, 2009: 219ff.) Hornby also emphasizes in the forward of his 1975 book that “the most important patterns are those of the verbs”. Without knowledge of verb patterns, EFL students would have great difficulty in communication. It is natural that MLDs, which are expected to help EFL students express themselves in English, present as many verb patterns as possible on each verb entry with schematized easy-to-understand descriptions.

2.2 For Decoding Purposes

Atkins and Rundell (2008: 4) is right in saying that “[t]here is an enormous body of linguistic theory which has the potential to help lexicographers to do their jobs more effectively and with great confidence”. The enormous body of linguistic theory includes valency, frame semantics and so on (pp.130ff.). Included among them is the idiom principle, advocated by Sinclair and followed by his colleagues. Francis et al. (1995) and Hunston and Francis (1999) have clarified that there is a close connection between words (verbs, adjectives and nouns) and the patterns they reside in and that “patterns and lexis are mutually dependent, in that each pattern occurs with a restricted set of lexical items, and each lexical item occurs with a restricted set of patterns”. (Hunston and Francis, 1999: 3) They argue that a word with a particular meaning appears in a pattern with a particular meaning that is compatible with the meaning of the word. The association between words and patterns is close enough to argue that “the sense of a word appears to be determined by the pattern in which that word appears” (p. 105). When we come across sentences such as It was big of you to take the risk, which shows an idiosyncratic use of big, we can correctly guess that the adjective has the evaluative meaning of courageous or generous, because it occurs in the pattern it v-link ADJ of n to-inf, which is used to
evaluate the action done by others. A word occurs with a particular pattern because the word is used in a particular meaning compatible with the pattern. To put it reversely, it is possible to identify the meaning of a word used in a particular pattern for the very reason that the word appears in the pattern. That is, patterns also help learners in terms of decoding purposes as well.

The knowledge of patterns and their semantic association with words can be a very powerful “weapon” for EFL students. When they meet sentences governed by polysemous verbs such as observe and appreciate followed by that-clauses, the very fact that they are followed by that-clauses serves as a determining factor regarding the meanings of the sentences.

It is no exaggeration to say, then, that the onus is on learners’ dictionaries to provide EFL students with detailed information about the patterns of each word. Verb patterns in particular are to be described in detail, because they affect the syntactic configurations of sentences. Hence, “[s]ince Palmer…first introduced the notion of verb patterns, the provision of syntactic information has been fundamental to the MLD tradition” (Rundell, 2008: 234) and “[t]he facts in the valency description are the most important facts that the lexicographer needs to be aware of when writing the dictionary entry” (Atkins and Rundell, 2008: 149).

3. ENGLISH-JAPANESE LEARNERS’ DICTIONARIES

While MLDs have been at the center of extensive discussions as to how they have developed into something useful for learners of English and how they should be revised into something far more useful for them, the current state of EJLDs has not been as widely recognized in the world as MLDs. This is probably due to the fact that they enjoy a limited circulation only in Japan and the fact that the detailed analyses of EJLDs are usually published in Japanese.

A brief history of the earlier days of English-Japanese Dictionaries is discussed by Kojima (1999). Ikegami (2006) gives a concise overview of the history of English lexicography in Japan. Yagi (1996) and (2006) (the latter includes an historical overview of EJLDs) point out a large number of flaws in their descriptions, and propose the means of rectifying them based on empirical data. Yagi argues that there is still much room for the improvement of EJLDs so that they are able to correctly mirror present-day English. Lexicographers in Japan, with the aid of data from corpora (such as the BNC or the ones of their own) and citations they have collected, try hard to load EJLDs with enough distinctive characteristics to stand out in the competitive market.
EJLDs, while influenced by MLDs published so far in Britain and the US, have followed the path Palmer and Hornby set out and are developing their own characteristics that are thought to benefit Japanese students of English in terms of enhancing their English ability for their decoding and encoding requirements. One of EJLDs’ characteristics is that dictionary makers in Japan have been devoting themselves to including phraseological aspects of language use in EJLDs (Yagi 2008). Hidesaburo Saito’s *Idiomological English-Japanese Dictionary* (published in 1915) provided some 5,000 collocations for the collocations project initiated by Palmer and later joined by Hornby (Cowie 1999: 53). *Sanseido’s Global English-Japanese Dictionary* (published in 1986) contained specially-made box spaces about collocations on some lexical entries. Syntactic patterns that verbs, adjectives and nouns take are described as accurately as possible by EJLDs. The *Shogakukan Youth Progressive English-Japanese Dictionary* (published in 2004) is loaded with meticulous information about syntactic patterns based on the data the dictionary makers gleaned. Finer pattern descriptions have been a tradition of EJLDs since the earliest stages of publication.

4. COMPARISON

In this section, comparisons are made between MLDs and EJLDs in terms of the pattern descriptions of English *that*-taking verbs. The EJLDs investigated in this study are targeted for students of intermediate to advanced levels and can be regarded as the counterparts of the MLDs used for this survey. They will be referred to with only part of their names (see references for their full names).

4.1 Verbs Of Manner Of Speaking

Verbs of manner of speaking are verbs that refer to manner of vocal production. A plethora of verbs are included in this semantic category, and they are classified with various names. Levin (1991) calls them verbs of manner of speaking, while Dixon (2005) uses the term SHOUT subtype. Whatever name they are given, verbs that can be classified as such show a wide range of syntactic behavior. It is argued in the literature that members in the same semantic group share syntactic behavior with one another; however, it is hard to identify which is a *that*-taking verb of manner of speaking (Sumiyoshi 2003). This being the case, dictionaries do not always concur in terms of which should be given as a *that*-taking verb.

EFL students may well come across the following examples where the verbs *shout* and *bellow* are followed by *that*-clauses:

(1) About once a week, Uncle Vernon looked over the top of his newspaper and **shouted** that Harry needs a haircut. (J. K. Rowling, *Harry Potter and the Philosopher’s Stone*)
(2) She kept **bellowing** at me **that** she should call the doctor and I kept telling her not to. (S. King, *The Body*)

Let us see to what extent MLDs can be helpful to a foreign learner of English who encounters these sentences and wants to check the uses or meanings of these verbs with *that*-clauses. “Helping learners” is what MLDs are for. *OALD₇* and *CALD₃* give information about the verb *shout* taking a *that*-clause, while *LDCE₅*, *LAAD₂*, *Cobuild₆* and *MED₂* do not. The MLDs unanimously give no information of *that*-clause patterns in the entry for *bellow*. This discrepancy regarding descriptions and the lack of information is probably due to the results of corpus consultation being made while each dictionary was produced. However, the exclusion of the *that*-clause pattern from these verb entries reduces the dictionaries’ usefulness for EFL students, who may well read the Harry Potter series and other novels.

Table 1 shows whether each MLD has the *that*-clause pattern description in each entry for the verbs indicated. A + mark in the table indicates that the dictionary gives information about *that*-clause patterns for each verb entry. Only *OALD₇* describes the *that*-clause pattern within all the entries, whereas others sporadically mention the pattern. The scant *that*-pattern descriptions in MLDs sharply contrast with Levin (1993), which argues that all verbs of manner of speaking in her list are perfectly compatible with *that*-clauses. They also stand in striking contrast with *that*-pattern descriptions in EJLDs shown in Table 2.

Table 2 shows that EJLDs provide more meticulous information about the *that*-clause pattern in the cases of verbs of manner of speaking surveyed. *Shogakukan Youth Progressive English-Japanese Dictionary* offers the most in-depth *that*-clause pattern descriptions. By and large, Japanese students of English find it more helpful to depend on EJLDs than MLDs in these cases.

<table>
<thead>
<tr>
<th>Verb</th>
<th>LDCE₅</th>
<th>LAAD₂</th>
<th>OALD₇</th>
<th>Cobuild₆²</th>
<th>CALD₃</th>
<th>MED₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>bellow</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>grumble</td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>mumble</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>murmur</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>matter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shout</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>whisper¹</td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Verbs of Manner of Speaking and That-clause Patterns in EJLDs

<table>
<thead>
<tr>
<th>Verb</th>
<th>Genius₄</th>
<th>Wisdom₂</th>
<th>Youth Progressive</th>
<th>O-Lex</th>
<th>Luminous₂</th>
<th>Royal</th>
</tr>
</thead>
<tbody>
<tr>
<td>bellow</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>grumble</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>mumble</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>murmur</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>mutter</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>shout</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>whisper₁</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

4.2 Gesture Verbs

Some verbs relate to people moving their hands, arms, head and so on. They are semantically defined as gesture verbs. Included in this category are verbs such as nod, gesture, motion, wink, wave and so forth. Gesture verbs have received unfair treatment in the literature including pattern/valency dictionaries. For example, no mention is made of gesture verbs in Levin (1993) despite over 3,000 verbs being dealt with in the book. No mention, either, is made of them in Herbst et al. (2004). It is natural, then, that EFL students expect these verbs to be dealt with at length in dictionaries when they look for information concerning their behavior.

Suppose that learners want to get the information about the pattern used in the following citations, which come from popular novels. They would probably be disappointed to find that no information is available about this pattern in LDCE₅, LAAD₂, OALD₇, Cobuild₆, CALD₃ and MED₂. The pattern nod that… is not described in the verb entry of the Oxford English Dictionary, either:

(3) “They off? Huh? Huh? They off me, Gordie?”

His eyes kept going past me, as wide and white as the eyes of a plaster horse on a merry-go-round.

I nodded that they were and just kept crying. (S. King, The Body)

(4) “Ready?” questioned Dr. Navarre.

Although Marissa nodded that she was, she would have preferred to put off confronting the patients. (R. Cook, Outbreak)

Instead, EJLDs serve as a primary and valuable information source for Japanese learners of English. All the EJLDs but Luminous₂ in Table 2 give this pattern in the entry nod, some of them illustrating this use of the verb in sentences such as He nodded that he understood/She nodded that she was pleased with it.
The other gesture verbs are expected to be dealt with in learners’ dictionaries, too, since, as mentioned above, valency dictionaries, pattern grammars and other materials do not describe the uses of these verbs in as much detail as learners expect. *Gesture*, for example, receives proper treatment only in *OALD*7, which provides the verb pattern [V that] followed by the illustrative examples *He gestured (to me) that it was time to go*. All the EJLDs but *Luminous*2 again give this pattern description in the entry for the verb.

### 4.3 Apologize

It is intriguing to keep track of the changes in pattern descriptions for the entry for *apologize* in MLDs. The *Cambridge International Dictionary of English* (published in 1995 and later revised as *Cambridge Advanced Learner’s Dictionary*) showed that *apologize* is used with a *that*-clause by giving the grammar pattern [+ that clause], which was instantiated by the sentence *He apologized that the statistics had been inaccurate*. This description was eliminated from the following revised editions. Other MLDs published so far do not include the *that*-clause pattern in the entry for *apologize*. This information is not mentioned in the *OED* entry for *apologize*. Editors of MLDs may have assumed that the use of the verb followed by *that*-clauses does not merit mention due to its low frequency. Interestingly enough, however, the *Oxford Learner’s Thesaurus*, which characteristically shows patterns and collocations alongside with synonymous words, has information on the *that*-clause pattern within the entry for *apologize*.

It cannot be denied that *apologize* occurs in the *that*-clause pattern, as witnessed by the following citation from a newspaper:

(5) Hadi answered with patience. He *apologized that* the union had yet to secure jobs in Baghdad. (*The WashingtonPost*, 08/March/2004)

This use of the verb with *that*-clauses can be attested in corpora such as the Oxford English Corpus.

Some of the EJLDs describe the use of the verb with *that*-clauses. *Youth Progressive* displays this information as follows in the entry for *apologize* (the Japanese translations are omitted in (6)):

(6) *apologize that*… (▶ also with speech) || He *apologized that* the report had been delayed.

The information given in (6) indicates that the verb can be used with direct speech as well as *that*-clauses. *Youth Progressive*’s description of this use of the verb is more meticulous than other EJLDs. This is attributed to the editor’s eagerness to reflect the current uses of English words in every single description of the EJLD in as much detail as possible. If
Japanese learners of English want to know how this verb is used with *that*-clauses and what it means in conjunction with the pattern, they have no choice but to consult EJLDs rather than MLDs, although both are expected to fulfill learners’ decoding requirements.

### 4.4 Other Verbs

There are many other *that*-taking verbs that merit mention. A few of them are dealt with here for a further comparison between MLDs and EJLDs, which will certainly underlie their future improvement.

*Advertise*, which means to tell the public about a product or service in order to encourage them to utilize or purchase it, can be seen to occur in the *that*-clause pattern, as shown by an illustrative citation from a weekly campus newspaper issued at a university in the US:

(7) Such beverages, including Red Bull, Venom, Adrenaline Rush, 180, ISO Sprint and Erektus, **advertise that** people who consume these drinks will have increased energy to help them through the day. (*The Lumberjack*, Jan.16-22, 2002)

Since the newspaper is targeted at university students, EFL students studying on campus are highly likely to read it. Since MLDs do not record the usage of the verb in the entry, however, they do not have any clues they can look to when trying to detect the meaning of the construction where *advertise* is followed by *that*-clauses. Note that some MLDs such as *LAAD*₂ and *Coubild*₆ show that the verb is used in the phraseology *advertise the fact that*…, although this means something quite different from *advertises that*… (the former means to show or tell people something about yourself or your secret) and usually occurs in a negative context (*Cobuild*₆). It is impossible for learners to identify the correct meaning of *advertise that*… just by deleting the noun phrase *the fact from advertise the fact that*… Japanese learners of English, fortunately, can depend on EJLDs, most of which, giving the corresponding Japanese translation and an illustrative example, tell them that the verb occurs in the *that*-clause pattern with companies or goods as the subject.

Descriptions of *that*-taking phrasal verbs are to be improved in MLDs. Hoey, in his introduction of lexical priming written for the study section of *MED*₂ (pp. LA12-LA13), carries out a brief case study on the expression *the time has come*. In the explanation, he uses the phrasal verb *work out* followed by a *that*-clause:

(8) They would subconsciously have **worked out that** about a third of the time the phrase occurs with verbs of thought and feeling:....

Hoey’s article, included in the learners’ dictionary as one of its appealing features, is expected to be read by EFL students. However, *MED*₂ fails to explain this use of *work out* in the entry. It seems quite strange that a dictionary would leave unexplained an
expression used in the dictionary itself. Phrasal verbs constitute a difficult area for learners of English, because they often mean something quite different from the meaning obtainable as a result of combining the meanings of the verb and the particle used. The onus is on learners’ dictionaries to give detailed syntactic and semantic information about them. *Genius*, the most widely used EJLD in Japan, succeeds in recording this use of the expression in the entry.

5. LEXICOGRAPHICAL CONTRIBUTION TO THEORETICAL LINGUISTICS

I began Section 2.2 by quoting Atkins and Rundell (2008: 4). To put their statement differently, I here argue that lexicography has the potential to help theoretical linguists to do their job more effectively and with greater confidence.

As attested in Section 4, verbs such as nod and apologize do take that-clauses. However, Rudanko (1989: 80) judges nod that… ungrammatical, arguing “nod does not verbal communication only, and, as predicted, it is incompatible with a that complement”. Examples (3) and (4) clean up his misunderstanding of this use of the verb. Moreover, Searle (1979: 15), Vanparys (1993: 73) and some prescriptive usage dictionaries argue that apologize cannot take that-clauses, which can be refuted by example (5) (see Sumiyoshi (2003) for a detailed descriptive study of the verbs under discussion, including other verbs that have not been touched upon here).

Lexicographers, as empiricists, depend on authentic data for analysis, and theoretical linguists use intuition for data. Exclusive dependence on intuition for linguistic data has been criticized for distorting the actual uses of English words (Meyer and Nelson, 2006; Schütze, 1996). Lexicographers can make a tremendous contribution, by using data they collect from corpora and other sources for dictionary-making, to correcting misjudgments of theoretical (or prescriptive) linguists’ about the facts of English.

6. CONCLUSION

EJLDs provide more detailed information about verbs that are likely to take that-clauses, at least as far as the verbs investigated in this paper are concerned. EJLDs’ closer attention to pattern descriptions can be attributed to the following facts; (i) Japanese lexicographers have been deeply aware of the importance of patterns in English learning since Hornby’s *ISED* was published in Japan (ii) they did not stop gleaning citations even after corpora started to play the central role in data collection for dictionary-making (iii) EJLDs opt to record as many English constructions as possible regardless of their frequencies because they have to give Japanese learners of English Japanese translations corresponding to them for learners’ decoding purposes.
Collecting citations as data for dictionary-making has sometimes been criticized on the grounds that people engaged in this work tend to notice atypical uses of words rather than common usages that are needed for dictionary-making. However, as discussed in Section 2, patterns are important for both encoding and decoding purposes. For encoding purposes, EFL students can express themselves with patterns of high frequencies. For decoding purposes, however, high frequencies indicate some but not all. When learners read, they do not choose a book written exclusively with words and expressions of high frequencies. Authors write using words and expressions in a varying range of frequencies. Learners’ dictionaries have to include what is necessary for learners. It is sometimes argued that “valency dictionaries are primarily relevant as encoding dictionaries for foreign users” (Herbst and Klotz 2009: 244); however, patterns can also help learners to identify the meanings of sentences due to the close association between words and patterns. One of the EJLDs’ jobs is to give learners meticulous pattern descriptions and the Japanese translations for decoding requirements. It is not surprising that EJLDs pay more attention to patterns of varying frequencies than MLDS do. What EJLDs need is different, in parts of pattern descriptions, from what MLDS require.

NOTES
(1) In all the MLDS investigated, whisper has two senses. One is “to say something very quietly so that other people cannot hear you” and the other is “to tell other people a piece of news or information that may or may not be true” (MED2). Some dictionaries indicate that the that-clause pattern can be used with the verb in the latter sense. However, the verb used in this sense is not included in the semantic category of verbs of manner of speaking, which accordingly is excluded from Table 1.

(2) It is interesting to see how the series of Cobuild dictionaries has recorded the that-clause pattern in the entries for shout and bellow. In the case of shout, the first edition (1987) made no mention of this pattern in its entry. The second edition (1995) mentioned in passing the fact that the verb is followed by that-clauses by showing [also V that] in the extra column. Francis et al. (1996: 99), which is based on the research carried out for the second edition (Hunston and Francis 1999: 32), accordingly includes shout in that-taking verbs. Notwithstanding that, the following editions (2001, 2003, 2005, 2009) exclude this pattern description. No information has been available on bellow used with the that-clause pattern in its entry since the first edition, although Francis et al. (1996: 99) points out that the pattern can be used with the verb.
REFERENCES


EJLDs


Abstract: Writing is seen as a difficult skill for students of English because they encounter many problems. However, to deal with the problems, a dictionary seems to be one of the essential language tools to help them. This study, therefore, was conducted in order to investigate the correctness of words chosen from Thai-English electronic dictionaries for writing. The subjects were six third-year KMUTT undergraduates majoring in science. The instruments used for data collection were the subjects’ written work, observation sheets and semi-structured interviews. The subjects were asked to do a writing task by using their Thai-English electronic dictionaries as a technical support. After the subjects finished the task, the researchers analyzed the data for correctness of words chosen from the Thai-English electronic dictionaries in terms of form, meaning and use. Furthermore, this study reveals the subjects’ stated reasons for their word choices. The results show that the subjects’ difficulties with vocabulary use derived from their own problems and from problems concerning the electronic dictionary. Moreover, several reasons were found for the subjects’ word choices: they had seen the chosen words before, they used their intuition to choose words and the chosen words matched the Thai meanings they wanted to convey. Suggestions and implications are provided with valuable directions for dictionary compliers and teachers of English.

Key Words: writing; electronic dictionaries; vocabulary use; word choices

1. Introduction
For Thai undergraduate students, writing seems to be a very difficult skill. What is more, they have to write a lot of assignments in English. One important language tool to help them produce English assignments is dictionaries. Nowadays, most students prefer to use electronic dictionaries because they are modern, small, light, and apparently easy to use.

Prior to conducting this study, the researchers observed that, when students write their assignments in English, they often seem unable to make optimal word choices. Students say that they use electronic dictionaries as the main tool to help them find target words. Nonetheless, the researchers wondered if students’ misuse of electronic dictionaries might precipitate mistakes in their English writing. One reason for this is that they may not know how to use electronic dictionaries very well.

This study thus aims to find out the answers to the following research questions:
1. Do students use translated equivalents chosen from Thai-English electronic dictionaries correctly? If not, why not?
2. What are their reasons for their word choices?
2. Literature review

This literature review focuses on problems, vocabulary and dictionary use in writing.

2.1 Writing problems

Writing in a foreign language can be problematic for language learners. Learners might encounter several problems throughout the process of writing a document. Reflecting that second-language (L2) writing problems are widespread in Thailand, there are many studies investigating problems in Thai university students’ writing.

Chinnawong (1999) studied students’ ability in academic writing and found that, after grammatical errors, the students made most mistakes with vocabulary (e.g. spelling, word choices). Lakkhunaprasit’s (1999) study investigated the writing ability of first-year students. She found six aspects where her subjects made mistakes: structure/grammar, run-on sentences, word choice, prepositions, spelling and capitalization. Likewise, Khao-Urai (2002) studied errors in the English essays of fourth-year students and found that the most frequent errors were grammatical while the next most frequent errors were syntactic and lexical. Her category ‘lexical errors’ included 1) spelling, 2) literal translation from Thai (first language, L1) to English (L2), 3) overgeneralization of the use of one translation equivalent, and 4) using general lexical items.

As we can see from this research, students have many problems in writing, including vocabulary use.

2.2 Vocabulary in writing

Read (2000) proposes that being proficient in L2 is not just a matter of knowing a lot of words or grammar rules but being able to use that knowledge effectively for various communicative purposes. Therefore, L2 learners need to show that they can use words appropriately in their own writing rather than just demonstrating that they understand what a word can mean. Thus, lexical knowledge, which L2 learners need to master, includes knowledge of form (pronunciation, spelling, word derivation), position in grammatical constructions, collocations, functions (frequency and appropriateness), and association (Nation, 1990; George, 1983, cited in Oxford and Scarcella, 1994). The aspects of word knowledge which are relevant to this study (form, meaning, and use) will be elaborated below.

Form. For the written mode, learners have to know how words should be spelled. Moreover, they need knowledge of how to combine elements of words to create other lexical items. For example, learners learn the rules of building different forms of related words such as forget, forgot, forgotten, forgetting and forgetful (Nation, 2001).

Meaning. When we think about words, we should think of their semantic features, too. For example, the words drink, sip, gulp...down, swig refer to the action of drinking but have different features, and each feature contributes to what are called ‘shades of meaning’ (Nation, 1990).

Use. There are two kinds of knowledge of word use relevant to this study: grammatical functions and collocations.

Therefore, in this study, we aim to investigate how students use vocabulary in writing by choosing words from Thai-English electronic dictionaries, and to check whether they can use these words correctly and appropriately in contexts.
2.3 Using dictionaries in writing

Nesi and Meara (1994) studied patterns of misinterpretation in the productive use of EFL dictionary definitions. Their subjects were 52 non-native speaker adults, most of whom studied in English language and study skills programmes. It was found that there were three categories of errors: semantic (the majority); grammar and usage errors (the next most frequent category); and failure to use the target word. Furthermore, they found that the subjects systematically misinterpreted dictionary entries and this affected the correctness of their writing. Christianson (1997) also studied dictionary use by EFL writers. The subjects were 51 Japanese university freshman EFL students majoring in computer science. They were assigned to do writing assignments while using a dictionary as a tool to assist them. They were also asked to underline all of the words that they looked up in any dictionary (L1-L2, L2-L1 and L2-L2). Forty-two percent of the underlined words were found to have been used wrongly in some ways; there were errors concerning word choices, prepositions, articles, pluralizations, spellings, word forms and tenses.

East (2006) studied the impact of bilingual dictionaries on lexical sophistication and lexical accuracy in tests of L2 writing proficiency. This research aimed to investigate whether using a bilingual dictionary enhances good writing in German writing tests. The subjects were 47 school students in New Zealand, aged around 17-18 years. The subjects took two tests, one with and one without a bilingual dictionary. The two tests were compared in terms of lexical sophistication, lexical accuracy and test score. The results revealed an increase in lexical sophistication in ‘with dictionary’ tests but frequent misuses of look-ups. There was also no improvement in test scores. Types of words looked up in the dictionary were nouns, verbs, adjectives, adverbs, phrases, and other items. It was found that 50% of words looked up were used wrongly, of which 51.5% were ‘wrong word’ and 48.5% were ‘wrong form’.

For Thai students, Boonmoh (2003) studied problems using electronic dictionaries to translate Thai written essays into English. The undergraduate students compose English assignments by writing in Thai first and then translating into English. His study found that, because of translation, their writing contained many mistakes, both in terms of forms and meanings (Boonmoh et al., 2004); moreover, the quality of their written work was quite poor partly because it contained many poor word choices as well as grammatical mistakes (Boonmoh et al., 2006).

In conclusion, to be able to write accurately, learners should have good knowledge of the target language; this means learners should have good understanding about its spelling, meaning, grammatical structure, collocation and word choices. Dictionaries have many valuable pieces of information which learners can use in writing.

3. Research methodology

3.1 Subjects

A questionnaire was used to survey how students use Thai-English electronic dictionaries and, from the students’ answers, the subjects were selected based on the criteria below.

1. The subjects’ grades on the previous English course they attended had to be between B and C+.
2. The subjects owned and used the same brand name and same version of electronic dictionaries.
3. The subjects always used Thai-English electronic dictionaries to help them complete their written work.

Finally, six subjects (5 males and 1 female) were selected. All the subjects took LNG 104 (Content-based Language Learning) as a compulsory English course in the first semester of 2006. They were experienced in writing short paragraphs. They were from the Faculties of Industrial Education, Engineering, and Science.

3.2 Technical support
According to the results of the survey questionnaire, one brand of the electronic dictionaries that every subject normally used was selected. This brand is used extensively because it is popular among students. However, the number of words contained and the corpus used are not mentioned either in its manual or on its box. There are three main functions in this electronic dictionary (each with sub-functions): dictionary (e.g. English-Thai Dictionary, Thai-English Dictionary, English-English), organizer and calculation. There are some additional functions such as Download, Games, and Reference.

3.3 Instruments
• Subjects’ written work
In order to answer Research Question 1, the subjects were asked to write for two hours on the topic of Computer Games. The content should contain a favorite computer game, its characteristics, how to play it and reasons why they like that computer game.

• Observation sheets
While the researcher was observing the subjects, she used blank observation sheets as supporting instruments to record four things: every Thai word which the subjects looked up in the Thai-English electronic dictionary; English word(s) which they found and did not find in the dictionary; the process which the subjects were doing on their written work; and the new word(s) which the subjects used to replace the old word(s) in their written work.

• Semi-structured retrospective interviews
The semi-structured retrospective interview was used to get information about the subjects’ reasons for using or not using particular words after consulting their electronic dictionaries (Research Question 2). With the aid of the data from the observation sheets, the first researcher interviewed each subject as soon as the writing activity was completed.

3.3 Procedures
The first researcher asked the subjects to come one by one to do the writing activity. First, the instructions were explicitly given. Then, the subjects were asked to do the activity in English, using their electronic dictionaries as much as they wanted; moreover, they could use whatever sub-functions of the dictionaries they wanted. However, the researcher would focus on words the subjects looked up in the Thai-English dictionary only.

While writing, the first researcher did not interrupt the subjects. They had to underline words which they looked up in their dictionaries and write Thai above those words to show what meaning they really wanted to convey. The subjects had to write numbers in chronological order to identify the word which they looked up first and those they looked up subsequently. The first researcher sat beside each subject to jot down any words they looked up.
Each subject was then interviewed by the first researcher in Thai for about thirty minutes. During the interview, the researcher asked each subject about their word choices in terms of form, meaning, and use. The subjects were also asked to clarify any unclear points while they were using the electronic dictionaries.

3.4 Data analysis

- **Data from the subjects’ written work**
  Each aspect of the words chosen (form, meaning and use) was calculated into percentages to show what aspects of words the subjects used correctly most often and what aspect they needed to improve most.

- **Data from the observation sheets**
  To analyze the data from this instrument, some important terms used in this study must be defined: words the subjects looked up in their dictionaries are called ‘words sought’; if the dictionary provided one or more translated equivalents, these words are then called ‘words found’; and if the subjects chose any of the words found and used them in their written work, they are called ‘words chosen’.

  Analysis focused on the total number of words the subjects sought in their dictionaries and, to show whether or not the dictionaries can provide words the subjects needed, differentiation was made between the words found and those not found. Then, from the words found, those chosen and not chosen by the subjects for their written work were counted. The frequencies of each word category mentioned were calculated into percentages.

- **Data from the interviews**
  The subjects’ reasons for their word choices were categorized, counted and calculated into percentages. The subjects’ word choices came from two sources: words from their own lexicons and words chosen from the dictionaries.

4. Data presentation

The data obtained from the record forms of word searching, the observation sheets, the subjects’ written work and the retrospective interviews are now presented.

4.1 Word searching via the Thai-English electronic dictionary

From Table 1, it can be seen that, from the subjects’ written work, there were 39 words the subjects sought in their dictionaries. According to the frequencies of word searches shown, the subjects could be divided into two groups: the group that searched more words, and the group that searched fewer words. Subjects F, B, D and E had similar frequencies of word searches, which were from three to five words, while Subjects A and C had frequencies of word searches from nine to fourteen words. Thus, although they wrote on the same topic, there was a big gap in the frequencies of word searches between the two groups.
Table 1: Frequencies of subjects’ word searches

<table>
<thead>
<tr>
<th>Subject</th>
<th>Words sought</th>
<th>Words found</th>
<th>Words not found</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>9</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>14</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Total (%)</td>
<td>39 (100.00)</td>
<td>30 (76.92)</td>
<td>9 (23.08)</td>
</tr>
</tbody>
</table>

From Table 2, another way to look at the words found is that 70% (21/30) were chosen by the subjects whereas 30% (9/30) were not; however, there was considerable individual variation among the subjects in terms of the proportion of words found that were actually chosen. While Subjects A, B, C and D did choose all the words given in the dictionary, Subjects E and F chose every word they found.

Table 2: Frequencies of words chosen and not chosen in written work

<table>
<thead>
<tr>
<th>Subject</th>
<th>Words found</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Words chosen</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>9</td>
</tr>
<tr>
<td>D</td>
<td>3</td>
</tr>
<tr>
<td>E</td>
<td>4</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
</tr>
<tr>
<td>Total (%)</td>
<td>21 (53.84)</td>
</tr>
</tbody>
</table>

Note: The percentages in the table are calculated based on words found, which were 76.92% of words sought (see Table 1).

4.2 Correctness of words chosen in terms of form, meaning, and use

Table 3 shows the frequencies of words chosen from the electronic dictionary that were correct in the students’ work in the aspects of form (100.00%), meaning (90.48%) and use (57.14%). According to these results, they can be categorized into three groups: correct form, meaning, and use; correct form and meaning; and correct form only.

Table 3: Words chosen from Thai-English electronic dictionary

To elaborate and make the description clear, there are examples from the subjects’ written work to support. The bolding and underlining in the extracts relates to the words chosen in the subjects’ written work.
Subject C used the word ‘murderers’ (ฆาตกร) in a sentence. He wanted to convey that the main character, a police officer, saw his wife and son who had been killed by murderers. He also intended to convey the word ‘murderer’ in its plural form to tell readers that there were many murderers.

“He encountered his son and wife be killed by murderers.”

(Subject C)

Subject D used the word ‘create’ (สร้าง) in a sentence. He wanted to tell his readers how to play the game. The sentence is wrong as ‘s’ should not be added to the word ‘create’ because it follows the word ‘must’.

“The player must creates and upgrades the army for destroys the enemy.”

(Subject D)

Subject A used the word ‘act’ (ท า) in the sentence below. It is wrong because, to convey his intended meaning, he should have used the word ‘follow’. Actually, he wanted to say that, if a player wanted to change his level, he should follow the requirements of the game. Furthermore, ‘ing’ should be added to the word ‘act’ because it follows the word ‘by’.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Word chosen</th>
<th>Form</th>
<th>Meaning</th>
<th>Use (Types of mistakes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>act (ท า)</td>
<td>√</td>
<td>X</td>
<td>X (gerund)</td>
</tr>
<tr>
<td>B</td>
<td>rifle (ไรเฟิล)</td>
<td>√</td>
<td>√</td>
<td>X (plural noun)</td>
</tr>
<tr>
<td>C</td>
<td>upright (ซื่อตรง)</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>encounter (พบ, พยายาม, ประสบ)</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>murderer (ฆาตกร)</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>try (พยายาม)</td>
<td>√</td>
<td>√</td>
<td>X (parallel structure)</td>
</tr>
<tr>
<td></td>
<td>storm (พายุ)</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>role (บทบาท)</td>
<td>√</td>
<td>√</td>
<td>X (collocation)</td>
</tr>
<tr>
<td></td>
<td>scene (ฉาก)</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>graphic (กราฟิก)</td>
<td>√</td>
<td>√</td>
<td>X (plural noun)</td>
</tr>
<tr>
<td></td>
<td>exciting (น่าตื่นเต้น)</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>D</td>
<td>create (สร้าง)</td>
<td>√</td>
<td>√</td>
<td>X (modal verb)</td>
</tr>
<tr>
<td></td>
<td>enemy (ศัตรู)</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>quality (คุณภาพ)</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>E</td>
<td>restricted (ก าหนด)</td>
<td>√</td>
<td>X</td>
<td>X (collocation)</td>
</tr>
<tr>
<td></td>
<td>apply for (สมัคร)</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>fill in (กรอก)</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>enjoy (สนุก)</td>
<td>√</td>
<td>√</td>
<td>X (gerund)</td>
</tr>
<tr>
<td>F</td>
<td>consolidate (เสริมให้แข็งแกร่ง)</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>position (ฐานะ)</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>achievement (ผลงาน)</td>
<td>√</td>
<td>√</td>
<td>X (plural noun)</td>
</tr>
<tr>
<td>Total (%)</td>
<td>21 (100.00)</td>
<td>19 (90.48)</td>
<td>12 (57.14)</td>
<td></td>
</tr>
</tbody>
</table>
When level of your character is level 15, you can change class by talk with NPC (non player character) and act with quests from NPC.”  

(Subject A)

In conclusion, the subjects’ written work appeared good in terms of word form and fairly good in terms of word meaning because they could see the spelling of the words and check the meaning of the words from the electronic dictionary; however, it was less good in terms of word use.

4.3 Reasons for choosing words in Thai-English electronic dictionary

This section presents the subjects’ reasons for their word choices as revealed by the interview data.

According to Table 4, the subjects gave 12 reasons for their word choices, though it can be seen that there were three main reasons. Firstly, the subjects had already seen how the words were used (23.53%); equally, the subjects used their intuition (23.53%) to consider whether they should use the words or not; next, the subjects pointed out that they chose words because these words matched the Thai meanings they wanted to convey (17.65%). For the least frequent categories (1.47%), it was found that the subjects used the chosen word because it matched the part of speech they wanted; the chosen word was the first word given in the electronic dictionary; and the dictionary also showed how to use the word.

Table 4: Reasons for choosing words in Thai-English electronic dictionary

<table>
<thead>
<tr>
<th>Reasons</th>
<th>Subject</th>
<th>Total (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S has seen the chosen word before.</td>
<td>A: 0 B: 1 C: 7 D: 3 E: 3 F: 2</td>
<td>16 (23.53)</td>
</tr>
<tr>
<td>S’s intuition (e.g. without providing clear evidence, S feels a chosen word is better than other words in the context).</td>
<td>A: 3 B: 0 C: 3 D: 0 E: 7 F: 3</td>
<td>16 (23.53)</td>
</tr>
<tr>
<td>The chosen word matches the Thai meaning S wants to convey.</td>
<td>A: 0 B: 0 C: 7 D: 3 E: 1 F: 1</td>
<td>12 (17.65)</td>
</tr>
<tr>
<td>S has checked the meaning of the chosen word in the English-Thai electronic dictionary.</td>
<td>A: 0 B: 2 C: 0 D: 0 E: 3 F: 5</td>
<td>5 (7.35)</td>
</tr>
<tr>
<td>Only one word is given in the e-dict.</td>
<td>A: 0 B: 1 C: 1 D: 1 E: 0 F: 2</td>
<td>5 (7.35)</td>
</tr>
<tr>
<td>S has used the chosen word before.</td>
<td>A: 0 B: 0 C: 2 D: 1 E: 0 F: 1</td>
<td>4 (5.89)</td>
</tr>
<tr>
<td>S’s background knowledge (e.g. S knows/remembers).</td>
<td>A: 1 B: 1 C: 1 D: 0 E: 0 F: 0</td>
<td>3 (4.41)</td>
</tr>
<tr>
<td>S has already planned to use the chosen word since the beginning.</td>
<td>A: 0 B: 0 C: 1 D: 1 E: 0 F: 0</td>
<td>2 (2.94)</td>
</tr>
<tr>
<td>S is familiar with the chosen word.</td>
<td>A: 0 B: 1 C: 1 D: 0 E: 0 F: 0</td>
<td>2 (2.94)</td>
</tr>
<tr>
<td>The chosen word matches the part of speech S wants.</td>
<td>A: 0 B: 0 C: 0 D: 1 E: 0 F: 1</td>
<td>1 (1.47)</td>
</tr>
<tr>
<td>It is the first word given in the e-dict.</td>
<td>A: 0 B: 0 C: 0 D: 0 E: 1 F: 1</td>
<td>1 (1.47)</td>
</tr>
<tr>
<td>The e-dict shows how to use the chosen word.</td>
<td>A: 0 B: 0 C: 0 D: 1 E: 0 F: 1</td>
<td>1 (1.47)</td>
</tr>
<tr>
<td>Total</td>
<td>A: 4 B: 3 C: 5 D: 1 E: 1 F: 2</td>
<td>68 (100.00)</td>
</tr>
</tbody>
</table>

Note: S = subject

To make the important categories clear, there are examples from the interviews to support the subjects’ reasons for their word choices from the electronic dictionary.
Reason 1: The subjects had seen the word before.

**quality**

*Researcher:* Why do you think so? Have you ever seen the word ‘quality’ before?

*Subject D:* I have seen like… the quality of this electrical appliance is good. Something like that.

*Researcher:* I see.

*Subject D:* I think it can be used.

Reason 2: The subjects used their intuition.

**restricted**

*Researcher:* Does the word ‘restricted’ that you have chosen convey the meaning ‘Kamnot’ you really wanted?

*Subject E:* Maybe.

*Researcher:* Maybe. What makes you feel confident?

*Subject E:* After I consider the other words, I think that they are not suitable for this context.

Reason 3: The chosen word matched the Thai meaning the subjects wanted to convey.

**graphic**

*Researcher:* Is the English word of the word ‘Krafik’ you saw in [the e-dict] what you wanted to convey to readers?

*Subject C:* Yes, it is. I want to tell readers that I like this game because there are beautiful graphics. They look realistic.

Reason 4: The subjects checked the meanings of the chosen words in their dictionaries.

**upright**

*Researcher:* Did you check the meaning from English to Thai again? What function in [the e-dict] did you use to check English meanings to Thai meanings?

*Subject C:* Cross Search function.

*Researcher:* What is it for?

*Subject C:* This function is available in [the e-dict]. It is used to check meanings in order to make sure that English words found are what I really want to convey in Thai.

*Researcher:* After you checked the word ‘upright’, is it the meaning of the word ‘Suetrong’ in Thai that you want to convey?

*Subject C:* Yes, it is.

*Researcher:* When you used this function, does [the e-dict] show that the meaning of the word ‘upright’ is the word ‘Suetrong’ in Thai?

*Subject C:* Yes, it does.

4.4 Words not chosen from the Thai-English electronic dictionary

Words sought and the subjects’ reasons for words not chosen are shown in Table 5. From the table, we can see that there are three reasons why the subjects did not choose words from the dictionary: words found were not the meaning the subjects wanted to convey;
words found were not suitable in the context; and words found were not the part of speech the subjects wanted to use.

**Table 5: Reasons for not choosing words from the electronic dictionary**

<table>
<thead>
<tr>
<th>Reasons</th>
<th>Words sought</th>
<th>Subjects’ explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not convey the intended meanings</td>
<td>ปราบ /Prap/</td>
<td>Words found (to clear, to level) were not the meaning he wanted to convey.</td>
</tr>
<tr>
<td></td>
<td>พัฒนา /Phatthana/</td>
<td>The subject felt more confident to use a word (upgrade) he had in mind. He also thought that the word could convey the intended meaning rather than the word found (develop).</td>
</tr>
<tr>
<td></td>
<td>จัดการ /Chatkan/</td>
<td>The meanings of words found (to handle, to deal with) were not the meaning he wanted.</td>
</tr>
<tr>
<td></td>
<td>สายไฟที่มีกระแสไฟ /Saifai Thi Mi Krasae /</td>
<td>A word found (a live wire/a dead wire) was not the meaning the subject wanted.</td>
</tr>
<tr>
<td></td>
<td>สายไฟที่ไม่มีกระแสไฟ /Saifai Thi Mai Mi Krasae /</td>
<td></td>
</tr>
<tr>
<td>Not convey the meanings</td>
<td>พบ /Phop/</td>
<td>Words found (to meet, to find, to discover) were not close to the meaning the subject wanted to convey.</td>
</tr>
<tr>
<td></td>
<td>พบว่า /Phop Wa/</td>
<td></td>
</tr>
<tr>
<td>Not suitable for the context</td>
<td>การเคลื่อนไหว /Kan Khlueanwai/</td>
<td>The subject thought that a word found (movement) was not suitable for the context.</td>
</tr>
<tr>
<td>Not the right part of speech</td>
<td>สีสัน /Sisan/</td>
<td>A word found (a colour) was not the part of speech the subject wanted to use. The subject wanted an adjective instead of a noun the dictionary provided.</td>
</tr>
</tbody>
</table>

5. Discussion and implications

The discussion and the implications will be discussed and presented under the following headings:

5.1 Correctness of vocabulary chosen in terms of form, meaning and use

Among the three aspects investigated (form, meaning and use), in answer to Research Question 1, form is the aspect where the subjects had no problems. Therefore, in this narrow sense, it can be said that students such as the ones in this study can use electronic dictionaries reliably.

As for meaning, the subjects were able to choose words to convey intended meanings correctly. According to the result, there were only two mistakes found out of 28 words chosen. The mistakes occurred because the subjects translated literally. When they found an English word in the dictionary and the English word had a Thai meaning that was what they thought in their mind, they would choose that word immediately. They were probably not aware whether the word chosen could be used appropriately in that context or not.

Use is the aspect that the subjects need to improve most. It consists of two sub-aspects: grammatical functions and collocations. We found that the problem where the subjects made mistakes frequently was grammatical functions (gerunds, plural nouns, modal verbs, parallel structures, etc.). From the interviews, the subjects reported that they made
mistakes because they were careless, they forgot or they did not check their written work carefully after they finished it.

5.2 Reasons for subjects’ word choices

In answer to Research Question 2, the reason for word choices that most of subjects reported was because they had seen them before, which shows that the subjects trusted their own knowledge rather than the dictionary. It was found that there were three situations. The first situation was that the dictionary provided only one word, but actually there should be more words to choose. In this situation, the subjects did not choose the word provided and preferred to use a word they had in their minds. The second situation was that the dictionary provided only one word because there is one English translated equivalent. Again, the subjects did not choose the word provided if they felt more confident to use a word from their own lexicons. The last situation was that, although the dictionary provided many words, the subjects did not choose any of them. They used their own words instead because they had seen the words before.

Next, the subjects chose words because they used their intuition. They felt a chosen word is better than other words in the context. The point is that, for words chosen for this reason, the subjects were not sure because they rarely saw and could not remember them. That is probably why they used their own feeling to choose words with which they felt more familiar, or perhaps, albeit without providing clear evidence, they just thought that those words should be used.

Furthermore, the subjects chose words because the chosen words matched the Thai meanings they wanted to convey; for this, they focused on meaning. The strategy that they used to help themselves for this is to check the meanings of words found by referring back to the English-Thai dictionary, which worked very well.

5.3 Problems in using electronic dictionaries

From the findings, the subjects’ mistakes in terms of form, meaning and use in their written work might be because of the electronic dictionary itself. Five problems are discussed as follows.

a. For some word searches, the dictionary did not provide any English words. There are some Thai headwords where the dictionary did not provide translated equivalents such as ‘เหมือนจริง’ /Muean Ching/ , ‘สมจริง’ /Somching/ , or ‘เสมือนจริง’ /Samuean Ching/ (realistic). Eventually, the subjects used the words ‘reality’ or ‘realistic’ from their own lexicons to use in their written work.

b. The dictionary provided insufficient words found. For example, Subject A wanted to find the English words from the Thai headword ‘การเคลื่อนไหว’ /Kan Khlueanwai/ (motion). However, when his dictionary provided the words found in the entry, it came out that the word ‘motion’ that Subject A had expected did not appear. The dictionary provided the words ‘movement, activity’. As a result, Subject A gave up searching for this word and, instead, used the word ‘motion’ from his own lexicon.

c. The dictionary provided only one equivalent meaning. It is found that there are nine words for which the dictionary provided only one meaning. For example, the Thai headword ‘ฐานะ’ /Thana/ (position) is given the meaning in English as ‘a position’ but it can have another meaning, ‘status’. This case supports Summers (1988) and Thompson (1987), who state that giving single-word translations
makes learners think, often erroneously, that there is a one-to-one relationship between L1 and L2; as a result, learners tend to make mistakes.

d. The dictionary provided more than one word but failed to differentiate them. The subjects might not be able to know how each particular word can be used in a particular context since they might believe that all of the provided words have the same meaning. For this, Underhill (1985) suggests that, when the dictionary provides more than one meaning equivalent, learners need to know the difference; otherwise, they may make mistakes because the chosen word is not suitable to the particular context.

e. The dictionary did not provide part(s) of speech of the words found. For example, the meaning of the Thai headword ‘กราฟิก’ /Krafik/ is given meaning as ‘graphic’; or another Thai word ‘คุณภาพ’ /Khunnaphap/ is given meaning as ‘quality’. From the two examples above, there should be ‘(n.)’ following the words to indicate a part of speech.

From the problems presented in this section, it might be possible to conclude that, as long as electronic dictionaries provide insufficient information, they may not be the best choice for students to use as their main support, especially when they have to use them for writing where accuracy is sometimes vital.

5.4 Suggestions for compilers
From this study, the problems that dictionary users found will be useful for the compilers to improve and develop their products in the future. They should take these five problems into consideration in order to improve their dictionary.

a. The subjects could not find many target words. This implies that the electronic dictionary may not have been helpful and user-friendly because they could not provide many words to users. This shows that the compilers have to increase their vocabulary of words translated as much as they can. At least, they should include possible words or similar meanings of each target word based on a reliable corpus.

b. Since the dictionary did not provide words the subjects expected to see, the compilers should add more words to serve their needs. By doing this, users will have various words to choose.

c. To avoid reinforcing the belief in a one-to-one relation at word level, the compilers should provide full semantic, grammatical and stylistic information, and usage notes that are not available in traditional bilingual dictionaries (Thompson, 1987).

d. Compilers should provide examples to show how a particular word is used in a particular context since this might prevent students from using incorrect words.

e. Compilers should provide part of speech of each word since students might not have met some words before. This solution will help them to use words matching a part of speech they need to use.

5.5 Training students to use electronic dictionaries
The following points are suggestions for teachers to train their students to use electronic dictionaries.
a. Ask students to study the electronic dictionary manual carefully, since there are many functions that the electronic dictionary provides to assist learners.

b. Make sure that students know how to use important functions in the electronic dictionary such as ‘Cross Search function’.

c. Make use of computer skills such as typing and knowing alphabets. With these skills, students can use electronic dictionaries faster.

d. Train students to change part of speech of a word sought if the target word cannot be found; for instance, if a word sought is a verb, change it into a noun form or vice versa. When converting the word into another part of speech, the electronic dictionary might supply a meaning.

e. For connectors, inform students to use a formal word to start as a headword. For example, when Subject A wanted to link two sentences by using the word โดยที่ /Doithi/, he could not find any English words. In this case, teachers might tell students to look up this word by using the first syllable โดย /Doi/ and scroll down to see other words which begin with the word โดย /Doi/. This strategy can be helpful to look up similar Thai words the subjects wanted to convey.

f. Raise students’ awareness to make use of the Thai lexicon since the Thai language is very rich in terms of the variety of words which have more or less the same meaning. In short, teachers may have to guide them to use lots of synonyms when they cannot find the target word.

g. Teach students to break down one word into smaller units. This strategy can help students to convey meanings they want by themselves. For example, students can break the word  semblant  /Samuean Ching/ into two units:  semblant  /Samuean/ and  มี /Ching/. Then they look up English translations of each word in the dictionary: ‘like, as if’ for  semblant  /Samuean/ and ‘real, true’ for  มี /Ching/.

h. Suggest students to simplify a sentence they have problem with. To do this way, they can still convey the meanings they want.

i. Train students to keep track of words found (i.e., writing words found onto a piece of paper whenever they are searching for words). When the students cannot find the word that they wanted, they can backtrack to consider from the list of words found which word they can use in a particular context.

j. Suggest students to use a monolingual dictionary. By doing this, students will have an opportunity to see useful information contained in the monolingual dictionary. It will help them to select a particular word to write in a particular context correctly.

6. Conclusion
The results of this study on students’ utilization of electronic dictionaries show that the aspect where students are weakest is use, and this appeared to be for two main reasons: students’ own problems and problems inherent in the electronic dictionary used in this research. There were three reasons for students’ word choices. Apart from the reason that the dictionary sometimes did not provide any English words, they are: students had seen the chosen words before, they used their intuition to choose words, and the chosen words matched the Thai meanings they wanted to convey. Suggestions and implications are also discussed as a result of these problems. Finally, the researchers hope that this study provides valuable findings and some directions for teachers of English to implement when training students to use electronic dictionaries successfully.
References:

(1) Boonmoh, A. (2003) Problems with Using Electronic Dictionaries to Translate Thai Written Essays into English. Master’s Thesis, Master of Arts, English Language Teaching, Faculty of Liberal Arts, King Mongkut’s University of Technology Thonburi.


(6) George, H. V. (1983) Classification, Communication, Teaching and Learning, Unpublished manuscript, English Language Institute, Victoria University of Wellington.


(15) ดวงตา ลักคุณะประสิทธิ์ (2542) “ความสามารถในการเขียนแสดงความคิดเห็นของนิสิตครุศาสตร์ที่เรียนคณิตศาสตร์เป็นวิชาเอก การศึกษาข้อผิดพลาดในการเขียนเรียงความภาษาอังกฤษ: กรณีศึกษา”, วารสารภาษาปริทัศน์, ปีที่ 18, ฉบับที่ 2542-2543, หน้า 60-74.

(16) พิชญ์สินี ขาวอุไร (2545) “การศึกษาข้อผิดพลาดในการเขียนเรียงความภาษาอังกฤษ: การศึกษาข้อผิดพลาดของการเขียนเรียงความภาษาอังกฤษ ของนักศึกษา”, วารสารภาษาปริทัศน์, ฉบับที่ 20, ปี 2545-2546, หน้า 137-156.

(17) สุภาณี ชินวงศ์ (2542) “ความสามารถในการเขียนเรียงความภาษาอังกฤษ ของนักศึกษา”, วารสารภาษาปริทัศน์, ปีที่ 18, ฉบับที่ 2542-2543, หน้า 10-27.
THE USE OF DICTIONARIES OF LOW PROFICIENCY STUDENTS IN A NATURALISTIC SETTING

Atipat Boonmoh,
Wareesiri Singhasiri
King Mongkut’s University of Technology Thonburi, THAILAND
atipat_b@yahoo.com, wareesiri.sin@kmutt.ac.th

Dictionaries are regarded as an important tool used for finding out the meanings of unknown words or checking the usage of interesting words. At KMUTT, in a Fundamental English Course (LNG 102), students are required to do self-study in reading skill by filling in a task record form to reflect their learning process and report new words that they have learned from reading materials. From the task, the researchers observed that the chosen words were too easy or too difficult and not suitable for their level of study. Some students preferred to learn only one type of words, i.e. nouns or verbs, by ignoring other word types. This study, therefore, investigates how students decided to choose words they were interested in learning.

Forty undergraduates majoring in science and engineering participated as the subjects by filling in the task record form which included the names of the dictionary they were using, look-up words, meanings selected and their parts of speech. The look-up words were analysed to find out the frequencies of words, parts of speech, meanings, and senses by checking with an advanced learner dictionary. The findings show that the subjects mainly selected nouns, verbs, and adjectives. They tended to choose low frequency words and were not aware of inflected forms of the searched words. For their convenience, subjects used bilingual dictionaries through mobile phones, although other types of dictionaries were provided. The results may help us understand students’ behaviors in using dictionaries and provide interesting implications for teachers and dictionary compilers.

Key Words: dictionaries; reading; words; low proficiency students
A STUDY OF DICTIONARY USE AND DICTIONARY NEEDS OF ESP STUDENTS

Trinidad Fernández
Technical University of Madrid, SPAIN
trinidad.fernandez@upm.es

Abstract: This article describes a survey of the dictionary use of ESP students in the Building Engineering School of the Technical University of Madrid, (UPM). It includes the frequency of students’ use of dictionary information, the situations in which they use it, the type of dictionaries they use, and how useful such information is perceived to be. The aim is to find out which dictionaries are more helpful for the students’ ESP learning. The instruments for data collection were a questionnaire and the ‘Quick Placement Test’ of OUP. The study indicates that students in general make very limited use of dictionaries, and when they do use them they prefer bilingual dictionaries as they seem to satisfy their inquiry promptly. However, when presented with sample entries from three contrasting types of specialized dictionaries (monolingual, bilingual and bilingualized), the majority of students chose the one in the bilingualized dictionary as the best and most complete of the three entries. Only the more proficient students make fuller use of bilingualized or monolingual dictionaries. Implications of these and other findings are discussed, and suggestions made to encourage students to use the most appropriate dictionaries for language learning.

Key Words: dictionary use, learners’ dictionaries, users needs, LSP dictionary, pedagogical specialized lexicography, specialized vocabulary, terminography, building engineering

1. OBJECTIVES
This article aims to analyse the use of dictionaries by tertiary students of Building Engineering in the UPM, and the needs those students have for dictionaries. It seeks answers to the following questions:
- How frequently do students use dictionaries and to what extent do they find them useful?
- What types of information in the dictionary do the students make use of most often, and how useful do they perceive them to be?
- Are there any differences in dictionary use between students more and less proficient in English?

The main objectives of this research study were, on one hand, to familiarize students with dictionaries and their macro- and microstructures, and to analyse students’ needs regarding dictionary reference skills. The results taken from the questionnaire will help us to improve explanations of dictionary content and dictionary use. Another objective was to increase students’ awareness of the scope of an ideal dictionary as well as its limitations, through an analysis of entries in different specialized dictionaries: a bilingual, a monolingual and a bilingualized. We asked them to compare the effectiveness of these dictionaries and decide which one provided better information for their needs and why.
To find out if there were differences in dictionary use relating to students’ English language proficiency, they were tested with the ‘Quick Placement Test’ (QPT) of Cambridge ESOL Examinations. Table 1 shows a comparison of several scale maps, i.e. the Association of Language Testers in Europe (ALTE), the Common European Framework Levels and the Cambridge ESOL Examinations (1).

Table 1: Chart of English level equivalence

<table>
<thead>
<tr>
<th>User</th>
<th>ALTE Level</th>
<th>Cambridge ESOL Examinations</th>
<th>Common European Framework Level</th>
<th>Common European Framework Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proficient</td>
<td>5</td>
<td>Cambridge Proficiency (CPE)</td>
<td>C2</td>
<td>Mastery (Upper Advanced)</td>
</tr>
<tr>
<td>Proficient</td>
<td>4</td>
<td>Cambridge Advanced (CAE)</td>
<td>C1</td>
<td>Effective Proficiency (Lower Advanced)</td>
</tr>
<tr>
<td>Independent</td>
<td>3</td>
<td>First Certificate (FCE)</td>
<td>B2</td>
<td>Vantage (Upper Intermediate)</td>
</tr>
<tr>
<td>Independent</td>
<td>2</td>
<td>Preliminary English Test (PET)</td>
<td>B1</td>
<td>Threshold (Lower Intermediate)</td>
</tr>
<tr>
<td>Basic</td>
<td>1</td>
<td>Key English Test (KEY)</td>
<td>A2</td>
<td>Waystage (Elementary)</td>
</tr>
<tr>
<td>Basic</td>
<td>0.5</td>
<td>-</td>
<td>A1</td>
<td>Break through (Beginner)</td>
</tr>
</tbody>
</table>

2. DICTIONARY USE AND DICTIONARY NEEDS

Reading comprehension is an important focus of the syllabus for our students, and vocabulary is a good predictor of the reading comprehension level in a foreign language (Laufer, 1995). Reading in L2 depends on the meaning of words and the knowledge of the subject matter, and unknown words can be guessed by means of clues taken from the context. However when students need more accurate information for productive purposes, the dictionary may be the only help available.

It has been shown that second language learners hardly ever use dictionaries (Atkins, 1996; Marello, 1998) and when they use them, most of them (75%) prefer bilingual to monolingual dictionaries (Atkins & Knowles, 1990; Hartmann, 1983; and Tomaszczyk, 1983) because those dictionaries offer them assurance and foster self-reliance due to the prompt satisfaction of their need to understand. This finding can also be applied to ESP learners.

The literature published on dictionary use rarely refers to studies performed in English-speaking countries, where English is taught as a second language. A study on dictionary use by international undergraduate students from all over the world at a British University showed that the preferred dictionary for those students was not the bilingual but the monolingual dictionary. Only 9% of those students use bilingual dictionaries and 4% specialized dictionaries. Nevertheless in a specific assignment task to carry out by those students, only a few consulted a bilingual dictionary (Nesi and Haill, 2002). This might be due to the students’ higher level of proficiency of L2, and to the fact that they were in an English-medium university, i.e. in a fairly naturalistic setting for learning the language. Students become increasingly familiar with their dictionary and perform better as their language level increases (Lew, 2004). Learners with advanced L2 skills tend to use monolingual dictionaries (Atkins and Varantola, 1997). By contrast, bilingual dictionaries
remain the type most often used by beginners and intermediate students while the bilingualized dictionaries are used by average and good students. Although the students at the British University were skilled dictionary users, they occasionally experienced some kind of look-up failure e.g. choosing a wrong definition because they had misidentified the grammatical class of the look-up word (Nesi and Haill, 2002). It seems that grammar is not given the importance it deserves and this is the cause of many failures with my students, only 29.09% of them make use of grammatical information in their look-ups.

To use a monolingual dictionary successfully learners need to have an effective receptive vocabulary of 2000 words and need to be able to interpret definitions (Nation, 2003). Some students who start using monolingual dictionaries have not been trained in dictionary use and might not succeed in this task, which would take them back to their previous habit of using bilingual dictionaries, because they feel those dictionaries solve their problems better. Moreover, if the dictionary is not properly used a student’s confidence may be undermined on the value of the dictionary (Slagter, 2000). Paradoxically, students prefer bilingual dictionaries despite the fact that most teachers would like them to use monolingual dictionaries, to make them begin using L2 in a productive way.

It is understandable that our learners and L2 learners in general consult the dictionary most frequently to find out the meaning of new words. It is natural that they try to find this meaning by means of L1 equivalents although there is the danger of over-reliance on L1 equivalents of words that may lead to the neglect of other kinds of knowledge (Fan, 2000).

3. METHOD
The needs of dictionary users in different countries are different (Hartmann and James, 1998). Users’ purposes can also be different with regard to the different countries and also to different types of users. A context-specific questionnaire to retrieve information from the students is most necessary. This study was based on the data collected from a questionnaire which consisted of 23 questions on dictionary strategies, which was handed out in class. There were questions about the students’ profile, their typical use of dictionaries, skills more frequently practised in their English lessons, types of dictionary used, etc. We also wanted to know if they had received any training in dictionary use and how dictionaries helped them better in their L2 learning.

3.1 Subjects’ profile, L2 level, and dictionary use
The subjects under study included 110 second-year and third-year degree students of Building Engineering at UPM. The students’ (95%) average age is 19-24. The results from the placement test show that, proficiency L2 levels of the subjects are: elementary: 16; lower intermediate 44; upper intermediate 39; lower advanced 11 and only 1 subject presented upper advanced level. There are 25% subjects who have been studying English from 5 to 10 years and 52% from 11 to 15. Nearly 21% of the subjects have studied English for more than 15 years.

Regarding dictionary use, most of them (85%) started using dictionaries at primary level, 52% due to teachers’ advice, and 46% as a personal initiative. Their most frequent use of the dictionary was at secondary level when they received more English lessons. Grammar practice takes most of their time (90%) of those courses before university, followed by vocabulary exercises (74%). The subjects used the dictionary more frequently when translating and writing, two productive skills they can practise by themselves and take their time over look-ups. It is remarkable to find that 7% of the subjects use the dictionary
when listening; one wonders what the speaker would do in the meantime. As expected, the results showed that the most important reason for using a dictionary was to find the meaning of words (98%). Taylor (1988) assured that the most frequent use of dictionaries was to find out word meanings, and this is the same for our students. On the other hand, the least frequent look-up for our students was etymology.

Most people would think that at tertiary level students are familiar with dictionary use; however, students say that they do not always find the words they are looking up, because they might have not chosen the appropriate dictionary, or they do not know how to interpret the information dictionaries include. Nearly 80% of our students have not received dictionary training, something we could corroborate from some of their answers: 41% of the subjects would like to receive this training as shown in table 3, section 3.3.

Learners’ dictionaries, either monolingual or bilingualized, are seldom used because students do not know how to use them, or sometimes they do not even know those dictionaries exist (Calderón Campos, 1994). Most of our students own their bilingual (pocket or large) dictionaries, 70% use pocket bilingual and 58% large bilingual dictionaries; only 5% have bilingualized dictionaries, having followed their teachers’ advice to purchase them. The second most important look-up is examples of use and synonyms (54%). Hardly anyone looks at the appendices (9%). Most of their queries are to translate from Spanish into English (88%) and from English into Spanish (72%). It is positively surprising that in reading practice, besides looking up words to know their meaning, they also read examples of usage in the entry (68%), a sign of how important usage examples are for them. Technical words are the most frequently searched 88%. Only 13% of our students use pocket monolingual learners’ dictionaries, 24% use large monolingual learners’ dictionaries but only 5% use bilingualized dictionaries. The students who use monolingual and bilingualized dictionaries have B2 and C1 levels of English.

3.2 Subjects’ attitudes towards dictionaries
When students are given open questions, their replies are usually more difficult to analyse, however they can provide richer results. The following information is about their attitudes towards dictionaries, table 2 below.

Table 2: Subjects’ attitudes

<table>
<thead>
<tr>
<th></th>
<th>strongly agree</th>
<th>agree</th>
<th>disagree</th>
<th>strongly disagree</th>
<th>don’t know</th>
</tr>
</thead>
<tbody>
<tr>
<td>all dict. contain same info</td>
<td>1.82%</td>
<td>10.91%</td>
<td>53.91%</td>
<td>29.09%</td>
<td>0.00%</td>
</tr>
<tr>
<td>dict. use is boring</td>
<td>0.91%</td>
<td>30.00%</td>
<td>47.27%</td>
<td>18.18%</td>
<td>1.82%</td>
</tr>
<tr>
<td>to depend on one dict is not enough</td>
<td>20.00%</td>
<td>53.64%</td>
<td>16.36%</td>
<td>3.64%</td>
<td>2.73%</td>
</tr>
<tr>
<td>understand Arch. texts without using dict</td>
<td>1.82%</td>
<td>40.40%</td>
<td>34.55%</td>
<td>20.00%</td>
<td>0.91%</td>
</tr>
<tr>
<td>want dict. training</td>
<td>7.27%</td>
<td>33.64%</td>
<td>30.00%</td>
<td>18.18%</td>
<td>6.36%</td>
</tr>
</tbody>
</table>

It was positive to know that most students are aware that dictionaries can have variable information (83%). They think to depend on one dictionary is not enough (74%), which
might be interpreted as the ‘so much appreciated bilingual dictionary’ (Atkins & Knowles, 1990; Hartmann, 1983; and Tomaszczyk, 1983) is not working properly. Almost half of the subjects (42%) say that they usually understand Architecture texts without using dictionaries; however 56% think they need a dictionary for those specialized terms, so clearly this is something which requires more study. Only 41% think it important to have dictionary training, though my viewpoint is that without it they do not benefit from the information the different dictionaries can provide.

### 3.3 Essential information students require in a dictionary

This section compares information from tables 3 and 4. Table 3 shows the information students expect to find in a dictionary, ranked according subjects’ selections. Table 4 illustrates what our students regard as the most important information in a dictionary entry for ESP learners. The information on entries was provided in alphabetical order so as not determine students’ selections. The ranking appears on the right column, i.e. no. 1: examples, no. 2: L2 definition, no. 3: grammatical information, etc. What students expect to find and/or most times do find in dictionaries is compared to the essential information that a dictionary entry should provide:

a) Sixty seven subjects think that most dictionaries provide ‘phonetic information’ (pronunciation in e-dictionaries), which is number 4 (67%) in the ranking of most important information a good dictionary should have to help them learn L2. However, most specialized dictionaries do not provide phonetic information though I have always favored it. Knowledge of pronunciation is essential to the production of speech; it is a fact that most students have difficulty with phonetic scripts (Taylor, 1988). Campoy Cubillo (2002) thinks that because pronunciation appears after the entry in most dictionaries, students assumed that it is important.

b) Fifty three subjects chose ‘examples of use’, which is number one in table 4 chosen by 86% subjects, most of them with B1 and B2 levels of English.

c) Fifty one students chose ‘translation equivalent’ as something expected in a dictionary, though only 48% of them chose L1 equivalent, which is number 8 in the ranking. However, in Campoy Cubillo’s study (2002), L1 equivalents are the most important feature in dictionaries.

d) ‘Synonyms’ were another favorite selection, chosen by 49 subjects. In their ideal dictionary 61% thought it was also important.

e) ‘Definitions’ were chosen by 46 subjects as a feature dictionaries should have. However, definitions in L2 were no. 2 in the ranking, 77% subjects. Definitions in L1 were chosen by 41% of the subjects (no.10 in the ranking). This may reflect the importance students give to the practice of L2. However, it seemed that once the students had the L1 equivalent in L1 they did not think L2 was important, only 36% (Campoy Cubillo, 2002).

f) ‘Meaning’ was chosen by 28 students without specifying the way meaning should be provided.

g) Nearly half of the students (45%) thought ‘illustrations’ were important in dictionaries, which might be due to their domain of interest.

h) More than half per cent (55%) chose ‘context’ as an important feature.

i) **Gender, intonation, numerous.** The first characteristic is important in Spanish but not so much in English. Regarding ‘intonation’. I wonder how this can be provided in a print dictionary. The word ‘numerous’ used by students might refer to singulants and plurals or to count and non-count uses of nouns.
Table 3: Information students expect on dictionaries

<table>
<thead>
<tr>
<th>phon/ pronu</th>
<th>examples of use</th>
<th>translation</th>
<th>synonym</th>
<th>definitions</th>
<th>meaning</th>
<th>find the look-ups</th>
<th>antonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>subjects</td>
<td>67</td>
<td>53 s</td>
<td>51 s</td>
<td>49 s</td>
<td>46 s</td>
<td>28 s</td>
<td>18 s</td>
</tr>
<tr>
<td>illustrati</td>
<td>idioms</td>
<td>context</td>
<td>several meaning</td>
<td>words related to same topic</td>
<td>gender</td>
<td>spelling</td>
<td></td>
</tr>
<tr>
<td>ons</td>
<td></td>
<td></td>
<td>(polysem?)</td>
<td>(syn?)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>etymol</td>
<td>collocation</td>
<td>concepts</td>
<td>applications</td>
<td>suffixes</td>
<td>type of word</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(examples of use)?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>domain</td>
<td>homonyms</td>
<td>intonation</td>
<td>numerous</td>
<td>abbreviat.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Ranking of essential information in dictionary entries

<table>
<thead>
<tr>
<th>LEVEL</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>A2</td>
</tr>
<tr>
<td>Nº</td>
<td>Nº</td>
</tr>
<tr>
<td>%</td>
<td></td>
</tr>
<tr>
<td>Check the item(s) you consider essential in an entry of a dictionary for L2 learners (they are in alphabetical order)</td>
<td>Ranking</td>
</tr>
<tr>
<td>drawings</td>
<td>0</td>
</tr>
<tr>
<td>collocations</td>
<td>0</td>
</tr>
<tr>
<td>encyclopedic information (extra info for the non-expert)</td>
<td>1</td>
</tr>
<tr>
<td>etymology of the headword</td>
<td>0</td>
</tr>
<tr>
<td>examples</td>
<td>1</td>
</tr>
<tr>
<td>field (arch./maths)</td>
<td>1</td>
</tr>
<tr>
<td>grammatical information (word class/behaviour)</td>
<td>1</td>
</tr>
<tr>
<td>L1 definition</td>
<td>0</td>
</tr>
<tr>
<td>L2 equivalent</td>
<td>0</td>
</tr>
<tr>
<td>L2 definition</td>
<td>1</td>
</tr>
<tr>
<td>phraseology (idioms, expressions)</td>
<td>0</td>
</tr>
<tr>
<td>pragmatics: ways in which people use language</td>
<td>0</td>
</tr>
<tr>
<td>pronunciation</td>
<td>1</td>
</tr>
<tr>
<td>synonyms</td>
<td>1</td>
</tr>
<tr>
<td>word senses (depends on context)</td>
<td>1</td>
</tr>
</tbody>
</table>

4. COMPARISON OF THE DICTIONARY ENTRY FOR ‘BAY WINDOW’ IN THREE DICTIONARIES

To know which dictionaries were more efficient for our students and found effective in supporting their understanding, we asked them to analyze 3 entries of the same term ‘bay window’ from three different specialized dictionaries (print monolingual, print bilingual and electronic bilingualized), since one of our objectives was to increase awareness of the scope of an ideal dictionary for our students as well as its limitations. The highest scores were obtained with the bilingualized dictionary as seen in table 6 below.
4.1 Monolingual specialized dictionary

From my viewpoint this is one of the most complete and rigorous monolingual print dictionaries of Architecture and Construction. It has different senses in the same entry, which is the common practice of lexicography not of terminology where there is an entry for each sense of the term. It also provides a lot of illustrations. This edition has been revised and extended including the latest terminology related to building conservation, new technologies, techniques, materials and new tendencies. As is the common tradition of most specialized dictionaries, it does not provide information on phonetics, grammar, examples of usage, because it is not designed as a learner’s dictionary.

However the results from the questionnaire are very different from the ones I expected. No one chose the definition provided in this dictionary as a good one, probably because they found it too technical. Our students clearly need dictionary training as some of them (nearly 13%) thought this monolingual dictionary provides a ‘fine equivalent’ (8%) and 4% not ‘a so good Spanish equivalent’. Some others (8%) thought it contained phonetics, etc.

Figure 1: Dictionary of Architecture and Construction (Harris, 2000)
Table 5: Students’ evaluations of the monolingual dictionary entry

<table>
<thead>
<tr>
<th>Technical content</th>
<th>Language content (gramm)</th>
<th>Example of usage</th>
<th>Spanish equivalent</th>
<th>Phonetics</th>
<th>Illustrations</th>
</tr>
</thead>
<tbody>
<tr>
<td>complete</td>
<td>good</td>
<td>good</td>
<td>good</td>
<td>good</td>
<td>clear</td>
</tr>
<tr>
<td>19.09</td>
<td>0.91</td>
<td>0.00</td>
<td>0.00</td>
<td>0.91</td>
<td>£31.82</td>
</tr>
<tr>
<td>accurate</td>
<td>fine</td>
<td>fine</td>
<td>fine</td>
<td>fine</td>
<td>not clear</td>
</tr>
<tr>
<td>37.27</td>
<td>31.82</td>
<td>26.36</td>
<td>8.18</td>
<td>3.64</td>
<td>62.73</td>
</tr>
<tr>
<td>not complete</td>
<td>not so good</td>
<td>not so good</td>
<td>not so good</td>
<td>not so good</td>
<td>other</td>
</tr>
<tr>
<td>34.55</td>
<td>35.45</td>
<td>20.91</td>
<td>4.55</td>
<td>5.45</td>
<td>2.73</td>
</tr>
<tr>
<td>not accurate</td>
<td>not provided</td>
<td>not provid.</td>
<td>not provid.</td>
<td>not provid.</td>
<td>not provided</td>
</tr>
<tr>
<td>9.09</td>
<td>30.91</td>
<td>50.91</td>
<td>83.64</td>
<td>88.18</td>
<td>1.73</td>
</tr>
</tbody>
</table>

4.2 Specialized bilingualized dictionary

**WINDOW OPENINGS** opening characteristics <by Form>

**bay window**

A window or series of windows, rising from the ground, forming a bay or recess in a room and projecting outward from the main wall of a building

**Synonyms:** jut window

**Broader Term (G):** window opening

**Narrower Term (G):** bow window, cant bay window, compass window, oriel

**Notes:**

i. A *bay window* which occurs only on upper storeys is generally called *oriel window* (Chudley, 1989).

ii. Sometimes it is corruptly called *bow window* (a curved bay window) (BSI). However, bay windows can have different shapes not only curved.

**Context from the WebCorp**

http://www.twoshabbychicks.com/projectofthemonth_2 ...

…… I told him that I wanted some sort of shelf or valance to go over my bay window, so that I could have an area for display…..

Figure 2: Bilingualized dictionary of Architecture and Construction (Fernández, 2004)
This entry corresponds to the bilingualized specialized learners’ electronic dictionary, which presents systematic, as well as alphabetical entry to its macrostructure, and thus is very flexible for decoding and encoding language. It provides definitions and examples of usage in L2, which show the term in context retrieved from the WebCorp with links to the captured information. Laufer (1995) compared the effectiveness of having only a definition, or an example as well in a study of comprehension and production. The results showed that the combination of definition and examples gave the best general outcome in learning new vocabulary. However there were differences in that for subjects’ comprehension of new words the definition factor is more effective, and then the examples. But for the purpose of production, a combined entry (definition + example) is more useful. This dictionary (Fernández, 2004) provides also translation equivalents in Spanish where the matching is between concepts not between terms, and information about partial or total equivalence (Fernández, Flórez de la Colina, and Peters 2009). It offers multimedia information: visual, acoustic and textual. As we have seen, many students confirmed in the questionnaire that phonetics is important in specialized dictionaries for learners. Sometimes there are specialized terms which come from the same root with similar spellings in English and Spanish but different pronunciation, i.e. architecture (arquitectura), aerogel (aerogel) or visible transmittance (transmisión de la visibilidad).

Table 6: Students’ evaluations of the bilingualized dictionary entry

<table>
<thead>
<tr>
<th>Technical content</th>
<th>Language content (gramm)</th>
<th>Example of usage</th>
<th>Spanish equivalent</th>
<th>Phonetics</th>
<th>Illustrations</th>
</tr>
</thead>
<tbody>
<tr>
<td>complete 74.55</td>
<td>good 29.09</td>
<td>good 13.64</td>
<td>good 41.82</td>
<td>good 11.82</td>
<td>clear 84.55</td>
</tr>
<tr>
<td>accurate 23.64</td>
<td>fine 40.00</td>
<td>fine 56.36</td>
<td>fine 42.73</td>
<td>fine 18.18</td>
<td>not 8.18</td>
</tr>
<tr>
<td>not complete 5.45</td>
<td>not so good 6.36</td>
<td>not so good 20.91</td>
<td>not so good 10.91</td>
<td>not so good 4.55</td>
<td>other 3.64</td>
</tr>
<tr>
<td>not accurate 0.00</td>
<td>not provided 20.91</td>
<td>not provided 7.27</td>
<td>not provided 2.73</td>
<td>not provided 62.73</td>
<td>not provided 2.73</td>
</tr>
</tbody>
</table>
4.3 Specialized bilingual dictionary

This is a bilingual dictionary of Construction and Public Works which offers an extended Spanish definition on each term with different senses. It provides the equivalents in English, it is not bidirectional in both languages, but monodirectional L1>L2. It has also some illustrations. There are not many bilingual specialized dictionaries in English and Spanish of Architecture and Construction, so this is the one I chose despite some drawbacks, because it is very comprehensive.

Table 7: Students’ evaluations of the bilingual dictionary entry

<table>
<thead>
<tr>
<th>Technical content</th>
<th>Language content (gramm)</th>
<th>Example of usage</th>
<th>Spanish equivalent</th>
<th>Phonetics</th>
<th>Illustrations</th>
</tr>
</thead>
<tbody>
<tr>
<td>complete</td>
<td>47.27</td>
<td>good 17.27</td>
<td>good 11.82</td>
<td>good 50.00</td>
<td>good 0.91</td>
</tr>
<tr>
<td>accurate</td>
<td>35.45</td>
<td>fine 38.18</td>
<td>Fine 26.36</td>
<td>fine 43.64</td>
<td>fine 2.73</td>
</tr>
<tr>
<td>not complete</td>
<td>13.64</td>
<td>not so good 15.45</td>
<td>not so good 11.82</td>
<td>not so good 5.45</td>
<td>not so good 2.73</td>
</tr>
<tr>
<td>not accurate</td>
<td>5.45</td>
<td>not provided 26.36</td>
<td>not provided 48.18</td>
<td>not provided 1.82</td>
<td>not provided 89.09</td>
</tr>
</tbody>
</table>

On the basis of these results it is concluded that the combination of the monolingual information which contains a definition and examples together with the translation equivalent of the word/term back into the students’ L1 produces the best results. These success rates are comparable to those in a similar large-scale study by Atkins and Varantola (1998). Though university students are urged to use L2 dictionaries as we have seen, most of them use the bilingual dictionary because they do not only want to have a general idea of what the word means and they prefer to confirm their guessing of the meaning with the translation equivalent into their mother tongue. A good compromise to this problem would be the bilingualized dictionary (Campoy Cubillo, 2000). The use of...
this type of dictionary seems to reconcile the two points of view, the teacher’s preference for students using L2 monolingual dictionaries, and the students’ preference for the bilingual dictionary (Koren, 1997). Moreover, there is a need for a dictionary which includes both general and semi-technical specialized English (Campoy Cubillo, 2000), as students do not find specialized terms in the dictionaries they look-up. From time to time they use specialized dictionaries, showing that a bilingualized specialized dictionary is very valuable for students’ purposes and needs.

In answer to the question comparing the three dictionaries: ‘circle the dictionary entry you like most and explain why’, the results from the students were:

Table 8: Students’ reactions to the information in different types of dictionary entries

<table>
<thead>
<tr>
<th>No answer 4.01%</th>
<th>sub</th>
<th>Monoling 0.00%</th>
<th>subj</th>
<th>Bilingualized 89.09%</th>
<th>subj</th>
<th>Bilingual 6.90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Very poor</td>
<td>53</td>
<td>Complete</td>
<td>3</td>
<td>Good to have definition in Spanish</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Synonyms</td>
<td>3</td>
<td>Better to have meaning in L1, because it is easier to understand</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Good Imagines</td>
<td>2</td>
<td>Very easy to understand</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Good examples of use</td>
<td>2</td>
<td>Most useful because it is in Spanish</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Good translation</td>
<td>2</td>
<td>Needs synonyms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Clearer than the others</td>
<td>2</td>
<td>Good for different uses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Good definitions with easy words</td>
<td>1</td>
<td>Very good</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Does not provide phonetics</td>
<td>1</td>
<td>Drawback: no illus.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Good to provide notes Notes not necessary</td>
<td>1</td>
<td>Needs illustrations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Context</td>
<td>1</td>
<td>Good examples</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Helps think in English because def and ex are in this language</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Conceptual relation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Easy use</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Word senses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Electronic dict. relates terms better</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Good phonetics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Audio (good, better than phonetics)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Does not provide examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>polyvalent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Some of the reasons for choosing dictionary B are: more complete, easy to use, varied in its contents, and multimodal, all of which result from its being an electronic dictionary where terms can be related. Bilingualized learner’s dictionaries and monolingual learners’ dictionaries should form part of a staged learning progression from the fully bilingual dictionary to the monolingual dictionary (Cowie, 1999).

Laufer and Melamed’s (1994) study also showed that the best results were almost always obtained when the bilingualized dictionary was used. This was true for all learners in the case of comprehension, and for the good and average dictionary users in the case of production. Only the unskilled users did better on production with a bilingual dictionary. The authors argued that the most useful learner’s dictionary was the bilingualized type. They concluded that the bilingualized dictionary is very effective because of its capability with all types of individual preferences.

5. SUMMARY AND CONCLUSIONS
In this dictionary survey, as in most others, the majority of students use bilingual dictionaries and they find them useful; students make only limited use of monolingual dictionaries and even more limited for bilingualized.

It is understandable that our learners consult the dictionary most frequently to find out the meaning of new words. It is natural that, in general, they try to find this meaning by means of L1 equivalent, though in our study the Spanish students find L2 definitions very useful. The more proficient students were found to use collocations, pronunciation, word senses significantly more often than the lower level ones.

Students’ preference for the bilingualized dictionary, as found in the results, seems to contradict the dearth use they make of them, because of the fact that many of our students do not even know those dictionaries exist.

Teachers need to work very closely together with learners to show them ways to use dictionaries more efficiently for specific purposes. It is not sufficient to tell our students that dictionaries are good aids in learning L2. We need to help them reap the benefits of good dictionaries, with explicit teaching on how to use them (Hartmann, 1992). All the findings of the present study suggest that it is essential to train ESP students on dictionary use and skills. In all such teaching, it is important to note that ‘A good dictionary must offer users not only what they are aware they do not know but also that they are unaware they want’ (Atkins, 1985: 23).

NOTE
(1) The Common European Framework levels of Reference for Languages are used in the reorganization of national curricula and by international consortia for the comparison of language certificates.
REFERENCES


Acknowledgements
My thanks to Professor Pam Peters (Macquarie University, Sydney-Australia) for her valued comments on this paper. I also want to thank Dr. Flórez de la Colina and Ms Alicia de San Antonio PhD student, both from the UPM, for passing all the questionnaires to excel program and handing me the results.
Abstract: Malaysia and the Philippines have relatively similar concepts on things and traditions. Their histories prior to the arrival of the colonial powers in Southeast Asia are quite similar: form of government, rituals, feasts, etc. The two countries, together with the other countries in Southeast Asia, have had a very close contact with each other through trade and migration. However, when the colonizers arrived in Southeast Asia, their similarities in cultural and religious concepts gradually changed. This paper looks into the religious and cultural concepts in the Philippine and Malaysian languages and the challenges it brings in building a bilingual dictionary. Some lexical items are easily translatable but the religious and cultural concepts pose some problems. For example, the word *binyag* ‘baptism, ritual for coming of age’ in Filipino does not correspond to any word in Malay although the concept is shared. *Kitab* ‘holy book pertaining to Islam’ in Malay does not correspond to any word in Filipino as well. These examples of lexical items generate difficulties in language teaching and learning. This paper will identify the ‘challenging’ areas in building a bilingual dictionary of Malay-Filipino. This paper will also analyze the religious and cultural concepts in terms of semantic theory and historical linguistics and how these challenges can be addressed.

Key Words: Bilingual dictionary; religion; culture; Bahasa Malaysia; Tagalog; Filipino; Catholic; Islam.

1. INTRODUCTION AND BACKGROUND

Tagalog, the Philippine’s basis for its national language Filipino, and Bahasa Melayu, the national language of Malaysia, Singapore and Brunei, have been largely influenced by the dominant religions in these countries. The Catholic religion, brought by the Spaniards in 1521, has largely been ingrained in all cultural aspects of the Filipinos. This is reflected mainly in the language, particularly relating to religion and celebration. Islam on the other hand, has been present in Southeast Asia earlier than Christianity. Islam was established in Malaysia by the conversion of the king of Malaka in 1400 but its spread was mostly in the coastal areas of Malaysia, Indonesia, and the Philippines (Tarling, 1992). Although Islam also came to the Philippines, its influence was not as vast as that in Malaysia, Brunei and Indonesia. Islam was mainly practiced in southern Philippines.

Prior to the arrival of both religions, Hinduism and Buddhism as well as paganism and animism were practiced for hundreds of years. Some words in Tagalog and Malay have their origins from Sanskrit. Inspite of this, Tagalog and Malay share a lot of cognate words making their genetic relationship strong.
This paper will discuss the influence of Catholicism and Islam in Filipino and Malay languages. The aim of this paper is to identify the terminologies used in terms of religious practices and how these terms are translated in both languages. The paper will also find ways in explaining these religious terminologies in order to build a bilingual dictionary using Malay and Filipino as both source and target languages.

1.1. Brief History of Dictionaries in the Philippines and Malaysia
The first Tagalog dictionaries published in the Philippines were written by Fray Francisco Blancas de San Jose in 1610 and Fray Pedro de San Buenaventura in 1613. Both were published by Tomas Pinpin in Pila, Laguna (Dirk, 1992). Most of the Tagalog dictionaries were authored by Spanish friars and were used as tool to learn the language of the natives. Only during the 1900s that dictionaries written by American scholars were published. Most of these dictionaries are bilingual dictionaries: Spanish-Tagalog, Tagalog-Spanish, English-Tagalog, Tagalog-English, Tagalog-Cebuano, etc. Tagalog monolingual dictionaries were only published in 1958, when Tagalog was proclaimed as the national language. But only a handful of Filipino or Pilipino dictionaries are available.

Pigafetetta may have probably produced the first Italian-Malay dictionary during Magellan’s expedition in the Pacific in 1500s. But it was only in the 1800s when the first Malay dictionary was published. Nor Hamsiah (2007) noted that between the years 1631 to 1899, there were about 57 Malay glossaries and dictionaries published. The oldest monolingual dictionary was Kitab Pengetahuan Bahasa published by Raja Ali Haji in 1858. From 1900s to the present, there are a lot of dictionaries in Malay, monolingual, bilingual, and even trilingual, are available.

To date, the only available glossary for Malay-Tagalog is Alicio’s *Language Phrases in English, Bahasa Melayu, Tagalog, and Spanish* (2001). It cannot be considered as dictionary but more of a phrasebook to guide learners of any of these languages in any of the target languages. Attempts are being made to compile and publish a probable first Filipino-Malay or Malay-Filipino dictionary.

1.2. Statement of the Problem
As mentioned in the previous section, there is no available Filipino-Malay, Malay-Filipino dictionary. This has become a constant problem in the study and teaching of Malay and Filipino in the Philippines and Malaysia. Bahasa Malaysia (or Bahasa Indonesia) is being taught as one of the language electives in some universities in the Philippines, while Filipino is being taught at the University of Malaya. Without a proper dictionary to use, the learning and teaching of both languages have become a burden to both teacher and student. Certain concepts in the Filipino culture are difficult to explain most particularly in the area of religion and festivities. The same is also true for Malay. There are some ceremonies and rituals practiced in Malaysia that are difficult to explain or to give the equivalent in Filipino (or even English). Since both languages are genetically similar, would it be possible to find certain cultural or religious concepts similar as well that will aid in the development of a Filipino-Malay bilingual dictionary?

2. MALAY AND FILIPINO DATA
Several words in Malay and Filipino were identified relating to religion and cultural concepts. These words were taken from the texts used in teaching Malay and Filipino and were found to be difficult to ‘translate’ in each of the languages. The English explanations helped a little, however, when related to some cultural practices, were understood easily. This section is divided into two parts: religious aspect and the cultural aspect.
2.1. Religious Terminologies

Below are some of the data used for this paper (based on most frequently used words):

Table 1. Malay Words Pertaining to Religion.

<table>
<thead>
<tr>
<th>Malay</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Azan</td>
<td>Call for prayers</td>
</tr>
<tr>
<td>Rumah Terbukah</td>
<td>Open house (during Hari Raya)</td>
</tr>
<tr>
<td>Fatwa</td>
<td>Religious decision in relation to any practices in the religion (Islam)</td>
</tr>
<tr>
<td>Firman</td>
<td>Warning or command</td>
</tr>
<tr>
<td>Haj</td>
<td>Pilgrimage to Mecca</td>
</tr>
<tr>
<td>Hari Raya</td>
<td>Feast after Ramadan, end of Ramadan</td>
</tr>
<tr>
<td>Jemaah</td>
<td>Pilgrims, congregation of believers in Islam</td>
</tr>
<tr>
<td>Kadi</td>
<td>Judge of Islamic Laws</td>
</tr>
<tr>
<td>Kafir</td>
<td>Infidel</td>
</tr>
<tr>
<td>Khalwat</td>
<td>Illicit (illicit contact related to religion)</td>
</tr>
<tr>
<td>Kiblat</td>
<td>Direction to Mecca for praying</td>
</tr>
<tr>
<td>Kitab</td>
<td>A holy book or scriptures</td>
</tr>
<tr>
<td>Maghrib</td>
<td>West (Arabic origin, pertaining to evening prayers)</td>
</tr>
<tr>
<td>Makam</td>
<td>Tomb for Muslims</td>
</tr>
<tr>
<td>Ramadan</td>
<td>Ramadan, the month for fasting</td>
</tr>
<tr>
<td>Songkok</td>
<td>A headgear worn by men during Friday prayers</td>
</tr>
<tr>
<td>Surau</td>
<td>Chapel (for Muslims)</td>
</tr>
</tbody>
</table>

Some of these are unfamiliar to a person who has little or no background of Islam. There are words however, that have been used quite extensively in the literature such as Ramadan, Hari Raya and Haj. But the rest of the words need further elaboration and explanation for a Malay language learner. In the case of Filipino, the following words are frequently used in the texts used for language teaching:

Table 2. Filipino Words Pertaining to Religion.

<table>
<thead>
<tr>
<th>Tagalog</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antanda</td>
<td>Sign of the cross (for Catholics)</td>
</tr>
<tr>
<td>Ati-Atihan</td>
<td>Feast in celebration of the Sto. Niño (Holy Child)</td>
</tr>
<tr>
<td>Binyag</td>
<td>Baptism</td>
</tr>
<tr>
<td>Komunyon</td>
<td>Communion</td>
</tr>
<tr>
<td>Kumpil</td>
<td>Confirmation</td>
</tr>
<tr>
<td>Kumpisal</td>
<td>Confession</td>
</tr>
<tr>
<td>Nobena</td>
<td>Novena</td>
</tr>
<tr>
<td>Palaspas</td>
<td>Palm used during Palm Sunday</td>
</tr>
<tr>
<td>Panata</td>
<td>Vow, a solemn promise</td>
</tr>
<tr>
<td>Piyesta</td>
<td>Feast, fiesta</td>
</tr>
<tr>
<td>Rosaryo</td>
<td>Rosary</td>
</tr>
<tr>
<td>Salubong</td>
<td>Celebration of Easter Sunday</td>
</tr>
<tr>
<td>Semana Santa</td>
<td>Holy Week</td>
</tr>
<tr>
<td>Sinakulo</td>
<td>Re-enactment of Jesus Christ’s suffering</td>
</tr>
</tbody>
</table>

3
For the Filipino language learners unfamiliar with the Catholic religion, these words are not easily grasped. Further explanation and elaboration are needed to make them understand the meaning of these words within the conceptual system of the Filipino language.

These are some of the sample words causing challenges in building a bilingual Malay-Filipino, Filipino-Malay dictionary. Having said this, by looking further into the meaning of the said words, we can see that there are certain similarities between these words. This paper makes use of Frame Semantics\(^1\) initiated by Fillmore (1968) and subsequently developed by other linguists. However, in this case, we will define our own frames according to certain primes. Given the word below, we can give a frame of related meanings according to:

### Table 3. Semantic Frame for ‘prayer beads’.

<table>
<thead>
<tr>
<th>Primers</th>
<th>‘prayer beads’</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Filipino – rosaryo</strong></td>
<td><strong>Malay – tasbih</strong></td>
</tr>
<tr>
<td>Length (no. of beads)</td>
<td>60 beads + 1 crucifix</td>
</tr>
<tr>
<td>Use</td>
<td>To count prayers and recitations</td>
</tr>
<tr>
<td>Prayer</td>
<td>3 Mysteries (Joyful, Sorrowful and Glorious), Hail Mary, Our Father, Glory Be and Apostle’s Creed</td>
</tr>
</tbody>
</table>

From this frame, we can see that *rosaryo* and *tasbih* share related meanings, although the prayers used may be different. Thus, we can propose an entry such as:

**Rosaryo** – alat untuk mengkabulkan doa dalam agama Katolik; seperti tasbih dalam agama Islam\(^2\).

**Tasbih** – butil-butil na nakatuhog gamit sa pagdadasal sa Islam; tulad ng rosaryo\(^3\).

### Table 4. Semantic Frame for ‘a month-long fasting and prayer’

<table>
<thead>
<tr>
<th>Primers</th>
<th>‘a month-long fasting and prayer’</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Filipino – kuwaresma</strong></td>
<td><strong>Malay – Ramadan</strong></td>
</tr>
<tr>
<td>Length</td>
<td>40 days</td>
</tr>
<tr>
<td>Reason</td>
<td>Commemorates Jesus’ retreat in the desert and his sufferings before crucifixion, and death.</td>
</tr>
<tr>
<td>Activity</td>
<td>Prayer, fasting, abstinence from meat, alms-giving</td>
</tr>
<tr>
<td>Final celebration</td>
<td>Easter Sunday</td>
</tr>
</tbody>
</table>

From this we can propose an entry:

**Kuwaresma** – dalam agama Katolik, orang-orang berpuasa ataupun menahan diri makan daging, bersembahyang dan membagi sedekah; seperti Ramadan\(^4\).
Ramadan – pag-aayuno tuwing ika-9 na buwan ayon sa kalendaryo ng Islam; tulad ng kuwaresma³.

2.2. Cultural Terminologies

In this section, we will look into some terminologies relating to culture. There are certain words, despite sharing similar cultural background, still do not have an almost similar meaning. Malaysia and the Philippines share a lot of cultural practices prior to the arrival of Islam and Christianity. For example, the concept of marriage, both countries practiced certain customs and traditions before and during the wedding ceremonies. But due to the influence of religion, in the Philippines for example, the practice is reduced to the engagement (only when necessary) and the wedding ceremony itself. Also, due the influence of Western culture, bridal showers and bachelor’s parties are being practiced nowadays.

Typically, before marriage, the concept of panligaw ‘courtship’ is practiced among single men and women. Eligible, single women are courted with harana ‘serenade’, then when they accept, the lucky guy will have to do paninibihan ‘servitude’ to prove his love and faithfulness to the woman and to her family. When they all agree, pamamanhikan ‘engagement; meeting of families’ is set. Parents of both parties meet and set the date of the wedding day. The wedding day itself is called Araw ng Kasal or kasalan which is taken from the Spanish word casar ‘to wed’.

In Malaysia, the traditional wedding ceremonies are still being followed. After the adat merisik where the bachelor’s family found a suitable bride, will the process of marriage begin. The concept of engagement is called (adat) bertunang normally held at the bride’s house. It is then followed by akad nikah ‘signing of marriage contract, solemnization’, and then followed by the wedding ceremony itself, persandingan or bernikah⁶.

Malay weddings are more complicated compared to a Filipino wedding. However, in setting up a semantic frame for the words related to marriage, there seem to be some overlap.

![Semantic distribution of the concept ‘engagement’](image)

However, the word pertunangan is a more appropriate equivalent of pamamanhikan than the other two. In this case, engagement in Malay is pertunangan and in Filipino pamamanhikan.

Pamamanhikan – pertunangan
Pertunangan – pamamanhikan

In terms of supernatural beings, Malaysia and the Philippines share quite a number of them. Although there are variations, the ideas are somewhat the same.
Table 5. Semantic Frame for ‘banshee’

<table>
<thead>
<tr>
<th>Primes</th>
<th>‘a changeling; a demon-child’</th>
<th>‘banshee or a ghoul’</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Filipino – tiyanak</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appearance</td>
<td>Normal and cute baby</td>
<td>Beautiful woman</td>
</tr>
<tr>
<td>Sign of existence</td>
<td>Sound of crying baby</td>
<td>Sound of crying baby</td>
</tr>
<tr>
<td>Origin</td>
<td>A baby who died before being baptized; an aborted fetus; spirit of an unborn child whose mother died during labor.</td>
<td>A stillborn child; a woman who died while in labor.</td>
</tr>
<tr>
<td>Food</td>
<td>Feeds on blood</td>
<td>Feeds on blood by sucking through its long nails which functions like a straw.</td>
</tr>
</tbody>
</table>

Although we see that there are some similarities, one striking difference is the appearance or shape of the banshee. A *tiyanak* is a child or an infant while a *pontianak* is a woman. However, another mythical being in Filipino, the *aswang*, is more related to the Malay *pontianak*. The Filipino *aswang* is a woman who practices black magic or sorcery or *kulam* and transforms into a ghoul at night to feed on human blood. This is a very important category in our frame. Thus, we cannot put them as co-equivalents since a very important is not satisfied.

*Aswang* has other variations which are shown in Figure 2. *Aswang* is sometimes taken as a general term for any of these. The general feature is that all of these are blood-eating creatures.

![Figure 2. Variations of *aswang*.](image)

A *pontianak* has another variant which is a *langsuir*. It is said to be the most violent of the *pontianaks*. The same with *aswang*, it feeds on humans and human blood.

Table 6. Semantic Frame for ‘banshee, a ghoul or a vampire-like creature’

<table>
<thead>
<tr>
<th>Primes</th>
<th>‘banshee or a ghoul’</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Filipino – aswang</strong></td>
<td></td>
</tr>
<tr>
<td>Appearance</td>
<td>Beautiful woman</td>
</tr>
<tr>
<td>Sign of existence</td>
<td>a boiling oil mixed with herbs and incanted with prayers</td>
</tr>
<tr>
<td>Origin</td>
<td>A normal person practicing black magic and transforms</td>
</tr>
<tr>
<td><strong>Malay – pontianak</strong></td>
<td></td>
</tr>
<tr>
<td>Appearance</td>
<td>Beautiful woman</td>
</tr>
<tr>
<td>Sign of existence</td>
<td>Sound of crying baby</td>
</tr>
<tr>
<td>Origin</td>
<td>A stillborn child; a woman who died while in labor.</td>
</tr>
</tbody>
</table>
itself into a vampire-like creature

| Food    | Feeds on blood, human heart and liver, human fetus. | Feeds on blood by sucking through its long nails which functions like a straw. |

From this we can give an entry:

**Asuwang/aswang** – a vampire-like creature feeding on human flesh and blood. Usually associated with women practicing black magic (or a witch) but men can also be one. Similar to Pontianak.

**Pontianak** – isang aswang na nagpapangggap na magandang babae. Sinisipsip ang dugo ng tao gamit ang kanyang mga kuko. Tulad ng aswang.

3. DISCUSSION AND CONCLUSION

Given the samples in the previous section, we can see that some words have overlapping meaning while some have no similarities at all. To find their similarities by putting them in a semantic frame, we can set some categories to find a probable prime in order to give an appropriate definition. Although at first, the idea behind Frame Semantics is not like this. It was used to define words according to the categories it has (Petrick, 1996). However, in this paper, we used Frame Semantics as a guide to set up categories in defining the set of religious and cultural terms presented in this paper. By doing so, we would be able to find certain similarities in order formulate a ‘friendlier’ set of definitions.

There have been no published Malay-Tagalog or Tagalog-Malay bilingual dictionaries to date. This may be attributed to the audience this type of dictionary may cater to. Since the two languages are now being taught in universities and private language centers in the Philippines and Malaysia, the need is slowly developing. In the attempt to build a corpus for this dictionary, the above-mentioned difficulties were faced. Some words in both languages are not easily translatable despite their similarities. But, by developing a set of frames with categories, it might in developing the corpus of the dictionary.

This paper is just an initial analysis to the corpus being developed and collected for the said dictionary. Further analysis of the data at hand is being taken in order to come up with a good set of dictionary entries.

4. ENDNOTES

(1) A website devoted to the use of Frame Semantics can be found at [http://framenet.icsi.berkeley.edu/](http://framenet.icsi.berkeley.edu/).
Frame semantics refers to a different variety of approaches to systematically describe the meanings of words ([www.uni-stuttgart.de/linguistik/sfb732/files/hamm_framesemantics.pdf](http://www.uni-stuttgart.de/linguistik/sfb732/files/hamm_framesemantics.pdf)).
(2) Prayer beads (Catholic religion); similar to tasbih in Islam.
(3) prayers beads (Islam); similar to rosaryo.
(4) In the Catholic religion, people fast or abstain from eating meat, pray and give alms; similar to Ramadan.
(5) Fasting every 9th month of the Islamic calendar; similar to kuwaresma.

5. REFERENCES

(8) http://www.mythicalcreaturesguide.com/page/Aswang
Abstract: Sangam classics is an ancient and cultural identity of Tamil Language, it dates take to several centuries. It is having wealthy of words and their meanings. Lexicography an applied linguistic field is developing for dictionary making worldwide in all languages. So the present paper tries to explain the compilation of synonyms dictionary for sangam classics of Tamil. In sangam classics many lexical items were in practice even today. But the Learners of Tamil Literature students will face lot of problems in understanding the meaning of ancient lexical items. In this context the present paper depicts on the compilation of synonymous dictionary in Tamil. The synonyms means many lexical items having similar or related meaning or sameness of meaning. Further, this paper tries to bring out the sample entries in the compilation of Synonymous dictionary. This is a pioneering attempt in compilation of Synonym dictionary for Sangam classics. It will be useful to the learners of Sangam classics, I hope. Further all the sets of synonymous words were given as a separate entry according to their alphabetical order and in addition to their meaning and their citation reference were also given.

1. Introduction
Sangam Classics is an ancient and excellence cultural identity and antiquity of Tamil Language. The period of Literature of Sangam identifies and confirms the Golden Age of Tamil. The research work and study of Sangam Literary classics is basically on Grammar, Sociology and Human anatomy. Soon after introduction of Linguistics, simultaneously Grammatical research work also finds its place. Lexicography an applied linguistics field is developing for dictionary making worldwide in all languages. Sangam classics are having wealthy of words and their meanings. But as Department of Tamil is growing in its production of Dictionaries the research work on this is undoubtedly inevitable when comparing other languages. So that the compilation of Dictionary in Tamil is an essential need. On its base, the present paper depicts on the compilation of Synonymous dictionary in Tamil.
2. **Compilation of Synonymous Dictionary**

The various Nouns of Sangam classics are collected with utmost care and art of the collected nouns, synonyms are separated and compiled this synonymous Dictionary. Nikandus is a kind of collection of word groups of Synonyms, Polysemous, antonyms and Phrases but this dictionary differs from the above. Also this is pioneering attempt in compilation of synonyms dictionary for sangam classics.

3. **Aim of the Dictionary**

In Sangam classics many lexical items were in practice even today. But the Learners of Tamil Literature Students facing lot of problems in understanding the meaning of sangam poems and ancient lexical items. In this context the present paper depicts on the compilation of this dictionary. Secondly when comparing Tamil with other languages, lack of different kind of dictionaries are there and is in need of it. So the effort is made. There is possibility of learning the different words and the excellence of Sangam classics with this synonym Dictionary and it is also helpful for easy understanding of the meaning of poems in Sangam Literature for foreigners who learn Tamil also. From this it is as clear as crystal the essential need of synonyms Dictionary.

4. **Structure of Synonymous Dictionary**

The synonymous Dictionary is compiled out the consolidation of the nouns from Sangam classics. The sample entry is given below

```
Bζ ⊗ (¬φ.) — ε∫ | ζ@| σ∫≡ζ an wild animal

1. χΔφ ( ⊔μ.61)
   ¬φΩ ≡>ζ | :ζ| ⊃{, ¬Σ} ...Bζ[ χΔφ_ ( ⊔μ. 61)
2. χκζ ( A⊕. 374:10)
   :ζ[| ] :| ≡| . |ζ[| ...> | δ| → χκζ ( A⊕.374:10)
3. |B:ζ v φ \gR | |κζμ | ≡→ YΛ (φ ).20:19
   \ (φ ).10:42
   :φ v / | J | , κB φ | \ :ζ_ φv→ ( φ ). 10:42)
4. \ (A⊕. 4.10)
   \ =→ (A⊕. 4.10)
   \ =→ (A⊕. 4.10)
   \ =→ (A⊕. 4.10)
5. κζδΦ (A⊕.4:10)
6. ζ≡→ δ(δ | .145:15)
```
The above entry refers the entry word Elephant and its synonyms. Out of words / General common terms collected the well known, the best useful and mostly familiarized word are mostly will enter as head word or lemma. The number of synonyms of head word is also given in the brackets and the synonyms are listed and arranged in alphabetical order. The words occurrence is also noted in the lines of poems which are bracketed. All synonyms of a particular Head word, its number of poems and its occurrences in the poems are clearly shown. This is a pioneering and initial attempt in compilation of synonyms’ Dictionary for Sangam classics.

In continuation of this work, confirming that each word really belongs to synonyms then the entry is made. Due attention and care is to paid that among the confirmed words, some words are polysemous meaning. In addition the synonymous words of Birds, the synonymous of animals, the synonymous words of nature, the synonymous words of metonymy, the synonymous words of perception, Love and Bravery or heroism are prevailing in Sangam Classics. The recognition of these synonymous entry are separately isolated and identified and the research on this is being done now. Also the way of presenting these synonymous identify is being under research. Opinions valuable ideas and suggestions of eminent experts are humbly requested at the introduction of initial compilation of synonymous dictionary. Acquired valuable opinions on this really will induce to read out this article and this paper may be refined.

The Sangam classics are classified in two major concepts, akam (Love) and puram (Bravery). According to the needs and uses in the above two concepts all the words are created in Sangam classics and the due attention is to be paid in entries of these words. Like this, the words or the vocabulary of this increases and enriches day by day in Sangam classics also considerable. For this situation so many entries can be shown.
5. Uses of synonymous Dictionary

1. It removes the non-availability of synonymous dictionary in Tamil.

2. It brings out the prosperity of wealthy words from Sangam Literature and it proves immortality of ancient cultural identity.

3. It makes understanding of the meaning of the poems of Sangam Literature.

4. It is a small contribution to the Lexicography.

5. It makes interest of the Non-Tamilians to learn Tamil.

6. It makes the necessity of Synonymous dictionary to Tamil students and thereby the essential need of compilation of synonymous dictionary.

6 Conclusion

Compilation of Synonymous Dictionary is introduced through Sangam Literature in this paper. To make this article crystal clear, opinions and views from high learned persons are humbly solicited through this seminar.
SIGNIFICANCE OF CUTAMANI NIGHANTU IN TAMIL

C. Sarada
Vtminss College, INDIA
saradaramesh2003@yahoo.com

The word meaning is of primary importance in a nighantu. Tivakarar’s Tivakaram (9th century) which has the pride of place among Tamil Nighantus was followed by Pinkalar’s Pinkalam (10th century) were written in ‘Sutra’ style. In the 16th Century Mandala Purudar made an in-depth study of these two Nighantus and wrote Cutamani Nighantu comprising of 1197 verses in rhyme (‘ethukai’) and in viruttam metre. Simpler and easy to learn by rote has become so popular over the years that the very mention of Tamil Nighantu brings to mind no other word than Cutamani Nighantu. The present study is an attempt to bring it into the limelight of the world at large. This nighantu comprises twelve sections: section one to ten deal with synonyms, section eleven is devoted to polysemy and the last section is about collective nouns. The first section deals with the collection of nouns associated with divinity, heavenly bodies and related matters, and the second section to human beings, and the third fauna, fourth flora, fifth topology, sixth miscellaneous things, seventh artificial things of various shapes and material, eighth quality of abstract and concrete things, ninth nouns denoting action, tenth nouns denoting different sounds, eleventh section word meaning section of which the vowel consonant was given in the series from. Kakara to nakara ‘ethukai’ and the twelfth of collective nouns. The eleventh section served as the beacon light in the development of dictionaries and it is proposed to prepare headword Dictionary. This Nighantu is veritable treasure house of knowledge and vocabulary and for egrounds the socio-cultural heritage.
TEACHERS’ USE AND KNOWLEDGE OF ELECTRONIC DICTIONARIES

Atipat Boonmoh
King Mongkut’s University of Technology Thonburi, THAILAND
atipat_b@yahoo.com

Research shows that electronic dictionaries have become popular in many South and East Asian countries. Advances in technology have helped electronic dictionaries become easily available. In Thailand, a growing number of Thai students are using various modes of electronic dictionaries i.e. pocket electronic dictionaries, dictionaries on CD-ROM, or online dictionaries. It is assumed that teachers are not familiar with the electronic dictionaries their students are using. The knowledge of and attitudes toward electronic dictionaries among teachers are an important factor in the process of EFL learning and teaching, and the learners' success or failure with the language. This study, therefore, presents a survey on teachers’ own use and knowledge of electronic dictionaries. Seventy teachers of English from eight universities in Thailand were asked about their experience of, and attitudes toward electronic dictionaries. The results of the questionnaire show a current situation of dictionary use in Thailand that most teachers did not use electronic dictionaries and they were uninformed about them. They had little idea about the lexicographical content of electronic dictionaries. They were not aware of many technological features that were contained in electronic dictionaries. The results may help teachers be better informed about the electronic dictionaries so that they can be in a better position to provide authoritative advice on how to buy and how to train students to use electronic dictionaries effectively.

Key Words: electronic dictionaries; teachers’ attitudes; dictionary use
Abstract: The Latin-Portuguese-Japanese dictionary *Dictionarium Latino Lusitanicum, ac Iaponicum* (DLLI) was compiled and published by Jesuits in Japan in 1595. It was based on Ambrogio Calepino’s Latin dictionary (Calepinus), which was published in Europe. According to the title and preface, the DLLI was edited to serve as a guide not only for young Japanese students of Latin but also for Europeans wishing to study Japanese further. From the perspective of the historical process involved and judging by the actual text, it is clear that the DLLI—just like Calepinus—was mainly compiled for learners of Latin. The Jesuits edited the DLLI for Japanese students learning Latin in Jesuit schools. Compared with the 1570 edition of Calepinus published in Lyons, the DLLI contains many entries aimed at aiding learners of Latin: the DLLI contains the inflected forms of many headwords, mostly omitted in Calepinus. As stated in the notes of the DLLI, Japanese equivalents were added for listing synonyms to aid Japanese learners of Latin. However, we can find several inadequacies in the DLLI as a dictionary for Japanese learners in comparison with the Japanese-Portuguese dictionary *Vocabulario da lingoa de Iapam* (VLI) printed by Jesuits in 1603–1604; there is, among other things, a dearth of colloquial words in Japanese and lack of information on the usage of the entries. The Jesuits used their experience in editing the DLLI as a dictionary for learners of Latin in compiling the VLI as a dictionary for learners of Japanese.

Key Words: *Dictionarium Latino Lusitanicum, ac Iaponicum*; Calepinus; Jesuit mission press in Japan; *Vocabulario da lingoa de Iapam*

1. INTRODUCTION

Ever since the first Jesuit missionaries came to Japan in 1549, the Jesuits, most of whom were Portuguese, studied the Japanese language and culture to facilitate the propagation of the Catholic faith. It is believed that they have written manuscripts on grammar and dictionaries prior to the 1590s; unfortunately, these works cannot be found today.
In 1579, soon after the arrival of Visitador Alessandro Valignano, the Jesuits began to restructure their educational system; they began schooling the Japanese in a European style with the aim of training native priests. At the behest of Valignano, the Jesuits spent considerable time teaching Japanese students Latin in addition to Japanese literature and other subjects, while the European missionaries continued to study the Japanese language and culture.

Moreover, with the same aims, namely, the education of Japanese Christians and the study of the Japanese language and culture by missionaries, the Jesuits printed various books in Japan using European printing machines from 1590 onwards. Unfortunately, only around thirty titles survive today. These books can be generally classified into three categories: books on religion, literature, and language. The third category concerns the study of Japanese or Latin, and consists of dictionaries and books on grammar. A list of the grammar books and dictionaries that were printed by the Jesuits in Japan is given below:

1594 De institutione grammatica (Latin grammar book originally by Manuel Alvarez)
1595 Dictionarium Latino Lusitanicum, ac Iaponicum
1598 Racuyoxu (Dictionary of Chinese characters)
1603–1604 Vocabulario da lingoa de Iapam (Japanese-Portuguese dictionary)
1604–1608 Arte da lingoa de Iapam (Japanese grammar book by João Rodrigues)

In this paper, I would like to focus on one of the above-mentioned works, namely, the Latin-Portuguese-Japanese dictionary Dictionarium Latino Lusitanicum, ac Iaponicum (DLLI) as a dictionary for learners of Latin and Japanese. Several Jesuits, both Japanese and European, are said to have edited the DLLI; however, their names are not known.

2. DLLI AS A DICTIONARY FOR LEARNERS OF LATIN

The dictionary has a total of 456 folios, including the title page, preface, body, supplement, and errata. It has a lengthy title as given below, describing the original text, the method of editing, and its objective.

DICTIONARIVM / LATINO LVSITANICVM, AC / IAPONICVM EX AMBROSII CALE- / pini volumine depromptum: in quo omissis no- / minibus proprijs tam locorum, quàm homi- / num, ac quibusdam alijs minús usitatis, omnes vocabulorũ / significationes, elegantioresque dicendi modi apponuntur: / in vsum, & gratiam Iaponicae iuuentutis, quae Latino idiomati ope- / ram nauat, nec non Europeorũ, qui Iaponicũ sermonem addiscunt. /
The title clearly states that it is a Latin-Portuguese-Japanese dictionary based on the dictionary complied by Ambrogio Calepino, but with proper nouns and unusual vocabulary omitted and all the meanings of standard vocabulary, along with elegant examples of their usage, included. Moreover, there is evidence that the DLLI was edited to serve as a guide for Europeans who wished to study Japanese further, as well as for Japanese students of Latin. In the main body of the dictionary, the Latin entries are presented alphabetically, and they are followed by *Lus.*, which means ‘in Portuguese’, and *Iap.*, which means ‘in Japanese’.

In light of the fact that the Jesuits decided to edit and publish the DLLI to facilitate Latin education at their schools, we should regard it first as a dictionary for learners of Latin. Then, primarily on the basis of (Kishimoto, 2005a), I would like to compare the European dictionary, Calepinus—treated as the original edition in this paper—and the DLLI as a dictionary for learners of Latin, and discuss the findings. The title and preface of the DLLI give evidence that it was based on the Latin dictionary compiled by Ambrogio Calepino (1440?–1510?), of which more than 200 reprints were published between the sixteenth and eighteenth centuries in Europe; these works have been listed by (Labarre, 1975). (Bravi; Ceresoli; Lo Monaco, 2002) indicated six reasons why Calepinus received favourable reviews in those days. I think I can summarize that it was because the dictionary provided much information on the practical usage of the prescriptive Latin language based on classics. Later editions also incorporated equivalents in many other languages, but the Portuguese and Japanese equivalents were not included in any European edition. The Jesuits in Japan were the first to translate the entries into Portuguese and Japanese, and include them in the DLLI.

A comparison of the Latin entries and explanations reveals that the DLLI must have been edited on the basis of one of the Calepinus editions derived from the 1570 edition published in Lyons and which includes seven languages (Latin, Hebrew, Greek, French, Italian, Spanish, and German). There are known to be close to 17 editions derived from the 1570 Lyons edition, of which I have not yet decided on one; therefore, in this paper, I use the 1570 Lyons edition to quote the original European Calepinus.

When we compare the DLLI with the 1570 Lyons edition of Calepinus (CA70L), we find that the DLLI was considerably simplified and modified in various respects—entries, subentries, translations, Latin quotations, etc. (Kishimoto, 2007) summarises the rough process of the translation as follows: ‘the editors chose the Latin entries and the necessary parts of the Latin explanations, then translated them into Portuguese and Japanese.’ Though the Jesuits referred to other European dictionaries such as the first printed Latin-Portuguese
dictionary compiled by (Cardoso, 1570), the Portuguese and Japanese translations of the DLLI are very close to the entries in the original Calepinus.

(Kishimoto, 2005a) indicated several improvements made in the DLLI for learners of Latin from the European original: rearrangement of the order of entries, provision of the declension and conjugation of the entries, differentiation of meanings under an entry, and presentation of new entries derived from subentries. I would like to show this with examples. (1)

For the convenience of learners, the editors seem to have rearranged the order of the entries.

CA70L:8
Abominor, aris, [to loathe (verb)]...
Abominandus, [being loathed (gerundive)]...
Abominatus, [loathed (perfect particle)]...

DLLI:5
Abominandus, a. um. ...
Abominatus, a, um. ...
Abóminor, aris. ...

CA70L:1323-1324
Verbum, bi, [word]…
Verbiger o [to exchange words]…
Verbosus, a, um [containing many words, talkative]…
Verbosè [with many words]…

DLLI:862
Verbiger o. as. …
Verbosé …
Verbosus, a, um …
Verbum, i.

In the original Calepinus, words were not always placed in alphabetical order as indicated by the examples. In the above examples, the base words, Abominor and Verbum, were placed at the top and were followed by their derivatives. However, in the DLLI, the alphabetical system was employed. I have been unable to find this order in any other Calepinus edition. The editors possibly rearranged the original order of the entries to help
users look up individual words.

Furthermore, in the example of *Abominor* and its derivatives, we can see the difference in the form of the entries. In the European Calepinus editions, inflections are often omitted; however, in the Japanese edition, the inflections of each word—for example, *Abominandus, a, um*—are indicated. This must have been done to provide learners with a correct understanding of Latin declension and conjugation.

CA70L:4
Abecedarius, dicitur rei literariae tyrunculus, qui adhuc in literis cognoscendis haeret. (in Greek) Accipitur item pro eo qui aliquid per ordinem alphabeti digerit, vel pro eo quod alphabetico oridine digestum est. [*Abecedarius* means a beginner of studying letters who remains at the early stages of learning letters, and also means someone who arranges something in alphabetical order, or something arranged in alphabetical order.] Vide Caelium libro 14. capite 28. & Budaeum in Pandect.

DLLI:3
Abecedarius, ij. Lus. Menino que aprende as primeiras letras. [a child who learns the ABC] Iap. Yrofano narai fajimuru varambe. [a child who has begun to study the Japanese alphabet] Item, O que ordena algûa cousa por ordem do alfabeto. [same, a person who puts something in alphabetical order] Iap. Yrofano xidaini monouo caqitateraru fito. [a person who enumerates something in Japanese alphabetical order] Itê, Cousa ordenada por ordem do alfabeto. [same, something arranged in alphabetical order] Iap. Yrofano xidai vomotte caqitateraretaru mono. [something enumerated in Japanese alphabetical order]

The DLLI provides three meanings of the word *Abecedarius*; however, this distinction is not found in the 1570 Lyons edition. The editors of the DLLI should have analysed the Latin explanation of the three Portuguese meanings to provide learners with a clear definition.

The following example shows the new entries of the DLLI derived from the subentries of the original. The Calepinus has the subentry *Abauia* under *Abauus*; the DLLI, on the other hand, not only has *Abauus* [great-great-grandfather] but also *Abauia* [great-great-grandmother] as entries.

CA70L:3
Abauus, penult. cor. Proaui pater. ... ¶ Sic abauia, mater proaui. ...
It seems that the DLLI’s editors made considerable changes to the original Latin dictionary for the learners of Latin in Japan: they omitted explanations and quotations which they did not consider to be very useful, and rearranged the entries and subentries; they also added Portuguese and Japanese translations.

3. DLLI AS A DICTIONARY FOR LEARNERS OF JAPANESE

(Kishimoto, 2005a, 2005b, 2007) provided several examples of the Japanese translations used in the DLLI to aid the Japanese people who were not familiar with European culture: the DLLI’s editors often provided Japanese translations with additional explanations—which cannot be found in the original—or included translations that traversed cultural differences, by using words more suited to the Japanese or Chinese context and which could not be found in European languages.

3.1. Synonyms in the Japanese Translations

However, as the title and preface of the DLLI indicate, the dictionary aimed at not only the learning of Latin but also of Japanese. Their consideration for the European people who learned Japanese is quite evident in that they tried to provide synonyms in the Japanese translations. The explanatory notes that follow the preface read as follows:

¶ Latinis vocibus variae interpretationes sūt affixae, tū quia aliquādo ita latitudo Latini vocabuli postulabant, tum vt varia synonima adhiberentur, quò Europeis, qui Iaponīcè loqui velent, verborum copia affluentius suppreteret. [Various explanations have been appended to Latin words; sometimes, the Latin words require such explanations, and sometimes, various synonyms are provided to aid Europeans who wish to speak Japanese enrich their vocabulary.]

¶ Interdum in declaratione Iaponica aliquid minus politè, & eleganter dictum est, vt Latini vocabuli vis apertiūs, & enucleatiūs exponeretur. [Japanese declarations are sometimes delivered less politely and sometimes elegantly; therefore, the meanings of the Latin words should be expressed more clearly and in more simple terms.]

The aim of the DLLI’s editors becomes more evident when the dictionary’s entries are compared with the original Latin in Calepinus. In the entry Abbreuiio [to abbreviate], the
original Calepinus does not provide the meaning of the word but includes equivalents in other languages as well as quotations. The DLLI’s editors provided three Japanese synonyms for ‘to shorten’: *Riacusuru, mijicamuru,* and *tçuzzumuru.*

CA70L:3-4

Abbreuiio, as, are. { (equivalents in six languages) } Vegetius in prologo lib. 3. Quae per diuersos autores librósque dispersa, Imperator invicte, mediocritatem meam abbreuiare iussisti, &c.

DLLI:2


Moreover, I can prove that the basic Latin words are generally accompanied with synonyms and additional explanations in Japanese. As given below, in the original Calepinus, the example *Autem* [on the other hand, but] contains only Latin synonyms, brief grammatical information, and equivalents in six languages. However, the DLLI contains three Japanese equivalents—va, sarinagara, and *ni voiteua*—with examples *fitoua yuqu, vareua cayeru* and *fitoua yuqutomo, vareni voiteua cayerô* in Japanese. The two sentences seem to be original, that is, not based on the European Calepinus.

CA70L:141

*Autem, Pro, sed, siue, verùm, Coniunctio est tantùm subiunctiua. [A conjunction is only used after another word] { (equivalents in six languages) }*

DLLI:78

*Autem, Lus. Mas. Iap. Va, sarinagara, ni voiteua, vt fitoua yuqu, vareua cayeru, l, fitoua yuqutomo, vareni voiteua cayerô. [He will go and I will return, or I will return even if he goes.]*

The DLLI lists three Japanese synonyms for *Inhabito* [to dwell in, to inhabit], as shown below. The Japanese word *sumu* is the most basic word including the meaning of *sumai suru* and *qiogiù suru.* However, the Japanese verb *sumu* has two homonyms; the others mean ‘to become clear’ and ‘to dive’. We can assume that the editors tried not only to provide as many synonyms as possible but also to avoid misunderstandings and clarify the meaning of the entry.

CA70L:632

*Inhabito, as, are, In loco aliquo habito. [To dwell in some place] { (equivalents in six

DLLI: 375


3.2. From the DLLI to the Vocabulario da lingoa de Iapam (VLI)

As we have seen, the DLLI was compiled with the aim of providing a dictionary for not only learners of Latin but also of Japanese. However, when we consider whether it was more useful as a Latin dictionary or as a Japanese dictionary, it is clear that the DLLI was mainly edited to aid learners of Latin. This view is supported when we compare the DLLI with Vocabulario da lingoa de Iapam (VLI), the Japanese-Portuguese dictionary published more than eight years after the DLLI was compiled by the Jesuits.

The DLLI seems to have less than 30,000 Japanese words and the VLI, much more than that, as (Fukushima, 1979) surmised. Although it is clear that the VLI has more Japanese words than the DLLI, we should focus on the vocabulary each one includes.

It would not be an exaggeration to say that the VLI is one of the most valuable dictionaries in the history of the Japanese language; it contains various kinds of Japanese vocabulary such as technical terms (those of Buddhism, Japanese literature, etc.), dialects, and colloquial expressions. As far as I have seen, the VLI never consisted of only the Portuguese-Japanese parts derived from the DLLI, though as a matter of course, we can find many similar instances of equivalents.

The compilers of the VLI often used explanatory notes below each Japanese entry. Regarding “X,” at that time the capital city of Japan was Kyoto, and as (Rodrigues, 1604–1608) points out, although the colloquial language used around Kyoto was thought to be the model for spoken Japanese, many of the missionaries worked at Kyushu in southern Japan.

B. = Baixo [vulgar word]
Bup. = Buppô [technical term for Buddhism]
P. = Poesia [word used for poetry]
S. = Scriptura [word used only in books and letters]
X. = Ximo [dialect in Ximo area, probably the Kyushu dialect]
On the other hand, the DLLI contains less variety of Japanese vocabulary: one of the features of the DLLI is that it employs written Japanese. As (Rodrigues, 1604–1608) states in the preface, there was a vast difference between colloquial and written Japanese in those days. (Otsuka, 2006) simply points out the difference in the Japanese employed in the DLLI and that in the VLI; he states that the former was edited in written Japanese for the purpose of preaching and the latter, mainly to provide vocabulary to be used by Japanese penitents during confession.

Though we should be careful in determining whether or not the DLLI contained colloquial Japanese, it is clear that on the whole, written Japanese was employed in the dictionary. One of the evidences which shows that written Japanese was employed in the DLLI is the usage of Nari at the end of sentences. According to the VLI, Nari means “be” in written Japanese, and is the same as Degozaru in the colloquial language.

DLLI:838  
Tributio, onis. [distribution] Lus. O dar. Iap. Atayuru coto nari. [to give]

VLI:178  
Nari. Na escritura. i, Degozaru. Ser. As vezes não serve mais que de elegancia cõ que se acaba algüa oração da escritura. Vt, Yomiqeru nari. Leo. [Used in written Japanese; same as Degozaru, that is, ‘Be’. It often adds elegance at the end of written text, for example, Yomiqeru nari. (He/she read.)]

Another evidence is given by the adjective forms. In colloquial Japanese, the old and written form Fucaqi was no longer used for the attributive form; instead, as (Rodrigues, 1604–1608) points out, the new colloquial form, Fucai, was used; this is also the case in Fucaxi (written) versus Fucai (colloquial) for the conclusive form.

DLLI:165  
Conuallis, is. Lus. Vale. Iap. Fucaqi tani. [valley enclosed on all sides]

VLI:105  
Fucai. Cousa profunda, ou de muito fundo. [to be profound, or of deep bottom] …

As we have seen, the DLLI was less convenient as a Japanese dictionary than the VLI. However, a comparison of the body of the two dictionaries and an examination of the preface of the VLI indicates several facts, including that the editors (and probably even the printers) of the VLI employed some basic aspects of the DLLI.
I have previously discussed the order and form of the Latin entries in the DLLI. Such consideration for learners was adopted when employing Japanese in the VLI. The preface of the VLI, given below, gives us a clear idea of the order of the entries.

NA colocação, & ordem dos vocabulos seguimos o modo que se teue no dicionario Latino ja impresso, não indo pollas cabeças, & deriuacão das palauras, mas pella ordem do Alphabeto Latino, por que assi se acha melhor, & mais depressa o vocabulo que se busca. [For the collocation and order of words, we did not arrange the derivatives under the base words but placed them in alphabetical order following the style of the Latin dictionary which was printed before, because it is a better and faster method for searching a word.]

Moreover, (Chonan; Doi; Morita, 1980) and (Morita, 1993) indicate other similarities between the VLI and the DLLI, as follows:

(a) Style of printing
They have a rather similar appearance with almost the same format and printing style.

(b) Form of verbs of entries
The VLI always gives the root and the present and past forms of the Japanese verbs, for example, Sumi, u, unda. It is similar with the Latin verbs listed in the DLLI.

(c) Explanatory notes
The VLI often uses explanatory notes similar to that used in the DLLI, such as per met [as metaphor] and Apud poetas [as used by poets].

It is not easy to confirm whether the DLLI directly influenced (b) and (c), because there is a possibility that the editors referred to other books besides the DLLI. However, we can assume that several features of the DLLI as a dictionary for learners of Latin were employed in the VLI as a dictionary for learners of Japanese. The Jesuits in Japan arrived at their almost ideal dictionary for learners of Japanese several years after the DLLI.

4. CONCLUSION

It is quite evident that the DLLI’s editors considered the learners of both Latin and Japanese. The DLLI was a considerably revised version of the European Calepinus which contained extensive information and quotations in Latin. However, the DLLI still has several inadequacies as a dictionary for learners of Japanese. The effective style used in the DLLI for aiding learners of Latin —for example, the forms of the entries and the explanations of usage—was applied by the Jesuits to compile the VLI for aiding learners of
Japanese. In conclusion, the Jesuits in Japan first edited a practical dictionary for learners of Latin on the basis of the European Calepinus and then compiled a dictionary for learners of Japanese on the basis of the DLLI and using their more than fifty years of experience of living in Japan.

NOTES
(1) In the quotations, the text placed within [ ] is the English translation and that within ( ) is the author’s note. The Calepinus printed in 1570 in Lyons used { } for presenting equivalents in the six languages, but they have been omitted in this paper.

REFERENCE
Primary sources

Secondary sources
(10) Kishimoto, Emi. (2005b) “Translation of Dictionarium Latino Lusitanicum, ac
Iaponicum (1595)”. Congresso Internacional A Presença Portuguesa no Japão nos séculos XVI e XVII, 47-52, Lisbon: Tipografia Lobão.


*This research was supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology KAKENHI (21720163) in 2009.
NEW TRENDS IN DICTIONARY USE: FROM CUNEIFORM TO ELECTRONIC GADGET

Li Lan
Hong Kong Polytechnic University, HONG KONG
eglilan@polyu.edu.hk

Computers have changed the form of language study fundamentally, and dictionaries mirror in tandem with the advancement of modern technology, from desktop e-dictionary to palmtop dictionary to mobile phone dictionary (m-dictionary). This paper reports an empirical survey consisting of a questionnaire, interviews and an experiment to investigate sociopsychological, psycholinguistic and pedagogic factors concerning e-dictionary use. It is worthwhile to look into how dictionary medium change affects user behaviour. The questionnaire and interviews reveal that paper dictionaries have given way to electronic dictionaries among the younger generation; the mostly used dictionaries are online dictionaries, followed by paper dictionaries and electronic gadgets. Dictionaries on mobile phones have been rapidly improved and are catching up with other lexicographic formats. The problem solving experiment focuses on e-dictionary use and demonstrates levels of user satisfaction when the users consulted e-dictionaries for understanding English metaphors in reading. The author argues that e-dictionaries could work as effectively as their paper counterparts, and with improved dictionary contents shifted from paper media to electronic media, e-dictionaries will probably replace traditional hard copies to a great extent.

Key Words: e-dictionary, user, change, effectiveness
Abstract: The traditional division between grammar and lexicon results in the belief that any combination of lexicon and grammar is acceptable, and inflected forms of a lemma share the same meaning. However, corpus linguists are much concerned with naturalness, trying to distinguish natural uses from unnatural (but grammatical) uses. Corpus-based studies have shown that inflected forms of a lemma share no certain similarity if their frequency and collocational associations are taken into consideration; there is a direct association between language form and meaning in the sense that each inflected form can be correlated with a specific pattern of usage. This paper addresses the form/meaning relationship in great detail by reviewing previous corpus studies on this issue and inspecting certain new lemmas in the Bank of English. It confirms the claim that inflected forms of a lemma tend to favour different collocates, which are associated with distinct semantic preferences and semantic prosodies. It argues that an awareness of such differences is helpful for L2 learners to produce natural language instead of unnatural language because the natural use of words depends not just on the choice of appropriate words, but equally on the choice of their proper forms. Therefore, such information should be included in learners’ dictionaries so as to help learners increase their understanding of the lemma/forms relationship, and improve their command of different usage patterns of various inflected forms, thereby making word learning more effective and language production more natural. In lexicographic practice, such information can be integrated explicitly into usage notes, starting with some typical forms (of the same lemma) which have been found to have clearly distinct meanings, or communicated implicitly by exemplification.

Key Words: lemma; word form; meaning; collocation; learners’ dictionaries

I. INTRODUCTION

The traditional division between grammar and lexis has been rejected by Sinclair because he (1987) finds that each sense/meaning of a word is associated with a specific syntactic pattern; one use of a word always distinguishes it in syntactic pattern from those of the other uses of the word. Sinclair sees lexis and grammar as two inherently connected parts of a single entity which can not be dealt with separately because “a grammatical structure may be lexically restricted” (Francis, 1993:142) and lexical items are often grammatical in nature (Biber, Conrad & Reppen, 1998; Hunston & Francis, 2000). Hoey (2005), in line with Sinclair, but from a psycholinguistic perspective, points out that we are primed to recognize specific patterns for different senses of a word and to produce them. He claims that the collocations, semantic associations and colligations a word is primed for
will systematically differentiate its senses and “the meanings of a word will have to be interpreted as the outcome of its primings, not the object of the primings” (p81). In other words, the meaning of a word (sense) lies in its collocations, semantic associations and colligations, not elsewhere. This is much in line with Teubert’s claims that meaning is not something that pre-exists in the reality or people’s minds; meaning is in the discourse; meaning is paraphrase and usage (for the details see Teubert, 2004, 2005, 2007 etc.).

The traditional assumption that ‘lemma and inflected forms are bound to share the same meaning and differ only in their grammatical profile’ (Tognini - Bonelli, 2001) has also been called into question by Sinclair since he believes grammatical pattern and lexicon are interwoven. A lemma is a dictionary head-word, realized by various word-forms. In a dictionary, an inflected form does not have a lexical entry of its own but appears in the lexicon within the lexical entries of its base form. Of course, it can be said that words are arranged in this way for the purpose of convenience, but it has led dictionaries users to take it for granted that various grammatical forms of a lemma share the same meaning and never question the association between them. Previous corpus-based studies have shown that “there is no guaranteed similarity between inflected forms if we consider frequency and their collocational associations” (Tognini-Bonelli, 2001: 92); inflected forms of a lemma favour different collocates, which may be associated with distinct semantic preferences or semantic prosodies (O'Halloran, 2007 etc.). These findings suggest that there is a much closer association between word form and meaning than was ever thought. Such complex linguistic phenomena are less accessible through human intuition; only corpus-based studies enable the researcher to reveal their existence and describe them.

Sinclair views lexicography as “the legitimate mode of expression of linguistic description at the lexical level” (Krishnamurthy, 2008:237). His observation that each use of a word is closely associated with a specific collocation or pattern has had a profound impact on subsequent lexicographical practice. His idea that semantic prosody is an important notion for lexicographic characterisation of words has also informed modern learners’ lexicography. Some modern learners’ dictionaries have included semantic prosodic meanings of some words in them. For example, the Cobuild dictionary (1987) defines *scrawny* as “unpleasantly thin and bony” rather than “thin and bony” and adds “an informal word, often used showing disapproval” to the definition of *prattle*. The LDOCE (2001) definition of *set in* has become ‘if something sets in, especially something unpleasant, it begins and seems likely to continue for a long time’. However, Sinclair’s finding that the similarity between inflected forms of a lemma can never be assumed has not led to any change in lexicography. Probably this is because it is not an easy issue to be dealt with in lexicographic practice.

In this paper, I will address the form/meaning relationship in great detail by first reviewing previous corpus studies on this issue and then inspecting certain new lemmas in the Bank of English. The purpose of this study is to show once more that inflected forms of a lemma have different collocational profiles and favour different collocates,
which are associated with distinct semantic preferences and semantic prosodies. In line with Hoey (2005), it claims that the meaning of a word (form) can be interpreted as the outcome of its collocation, semantic preference and semantic prosody. It argues that modern learners’ dictionaries should provide a more accurate and useful description of words by including information concerning the differences in meaning between inflected forms, describable in terms of their collocates, semantic preferences and/or prosodies, so as to help learners increase their understanding of the lemma/form relationship, and improve their command of different usage patterns of various inflected forms, thereby making word learning more effective and language production more natural. In lexicographic practice, such information can be integrated explicitly into definition, starting with some typical forms (of the same lemma) which have been found to have clearly distinct meanings, or communicated implicitly by exemplification.

II. PREVIOUS STUDIES ON WORD FORM AND MEANING

It is Sinclair who first (Sinclair, 1985, 1987, 1991, etc.) questions the traditional lemma/forms association as problematic. He examines some common words in a general corpus, such as *decline* (1985, 1991:41) and *yield*, (1991:53) or phrasal verbs such as those with *set* (1991:67), and observes that each inflected form is directly associated with a specific pattern of usage. For example, he finds that, for the lemma DECLINE (in this article, I will cite lemmas in upper case, and italicize word forms), "the form *decline* is associated with nominal usage, *declining* with adjectival usage, and *declines, declined,* with verbal usage" (1991:12). Similarly, Stubbs (1996) investigates the lemma EDUCATE in a general corpus of 130 million words and summarises his finding as below:

In 130 million words, frequencies were: *education* 27,705, *educated* 3,450, *educate* 858, *educating* 463, *educates* 29. The form *education* collocates primarily with terms denoting institutions (e.g. *further, higher, secondary, university*). The form *educate* with approximate synonyms such as *enlighten, entertain, help, inform, train*. (…) The form *educated* frequently collocates with *at* (often in the phrase *he was educated at*) and then with a range of prestigious institutional names, including *Cambridge, Charterhouse, college, Eton, Harrow, Harvard, Oxford, school, university, Yale*. The collocates of the form *educated* therefore provide a little list of ways in which men (much less frequently women) are classified and talked about. (Stubbs 1996: 172-173)

Tognini-Bonelli looks at the differences between *faced* and *facing*, two inflected forms of the verb *to face*, with a specific focus on their collocational profiles in a general subcorpus and a specialized subcorpus of economics respectively. She observes that, in the general corpus, *facing* is clearly more associated with FACE’S concrete meaning indicating position and direction and *faced*, in contrast, is more related to its abstract meaning. In the subcorpus of economics, both *faced* and *facing* are consistently abstract. It is not hard to understand such a result because in the domain of economics physical direction or position are hardly talked about. What is interesting is she finds that in the
specialized subcorpus a specific grammatical pattern is closely associated with the word form *facing*, but not with *faced*: the subject can either precede (40%) or follow the verb (60%). It shows that different forms of a lemma have taken on different roles through the association with different patterns.

Following Sinclair, William (1998) examines the word forms *gene* and *genes* in the context of molecular biology research papers and finds that their collocates are quite distinct, both to the left and the right of the node word. Doyle (2003, cited in Hoey, 2005) likewise finds that grammatical related forms of lemmas share few collocates in scientific textbooks. Hoey (2005: 8) gives a summary of Doyle’s finding:

… he looks, for example, at *amplifier, amplifiers* (only three shared collocates), *circuit, circuits* (only two collocates), *frequency, frequencies* (only one shared collocate) and *shift, shifts* when he finds no shared collocates at all.

Hoey argues that common collocates should never be assumed, though various forms of a lemma do share collocates sometimes, such as *training* and *trained*, sharing collocation with *as a teacher*.

In the 260 million words newspaper subcorpus of the Bank of English, O’Halloran (2007) examines the collocates for *simmering*, auxiliary+been *simmering*, *was simmering/simmered, erupted, erupted* in the past tense, *erupt(s), eruption(s)*, swept through with or without auxiliary respectively. He also extends the examination of *eruption(s)* and *had swept through* to the whole Bank of English. The investigation shows:

in the hard news register, ‘has been simmering’, as well as ‘erupted’ in the past tense, have a semantic preference for human phenomena, rather than for volcanoes, and carry a negative register prosody. The same is true for ‘erupts(s) although not to the same degree. However, there is evidence that ‘eruptions’ in collocation is much more likely to carry meanings associated with volcanoes inside and outside the hard news register. So, across different forms of the lemma, ‘erupt’ there would seem to be a cline of delexicalisation from ‘eruption(s)’ to ‘erupt(s)’ to ‘erupted’ (p20).

O’Halloran (2007) proposes the notion of register prosody based on his corpus finding that the semantic prosodies of grammatical word forms have a strong affinity with register, as mentioned in the quote above: *has been simmering* and *erupted* in the past tense are closely associated with a negative semantic prosody in the register of hard news. The finding on the lemma ERUPT seems to suggest that a cline of delexicalisation can be related to different inflected word forms of a lemma, with ‘literal’ and ‘figurative’ as the two extremities of a continuum of delexicalisation. For the lemma SWEEP, only 38 instances of *had swept through* are found in the Bank of English. Interestingly, among them, only one instance is associated with the *broom* meaning of *swept through*; all the others are used in its figurative sense, including 8 instances in the hard news register, such as ‘the flu epidemic which had swept through his squad’. O’Halloran emphasizes
that the number of instances of had swept through is small in the Bank of English, so there might be a danger of overgeneralising the data. He also notes that about 30 percent of the collocates of swept through in the newspaper subcorpus (437 instances in total) are associated with natural forces. Half of them (15 percent of the total number) semantically prefer fire, including 42 instances of fire, 10 of blaze and 8 of flames, with significant t-scores of 6.5, 3.2, and 2.8 respectively. He claims “there is indeed the possibility that ‘swept through’ is partially imbued with the meaning of ‘fire’” (p19) in the register of hard news.

Although various inflected forms of a lemma “show some overlap in meaning, it is clear that specific meaning dimensions and different connotations are associated with each of them” (Tognini-Bonelli, 2001: 99). In this respect, Sinclair says, “[t]here is a good case for arguing that each distinct form is potentially a unique lexical unit, and that forms should only be conflated into lemmas when their environments show a certain amount and type of similarity”. (1991:8). He thinks that “the distinction between form and meaning is only a methodological convenience” (Tognini-Bonelli, 2001: 99) and this leads him to “posit formal observations as criteria for analysing meaning” (ibid). He goes on to claim that “form could actually be a determiner of meaning”, and “[t]here is ultimately no distinction between form and meaning” (Sinclair 1991:7). Tognini-Bonelli (2001), in line with Sinclair, argues that language form is directly associated with its meaning; form and meaning are two aspects of the same phenomenon, and “to analyse the former according to objective criteria will yield insights into the latter” (2001:99). Teubert (2005:3) shows a similar idea by saying that “[e]very text segment, word, multi-word unit, phrase etc., can be viewed under the aspect of form and the aspect of meaning; [t]he form is what represents the meaning, and there is no meaning without the form by which it is represented”.

### III. A CASE STUDY

To explore further the relationship between inflected forms of a lemma and their linguistic behaviour, I will draw on the Bank of English, a well-known general English corpus containing 450 million words. In all of my collocation researches in this study, I will use its default span of 4:4. LookingUp, its searching software programme, provides the top 50 collocates by frequency, t-score and MI score within a span of four words to the left and right respectively of the node word. I will present the raw frequencies of the collocates. By looking at raw frequencies we can know roughly which collocations are frequently used. The statistical measure, t-score, tells us about “the certainty of collocation” (Hunston, 2002:73) by taking the size of the corpus into account. It provides a more precise measurement of the level of attraction between the node word and its collocates, reducing the probability of two words occurring together accidently rather than through semantic relations. T-scores of the collocates can be automatically generated by the LookingUp programme. Generally speaking, a t-score of above 2 is taken to be significant (ibid). I will present collocates above such a value.
I will illustrate and support Sinclair’s position with three examples, one verb and two nouns. The verb concerns its various grammatical forms and the nouns of its singular and plural forms. First, I will look at the inflected forms of the verb to flare. Their occurrence in the Bank of English is (the number before the slash is the raw frequency and the one after it is the proportion of total usage of each form):

<table>
<thead>
<tr>
<th></th>
<th>Verbal</th>
<th>nominal</th>
<th>adjectival</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>flare</td>
<td>146/13.7%</td>
<td>923/86.3%</td>
<td></td>
<td>1069</td>
</tr>
<tr>
<td>flared</td>
<td>1805/85.7</td>
<td></td>
<td>302/14.3%</td>
<td>2107</td>
</tr>
<tr>
<td>flaring</td>
<td>299/74.6%</td>
<td>17/4.2%</td>
<td>85/21.2%</td>
<td>401</td>
</tr>
<tr>
<td>flares</td>
<td>121/9%</td>
<td>1218/91%</td>
<td></td>
<td>1339</td>
</tr>
</tbody>
</table>

TOTAL 2371 2158 387 4916

From the proportion figures shown above, we can associate the forms flare and flares with nominal usage, flared and flaring with verbal usage. I shall further investigate the relationship between word forms and their usages by choosing two forms and looking specifically at their collocational profiles in the corpus. The forms chosen to illustrate my point are flaring and flared as verbal. The various usages of the verb can be broadly classified into two types: one is related to concrete things and the other is used figuratively. I will consider the two meanings when looking at the most significant collocates of the chosen forms. Significant collocates for flaring (299) as verbal in the corpus are as follows. In the inspection of the verb in this article, only relevant nominals with t-score above 2 are shown; the number before the slash in brackets is the raw frequency, and the one after it is the t-score:

nostrils(48/6.9), eyes(19/4.2), tempers(13/3.2), gas(10/3.1), nostril(6/2.4), waist(6/2.4), skirts(5/2.2), head(5/2.1), risk(5/2.1), sun(5/2.0)

From the above we can see that, of all the nominal collocates with t-score above 2, the great majority are concrete things, such as nostrils, eyes, gas, skirts etc. It should be noted that a t-score above 2 does not guarantee that it is semantically relevant to the node word. The figures above show that waist occurs 6 times with flaring as verbal, with a t-score of 2.4. However, a careful examination of the concordance lines indicates all instances of waist have nothing to do with flaring, as shown in the example: “There is also a flirtier short, shaped on Lana Turner lines, with a defined waist and flaring legs.” The figures also show that head co-occurs 5 times with flaring, with a t-score of 2.1. All the instances are shown below:

back her head, Szabla inhaled, flaring her nostrils. Justin wrinkled his
tilted her head back, her nostrils flaring as she tried to calm herself. Rex
Jerking up his head, nostrils flaring, neck magnificently arched, he
head, your shoulder moves, usually flaring to the left. If it were a full
voice, his head back, his white eye flaring. <p> `You don't put me out to
From the above, we can see that *flaring* is not related to *head* at all. Therefore, the appearance of *waist* and *head* in the top collocates of *flaring* is just because that they happen to occur with the node word in the span of 4:4 many times. Similarly, if *eyes* occurs with *flaring*, it does not necessarily follow that *flaring* is used to describe eyes, so we have to investigate the concordancing lines further. Among the 19 instances, *eyes* occurs after *flaring* in 5 of them, as shown below:

```
d charge in, knees pumping, nostrils flaring, his eyes afire, his long black
She put down her fork, nostrils flaring, eyes wide. "How the fuck did you
before, deep in his trance, nostrils flaring, eyes open now but unseeing, his
to leap. Smoke billowed, a light flaring in his eyes blinded him. out of
Position your lamp so that it's not flaring into your eyes, otherwise third
```

As can be seen, in the first three instances *flaring* is used to describe nostrils, and the other two examples are associated with light flaring. In 14 instances of *eyes* occurring prior to *flaring*, only 10 of them really talk about eyes flaring. In total, there are 12 instances in which *eyes* are associated with *flaring*. So *waist* and *head* should be removed from the figures above, and the frequency of *eyes* should be reduced to 12.

The only collocates that are related to abstract notions are *tempers* (13) and *risk* (5), with t-scores of 3.2 and 2.1 respectively. After the examination of the concordance lines, I finds that *risk* occurs with *flaring* 5 times is all because the risk of a disease or war flaring up is talked about in the corpus. Anyway, in all these cases the node word is used in its figurative meaning. However, as we can see that the frequencies of *tempers* and *risk* are still relatively lower compared to the concrete group (nostrils/48, eyes/12, gas/10, nostril/6, skirts/5, and sun/5), so we could conclude that *flaring* as verbal is much more associated with a concrete notion than an abstract one; it has a semantic preference for physical things, such as *nostril(s)*, *eyes*, *gas*, *skirts*, *sun*, and a neutral semantic prosody since descriptions of such phenomena can not fall into the positive or negative categories.

For the form *flared* (1805) as verbal, the following collocates are found:

```
trouble(183/13.4), violence(137/11.6), tempers(101/10.02), nostrils(69/8.3), fighting(53/7.1),
eyes (55/7.1), fire(37/5.7), anger(31/5.5), row(31/5.4), injury(31/5.3), trousers(25/4.9),
problem(30/4.8), tensions(22/4.6), temper(21/4.6), knee(18/4.1), match(22/3.9),
dispute(15/3.8), light(19/3.7), flames(14/3.7), war(22/3.4)
```

A simple glance of the collocates shown above will lead to the understanding that *flared* as verbal occurs very frequently with abstract things. The highest co-occurrences are *trouble*, *violence* and *tempers*, with t-scores being 13.4, 11.6 and 10.02 respectively. A t-score over 10 is very significant. From the above list, we also find some concrete nominal collocates: *nostrils*, *fire*, *trousers*, *match*, *light*, *flames*, but we can see that compared to the abstract group their t-scores are relatively lower at 8.3, 5.7, 4.9, 3.9, 3.7
and 3.7 respectively. For *eyes*, in 30 instances of 55, *eyes* occurs after *flared*. Of them, only 6 instances relate to concrete things, including “saw the sparks that flared in her eyes”, the whole of which is more interpreted as a metaphor, but we do not include such a use into our figurative group. All the other 24 instances talk about abstract things, such as *desire, determination, temper, terror* etc. For example:

New distrust flared in her eyes
Desire flared in his eyes.

It is interesting that *knee* occurs 18 times with *flared* as verbal in the corpus. After examining the concordances, I find that all of them talk about knee problems, knee injuries or knee conditions. So in all these instances, *flared* is used figuratively. As such, we can see that, overwhelmingly, *flared* as verbal has a semantic preference for abstract notions, such as *trouble, violence, temper(s), fighting, anger, row, problem, tensions, dispute, war* etc. We could conclude that in contrast to *flaring* we discussed earlier, *flared* as verbal is more related to unpleasant abstract concepts and has a negative semantic prosody. It seems that the two inflected forms of the lemma *flare* have taken on different roles.

For the nouns, I will investigate *disadvantage/disadvantages* and *opportunity/opportunities* by looking at their collocational profiles in the Bank of English. First, let’s look at *disadvantage/disadvantages*. Their significant collocates are shown in the following table (only the top 24 collocates by t-score are shown here). For the inspection of nouns in this article, the number after the word is the frequency, and the number following the frequency is the t-score.

<table>
<thead>
<tr>
<th>Significant Collocates (by t-score)</th>
<th>Disadvantage/3375 hits</th>
<th>Disadvantages/1912 hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1856</td>
<td>29.096457</td>
</tr>
<tr>
<td>at</td>
<td>1083</td>
<td>28.569265</td>
</tr>
<tr>
<td>is</td>
<td>665</td>
<td>16.289602</td>
</tr>
<tr>
<td>competitive</td>
<td>130</td>
<td>11.303084</td>
</tr>
<tr>
<td>be</td>
<td>330</td>
<td>10.394609</td>
</tr>
<tr>
<td>that</td>
<td>479</td>
<td>9.328815</td>
</tr>
<tr>
<td>compared</td>
<td>80</td>
<td>8.727665</td>
</tr>
<tr>
<td>because</td>
<td>127</td>
<td>8.724275</td>
</tr>
<tr>
<td>advantage</td>
<td>74</td>
<td>8.397821</td>
</tr>
<tr>
<td>put</td>
<td>95</td>
<td>8.380822</td>
</tr>
<tr>
<td>of</td>
<td>923</td>
<td>7.952436</td>
</tr>
<tr>
<td>distinct</td>
<td>61</td>
<td>7.758104</td>
</tr>
<tr>
<td>being</td>
<td>97</td>
<td>7.562197</td>
</tr>
<tr>
<td>the</td>
<td>1779</td>
<td>6.896351</td>
</tr>
<tr>
<td>are</td>
<td>215</td>
<td>6.759874</td>
</tr>
<tr>
<td>major</td>
<td>57</td>
<td>6.640405</td>
</tr>
</tbody>
</table>

Table 1: Collocational Profiles with Frequencies and T-scores of *Disadvantage* and *Disadvantages*
From the above table, it can be easily seen that *disadvantage* and *disadvantages* tend to occur with different group of words. They shared only a few top collocates, such as *are, the, of, social* etc. Of them, the first three are all grammatical words and only *social* is lexical. For *disadvantages*, it is interesting to note that *advantages* tops its collocates, even more significant than the grammatical words like *of, are, and, the, there* etc. *Outweigh* is listed as its second meaningful collocate. *Possible* is also worth special attention. It should be noted that *transactions* and *separate* are invalid and should be removed from the table because they occur in the same sentence repeatedly probably due to some errors in the corpus. For *disadvantage*, it has *competitive, distinct, serious, major* and *big* as its most frequent modifiers. There is little overlap between their top 24 collocates, with *social* as the only exception. Of course, the great majority of the collocates for the singular form *disadvantage* listed above, theoretically speaking, can be used for its plural form *disadvantages*, and vice versa. The difference lies in how frequently they are used with them. Take *outweigh* for example, it co-occurs with *advantages* 81 times, with a very significant t-score of 9.0, but it only co-occurs with *disadvantage* 2 times in the large corpus of 450 millions of words. *Possible* collocates with *disadvantages* 48 out of 1912 times, with a high t-score of 6.4, while it co-occurs with *disadvantage* only 13 out of 3375 times in the corpus. The former is 6 times in percentage more than the latter. Obviously, it has different values for the two inflected forms of the same lemma.

Next, let’s examine *opportunity* and *opportunities*. Their typical collocates (top 24) are shown in the following table:

<table>
<thead>
<tr>
<th>Significant Collocates (by t-score)</th>
<th>Opportunity/51357 hits</th>
<th>Opportunities/26696 hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>to</td>
<td>33580</td>
<td>for</td>
</tr>
<tr>
<td>an</td>
<td>12507</td>
<td>equal</td>
</tr>
<tr>
<td>for</td>
<td>9324</td>
<td>and</td>
</tr>
<tr>
<td>the</td>
<td>31959</td>
<td>new</td>
</tr>
<tr>
<td>have</td>
<td>5001</td>
<td>business</td>
</tr>
<tr>
<td>given</td>
<td>1909</td>
<td>to</td>
</tr>
<tr>
<td>this</td>
<td>3987</td>
<td>are</td>
</tr>
<tr>
<td>every</td>
<td>1518</td>
<td>there</td>
</tr>
<tr>
<td>give</td>
<td>1454</td>
<td>employment</td>
</tr>
<tr>
<td>equal</td>
<td>1084</td>
<td>many</td>
</tr>
</tbody>
</table>

| transactions                      | 5.622567               | separate                 |
| no                                | 6.385679               | being                    |
| serious                           | 6.233326               | overcome                 |
| would                             | 6.140092               | any                      |
| big                               | 6.056221               | however                  |
| competitors                      | 5.991264               | also                     |
| having                            | 5.976879               | social                   |
| social                            | 5.800137               |                         |
From the above table, we see that *opportunity* and *opportunities* do not exhibit completely distinct collocational profiles as they do share some lexical collocates, such as *equal, provide, missed, business*, besides the grammatical words. It should be noted that with *opportunity* we find in its top collocates some modifiers, such as *great, golden*, that have to do with the judgment or evaluation of an opportunity, while for *opportunities* we find *employment, franchise, investment, job, training, career, educational*, which are obviously more related to the classification of opportunities. Hence, they tend to have different semantic preferences by favouring a certain group of words, displaying different semantic prosodies. By ‘semantic prosody’ here, I mean how it is used in real language, in its broad sense because sometimes we can not classify things exactly into three broad categories of positive, negative and neutral. The tendency we would like to describe might mean something more concrete than that, but it can not be labeled as any one of them. For example, we can say that unlike *disadvantage, disadvantages* has a semantic prosody of being more associated with classifying categories of society, neither positive nor negative. It is true that these two forms exhibit some overlap in meaning, for example, by collocating with the same semantic set of verbs, including *provide, miss* etc.. However, they have taken on different roles and are associated with their own “specific meaning dimensions and different connotations” (Tognini-Bonelli,2001:99), which is clearly visible in their collocational profiles. It should be stressed that we are talking about tendencies when we deal with corpus data. The concept of semantic preference or semantic prosody is only a probabilistic concept. What I want to emphasise is that, in line with Sinclair (1987; 1991) and Hoey (2005), common collocates of different word forms should never be assumed; it seems unreasonable to believe that, in all cases, they share the same meaning and only differ in their grammatical forms. On the basis of the previous studies sketched earlier and the new corpus data presented in this paper, we could argue that different word forms of the same lemma appear to favour different collocations, associated with distinct semantic preferences and semantic prosodies and every word form may have its own grammar in these respects.
IV. IMPLICATIONS FOR LEARNER’S LEXICOGRAPHY

The learner’s dictionary plays an important role in helping learners solve lexical problems in the second language (L2) learning process. Ideally, it should include a full-fledged lexicographic description of words to facilitate L2 vocabulary learning. Then what should count as relevant and useful information and be included in the learner’s dictionary? Before answering this question, let’s look at another relevant notion: naturalness, which Sinclair and many other corpus linguists are much concerned with. The traditional division between grammar and lexicon results in the belief that “any combination of lexis and grammar is acceptable” (Hoey& O’Donnell, 2008:294), exemplification in the dictionary is “used only to illustrate the meaning of the word” (ibid). However, corpus linguists are trying to distinguish natural uses from unnatural (but grammatical) uses.

The traditional assumption that different inflected forms of a lemma have the same meaning has been questioned as groundless by corpus linguists. In traditional lexicography, different inflected forms have been conflated into a lemma because they are similar in semantic content. However, corpus evidence shows word forms of a lemma favor different collocates and have distinct uses; although different forms of a lemma may have some overlap in meaning, they are associated with distinct meaning dimensions, and a clear dividing line can be drawn between them in terms of their uses and meanings, i.e. their collocates and semantic associations. For example, flared as verbal is more associated with abstract notions, while flaring as verbal is more likely to be related to concrete things; unlike its singular form, disadvantages favours outweigh very much and tend to collocate with the group of words indicating the classification of opportunities rather than the judgment or evaluation of an opportunity, which is favoured by its singular form; ‘eruptions’ is much more likely to carry meanings associated with volcanoes; across different forms of the lemma, ‘erupt’, there would seem to be a continuum of delexicalisation from ‘eruption(s)’ to ‘erupt(s) to ‘erupted’ (O’Halloran, 2007); faced relates more to the abstract notion, while facing, in the general language, is more associated with position and direction (Tognini-Bonelli, 2001). With such knowledge, SLA learners would be less likely to produce unnatural sentences like “outweigh its disadvantage”. So the natural use of words depends not just on the choice of appropriate words, but equally on the choice of their proper forms. Corpus linguistics tries to show that not every combination of grammar and lexicon is natural and acceptable, but only the naturally recurrent patterns or uses in real language should be “the concern of lexicographers of the contemporary languages” (Sinclair, 2001:44). The main objective of the learner’s dictionary is to tell its users what has been found about natural language.

The differences in meaning and use between inflected forms, i.e. their distinct tendencies in terms of their collocates, semantic preferences and prosodies, have not been mentioned in modern dictionaries. Drawing on the corpus evidence presented earlier, this paper argues that learner’s lexicography should take such information into consideration since an awareness of such differences is more likely to enable L2 learners to produce natural language. If this is not available in the learner’s dictionary it is arguable that some
difficulty may arise in the process of production.

Then the problem begins to arise how such information should be handled in the learner’s dictionary. The solution Hoey & O’Donnell provides when they discuss the relationship between lexis and textual position seems to be also plausible to the problem we are faced with here:

...Sinclair shows that the contexts for interpreting *naked eye* extend throughout the sentence in which the phrase appears and are describable in terms of a number of different relationships that the kernel words form with other choices in the environment (collocation, colligation, and semantic preference), which together have an interpretative outcome (semantic prosody). No dictionary, however well informed, could find convenient ways of communicating such a rich array of lexical relationships, except implicitly by exemplification, which has become a dominant element in the post-Sinclair dictionary (2008: 295).

Sinclair’s discussion mentioned above focuses on the (extended) unit of meaning rather than simple words, though sometimes a word is really a unit of meaning. His notion of the lexical item “represents a further challenge to the autonomy of grammar” (Hoey & O’Donnell, 2008:295) and opens up a new alley to lexical description. Many researchers apply this lexical model, especially the four categories, or some of them, to single words. Clearly lexicographic description of words in terms of collocation, colligation, semantic preference and semantic prosody would be insightful, rich and highly useful for SLA learners, but just as Hoey & O’Donnell said, it is more than difficult to find ways to deal with such rich information in a space-limited print dictionary. The same difficulty applies to the issue we are discussing in this paper. Therefore, one possibility to handle different inflected forms of a lemma in the learner’s dictionary would be to implicitly integrate such information into examples. Of course, guidance on how to read examples should be given in the Guide to the Dictionary. So the purpose of exemplification in the dictionary is not only to illustrate the meaning of the word, but more importantly, to tell the users what collocates or patterns are more significant for or strongly favoured by which inflected form.

Another possibility would be to deal with it in a similar way as some modern learners’ dictionaries do with semantic prosody, starting with some typical words. Such dictionaries explicitly integrate the semantic prosodic information in the definition if the word has a clearly positive or negative semantic prosody. For example, as we mentioned earlier, the Cobuild dictionary (1987) defines *scrawny* as “unpleasantly thin and bony”, where the semantic prosody information of the word is apparent in its definition, rather than “thin and bony”. Only a small number of words have been dealt with in this way because they are the only words which have been found to have significant semantic prosodies. Not all senses of the words in a dictionary can be clearly defined in terms of semantic prosody since word as a basic unit of meaning has been questioned (Sinclair, 1987, 1991,1996; Tognini - Bonelli, 2001;Teubert, 2004, 2005, 2007 etc.). In spite of this,
the inclusion of such information in a dictionary can still be viewed as an important step in modern learner’s lexicography because it offers more useful information about at lease some words, and leads dictionary users to a better understanding and more natural use of these words. More importantly, it has paved the way for the absorption of more such corpus findings by future learners’ dictionaries. These two issues are strikingly similar. The only difference is that the explicit explanation of the differences in meaning between inflected forms should be offered in Usage Notes of a dictionary, rather than definitions.

V. CONCLUSION

Following John Sinclair, many corpus linguists have rejected the traditional grammar/lexicon dichotomy and questioned the traditional assumption that “lemma and inflected forms are bound to share the same meaning and differ only in their grammatical profile” (Tognini-Bonelli, 2001). This study, by looking at the collocational profiles of FLARE, DISADVANTAGE, and OPPORTUNITY in the Bank of English, has confirmed the claim that word forms of a lemma have distinct linguistic behaviour and “form is often in alignment with meaning” (Sinclair, 1991:7). In line with Hoey (2005) and Teubert (2004, 2005, 2007), this paper argues that the meaning of a word (form) is not something that pre-exists somewhere, but the outcome of its uses, or in Hoey’s words, its primings, describable in terms of its collocation, semantic preference and semantic prosody.

For Sinclair, lexicography is “the canonical mode of language description at the lexical level” (Krishnamurthy, 2008). The corpus research findings on the direct correlation between lemma and inflected forms can shed a new light on the description of lexicon, thereby assisting in the development of learners’ dictionaries. Being learning-orientated, such dictionaries should fulfil a valuable pedagogic as well as a descriptive role.

REFERENCES:

AN ANALYSIS OF THE COMPONENT PARTS AND STRUCTURES OF THE 1779 AND 1809 EDITION TAMIL- ENGLISH DICTIONARY

Saravanan Raja
Annamalai University, INDIA
kiranraja_ling@yahoo.co.in

Thomas Malten,
Cologne University, GERMANY
th.malten@uni-koeln.de

Abstract: The first dictionary printed in Tamil and English were compiled by missionaries, which were printed in Vepery, Madras (1779 / 1782). The 1779 edition was compiled by Johann Philipp Fabricius and Johann Breithaupt. It was further revised and enlarged by Carl Pazold and William Simpson which appeared in 1809. The recent research in metalexicography based on the structure of the dictionaries identified by Hausmann and Wiegand (1989) and Wiegand (1989) will be used in comparing these two editions. The component parts of a dictionary or other reference work in terms of its overall design, viz. macrostructure and contents of individual entries, viz. microstructure will be taken into consideration, while comparing FAB1779 and FAB1809. This paper will highlight the different perspectives of metalexicography with reference to component thoughts and the structure of Fabricius Tamil-English dictionaries. The main focus of this present paper will deal with presentation of polysemy, treatment of homonymy (that is selection of lemma), providing lexical description in both of these editions. The study will be quite interesting to check the structural development of the dictionary in respect to learner's point of view.

Key Words: Johann Philipp Fabricius; Tamil-English dictionary; macrostructure; microstructure; metalexicography.

1. INTRODUCTION

Lexicography is an applied field, the theoretical background for which is provided by linguistics. Almost all the branches of linguistics provide information for making a dictionary. It has three major functions viz. store-house of lexical information, retrievable function and the court-house function which is derived from the authenticity of its information. These functions of the dictionary, makes lexicography, an applied field in linguistics. According to Hoeningswald (1967), findings in linguistics can be utilized to solve practical problems in the real world like knowing the spelling of the words, pronunciation, its grammatical category, its meaning, etc. In his point of view, dictionary is an index to the language. The main aim of the present research paper is to compare the component parts and structure of the 1779 and 1809 editions of Fabricius Tamil-English dictionary.

Fabricius Johann Phillip (1711 –1791), linguist and Bible translator of the Danish- Halle Mission in South India after finishing a full course of studies in both law and theology, arrived in South India in 1740 and soon took charge of the Small Tamil Lutheran Congregation at Madras. Fabricius worked on the preparation of Dictionary from 1774-1786, when he was already over 60 years of age. His main interest in preparing this bilingual dictionary was to help the foreigners who came to India for the missionary activities. In 1779, he published the first Tamil-English dictionary (A Dictionary of Malabar and English) containing 9000 words. It was further revised by Carl Pazold and William Simpson which was published in 1809. Later on, E. Schaffer and Pakyam Pillay started the work of enlarging
this dictionary and it was published in 1897 as the “Tranquebar Dictionary” with the help of H. Beisenherz due to the death of E. Schaffer. The revised and expanded dictionary contained 11,590 primitives with 21,300 derivatives and phrases from Fabricius and Pazold’s works which has appeared in three further editions (1910 - Beisenherz, with the assistance of Rev. N. Samuel and Munshi Muthu Thandavaraya Pillay; 1933 - D. Bexell, with the assistance of Purushothama Ayer; and 1972- Carl Gastav Diehl with the assistance of Paavalar V.P.K. Sundaram and the Rev. S. Lawrence) in which idioms and phrases were incorporated and the archaic words which had become obsolete between 1779 and each new edition were eliminated.

2. COMPONENT PARTS AND STRUCTURE OF A DICTIONARY

The word ‘dictionary’ has two meanings in a textural manner i.e. (1) the whole book and (2) the word list, which constitutes the main part of the book. Both of them have two structures, the textural book structure and the textural word list structure respectively. The word has several important units and structures. The basic unit of the dictionary is the treatment unit. The treatment units have a form and information relating to that form is brought together. The relation of form and information is that of topic and comment (Hausmann and Wiegand, 1989).

A form and information relating to a form are brought together under the addressing procedure. Each information item is addressed to a form called address. In any dictionary the most important item is the definition, but there may be hundreds of other information types, i.e., items. The most important address is lemma (headword or entry word), because the lemma belongs to the alphabetical access structure of the dictionary. Normally all the ordered set of lemmata of the dictionary forms the macrostructure. The lemma and the whole set of information items which are addressed to the lemma, form the dictionary article. The macrostructure is the relative arrangement of the stock of lemmata in the word list (cf. Svensen 1993, Bergenholtz 1995). The macrostructure is the overall list structure which allows the compiler and the user to locate information in a reference work. “It is the macrostructure that determines under which lemma the lexicographical item is to be found” (Hausmann and Wiegand 1986). The macrostructure may have a single central word list or additional word lists. The arrangement of words in the macrostructure can be fully alphabetic or systematic (conceptual grouping) or a combination of both. The macrostructure of the dictionary has a central word list. Normally speaking, the structure of information within the article is called the microstructure.

3. STRUCTURE OF FABRICIUS DICTIONARY

Tamil Language is diglossic in nature which means that it has both the high variety (written form) and the low variety (spoken form). While compiling the Tamil English Bilingual Dictionary, Fabricius used both the written and spoken forms in the selection of lemma (headword). For example, the Tamil headword kaal has four different English equivalents (polysemy) which is shown as given below:

\[
\text{kaal} \quad \text{the foot; a brook, channel; the fourth part of a thing; a degree of consanguinity}
\]

In the above entry, the grammatical category like noun, adverb, adjective, etc. has not been mentioned and it has not been practiced throughout the dictionary.

In the same entry, different multi word lexical units or compound forms are given along with their English equivalents. For example,
kutikaal the heel
uLLankaal the sole of a foot
kanuukkaal the ankle
kanaikkkaal the thin bone
mulankaal the knee
munnankaal the fore feet of quadrupods
pinnankaal the hind feet

So far we have discussed the multiword lexical units or compound forms which are derived from the meaning of ‘foot’. Further, Fabricius extended this meaning into idiomatic expressions also.
For example,

kaaluNiniRkiRatu to stand upon your feet
orutan kaalaippiTittu (kaalil viluntu ) veNTukiRatu to supplicate; to fall at a person’s feet

Further, idioms and phrasal verb forms of Tamil language and their equivalents in English have also been explained by Fabricius as given below.

kaalnaTai what goes on foot; cattle
kaalnaTaiyaayp payaNam paNNukiRatu to travel a foot
vicelakkaalile orutanai naluva viTukiRatu to dispatch one with fair words

In the same entry, the free combination form like muTTukooTukkiRakaal which means a prop to support anything has been given as a sub-lemma, but it has to be treated in the main entry itself as the fifth meaning. Likewise, the Tamil subentry orukkaal which means once in English which functions as an adverb has been treated as the derivative of the Tamil lexical item kaal, but it must have been given as a separate entry, as the word orukaal is not the derivative of the Tamil word kaal.

In the Fabricius 1779 edition, the third part of the Tamil lemma kaal which refers to the fourth part of a thing and their derivatives (multiword lexical units or compound words) and idiomatic phrasal verbs are given the sub-lemma status in the same entry itself. For example,

kaal, kaalpanku a quarter
kaalpaTi the quarter of a measure
mukkaal three quarters
araikkkaal half a quarter, an eighth
naalekaal four and a quarter

velai kaalvaaci tiirntatu the fourth part of the work is done

In addition to this, some other multiword lexical item, inflected forms and their derivatives, derived idioms are also given in the same entry as a sub-lemma status. For instance,

kaalmadu the foot of a bedstead
kaalamaaTile at the foot of a bedstead (inflected form)
kaalaalL  a foot-man; one who is very young, having but the fourth part of a complete age

kaalaaLkkaL  infantry

kaalaaNi  a particular beam in the houses of Indians, that leaves for supporting the roof.

kaalaaNiyaRRuppoonavan  one who has become weak or poor (derived idiom)

naaRkaali  a chair

Finally, the fourth part of the meaning of the Tamil lexical item kaal is explained in terms of a degree of consanguinity. It is explained from their different derivatives and phrasal verb form, as follows:

kuntuukaalan (fem. kuntuukaali)  a hobbler
poykkaal  a stilt
poykkaalkaTTiiaaTukiRatu  to go upon stilt (phrasal form with free compound)
kaalvali  a degree of consanguinity (idiom)
muunaRumanaalaankaalile uRamuRaiyaanavarkaL  people of the third kind or fourth degree or remove of consanguinity

Modern dictionaries treat homonyms as a separate entry, but Fabricius (1779) and Pazold (1809) treated them in the same entry as they were of the idea that all derivatives and the phrasal forms and their polysemous meanings should be given in the same entry as a sub-lemma status. Some of their ideas and practices in the dictionary tradition are still maintained by the modern lexicographers. Structurally not much difference exists in the Fabricius (1779) and Pazold (1809) editions. However, diachritic markers have been given in Pazold’s edition and some of the lexical descriptions have been changed. For example,

kakkuvan  a hawking cough (1779)
           whooping cough (1809)

kacakkiRatu  to rub in your hands (1779)
           to rub with the hands (1809)

There are many interesting features in these two dictionaries. The derivative, suffix and phrasal forms of lemma are given. Fabricius has mainly concentrated on proper names. For example,

kaTTaaNi  an avaricious person
kaNakkaN  an accountant

kaNitaakkaaran  an astrologer
kacaTan (fem. KacaTi)  a bad person

kaNNATiyan  one of the Cannary – country, a cannariah
Fabricius has used very simple, elegant, lexical descriptions for better understanding of the learners of Tamil and English Languages.

\textit{kaLam}, a place in the field; a thresh-floor; shelves of the land, banks under water (1779)

\textit{kaLam}, a place in the field; a threshing – floor; shelves, banks under water (1809)

Fabricius has used strong verb first and weak form next in the entry. For example,

\textit{kaNaikkiRatu} to neigh, to retch or hawk
\textit{kaNaippu} neighing, hawking, retching, kecking

When compared to the fourth edition compiled by Carl Gastav Diehl (1972) with the assistance of Paavalar V.P.K. Sundaram and Rev. S. Lawrence, there has been a remarkable change in the structure of the dictionary. Grammatical categories, polysemous and homonymous meanings were clearly explained. Headwords and their derivative forms have been given in bold letters and the printing is also very elegant.

In the Fabricius 1779 edition, the Tamil headword \textit{kaal} and their derivatives have been given in the same entry itself, but in the fourth edition, it has been shown as homonymous form (given as two entries). The headword \textit{kaal} has seventeen polysemous meanings, along with their Tamil forms and their English equivalents as given below.

\textit{kaaI, s. leg, foot, paatam} 2. post, \textit{tuuN}; 3. support, prop, taankukaal;
4. the lower part, base, \textit{aTippakkam}; 5. the wheel of a car, terkkaal;
6. the spokes of a wheel, aarakkaal; 7. brook, channel, vaaykkaal;
8. the fourth part of a unit or thing, a quarter, \textit{kaaRpanku};
9. a degree of consanguinity, \textit{irattak kalappaamuRavu};
10. way, \textit{valli}; 11. a turn, time (as \textit{orukaal}, once);
16. a forest; 17. (gram.) \textit{vainaiyecca vikuti} for \textit{kaalam} if, provided
(as \textit{vantaakkaal, vantaakkaal}, if he should come).

In addition to the polysemous meanings, seventy one derivatives (compound forms), idiomatic phrasal expressions and their English lexical descriptions have also been given. Some of the examples are given below.

\textit{kaalaNi} ornament for the feet in general
\textit{kaalkaTTu} inseparableness (as of wife and husband). \textit{kaaRkaTTu}
\textit{kaalnaTai, kaanaTai} walking, going on foot; cattle, sheep, goats and oxen
\textit{kaalnaTaiyaaypp payaNam paNNa} to travel on foot [lame man]
\textit{kaalvay} a canal, a channel
\textit{mukkaali} a tripod
4. CONCLUSION

The Fabricius (1779, 1809, 1972) dictionaries were compared and it has been observed that some structural differences exist among them. The microstructure and macrostructure concept of Hausmann and Wiegand (1986) has been observed in these dictionaries. The study will pave a way for checking the structural development of the Tamil-English dictionary in respect to the learner's point of view. Such type of studies will help in compilation of dictionaries and form a new branch in the field of lexicography which can be termed as Missionary Lexicography.

REFERENCES


John Benjamins.


Malkiel, Yakov (1967) “A typological classification of dictionaries on the basis of the distinctive features” in Householder and Saporta (eds), Problems in Lexicography, pp. 3-24.
I trace the change in the dictionary use (preferences of genres, media, and titles) by Japanese college students of English over the past ten years on the basis of the questionnaires conducted at the beginning of each academic year. The questionnaires led to some interesting findings: the hand-held electronic dictionary overtook the print dictionary in terms of number of users in the 21st century and has been enjoying enormous popularity; Japanese-English dictionaries, monolingual English ones and those on CD-ROMs and the web are not used much; synonym and collocation dictionaries appear to be even less relevant for the students. A separate recent questionnaire has shown that the majority of students may not be able to take full advantage of the potential offered by hand-held electronic dictionaries since they are not familiar with the range of sophisticated functions these machines offer. Instruction is needed to familiarize the students with the models commercially available, their main functions and operation. Students also need to be made aware of the shortcomings of the electronic dictionaries.
A CORPUS-BASED ANALYSIS OF BUSINESS COMPUTER TERMS

Rinruedee Pattaradej
Suan Dusit Rajabhat University (Suphanburi Campus), THAILAND
iamrindy@gmail.com

Abstract: At present, it is widely accepted among the English language teachers that computers have been used in vocabulary studies and in lexicography for several decades. This paper focuses on a corpus-based research project for Business Computer students at Suan Dusit Rajabhat University, Suphanburi Campus. This paper, thus, aims to study the characteristics of vocabularies and special terms in Business Computer field. The data used for this study were the textbooks collected from Business Computer curriculum that is composed of general business and computer subjects from the Faculty of Management Science together with interviewing lecturers from Business Computer program. This research used the analysis tool which is a concordancing software and based on Biber et al. (1999), Coxhead (2000) and Nation (2001) to find out: (1) word frequency, (2) two major of word classes: lexical words and functional words, (3) collocation of terms from concordance, and (4) academic vocabulary and technical vocabulary. This study can be a guideline for lecturers in teaching the vocabulary for the students. Besides, the corpus-based data will be used to compile a dictionary of technical terms for Business Computer and English teaching materials.

Key Words: Business Computer, Special Terms, Textbooks, Curriculum

1. Introduction

Nowadays, English language is imperative of vital importance in many countries. It can be said that English language is able to be global language. As stated by Crystal (1997), approximately 375 million people speak English as their first language and 235 million people speak English as their second language. In addition, many countries taught English as a foreign language. In Thailand, English language is one of the most essential foreign languages. The role of English language in Thailand is quite important as it is in many other developing countries. New technology and adoption of the internet have resulted in a major transition in terms of business, education, science, and technological progress, all of which demand proficiency in English (Wiriyachitra, 2009: Online). As the growth of both international business and technology in Thailand, it is necessary to follow the changing world of business and technology. Regarded with Veerachaisantikul (2007:6), English is acquired as international language of technology and commerce because it is suitable for learners who want to know English for particular purposes. English for Specific Purpose is the approach of second language teaching and learning that relates to multi-disciplinary activities of particular learners.

The specific purpose learners, thus, are expected to have high quality both in English language and in specific filed, particularly technology filed as computer science and business computer. It is necessary that they have to know the vocabulary of their specific field to help their reading and writing texts. The acquisition of vocabulary has long been considered to be a crucial component of learning a language because the breadth and depth of a student’s vocabulary will have direct influence upon the descriptiveness, accuracy and quality of his or her writing (Coady et al. 1993, Nation 2001, and Read 1998). This research studied the Business Computer Terms: Lexical
Words and Function Words, Collocation Terms, Academic Vocabulary and Technical Vocabulary by using computer tool called a concordancing program.

1.1 Lexical Words and Function Words
In accordance with Biber et al (1999), lexical words are the main carries of meaning in a text. They are numerous and are number of open classes: nouns, verbs, adjectives, and adverbs. While lexical words are the main building blocks of texts, function words provide the mortar which binds the text together. They often have a wide range of meaning and serve two major roles: indicating relationship between lexical words or larger units, or indicating the way in which a lexical word or larger unit is to be interpreted. Function words are members of closed systems--auxiliaries, conjunctions, and prepositions.

1.2 Collocation Terms
Nation (2001:317) said that the term “collocation” is used to refer to a group of words that belong together, either because they commonly occur together like task a chance, or because the meaning of the group is not obvious from the meaning of the parts, as with by the way or to take someone in (trick them). Whilst, Biber et al. (1999:59) mentioned that collocations consist of independent words that tend to co-occur. Nesselhauf (2005:1) described that collocations are one type of a group of expressions whose importance in language has been increasingly recognized in recent years. This group of expressions has been variously called prefabricated units, prefabs, phraseological units, (lexical) chunks, multi-word units, or formulaic sequences. In addition, collocations can be either grammatical or lexical. In grammatical collocation, it is a phrase consisting of a dominant word (noun, adjective, and verb) and a preposition or grammatical structure such as an infinitive or clauses. In contrast to grammatical collocations, lexical collocations normally do not contain propositions, infinitives or clauses. Typical lexical collocations consist of nouns, adjectives, verbs, and adverbs (Benson et al. 1986).

1.3 Vocabulary Levels
Nation (2001) divided vocabulary into four levels: high frequency words, academic vocabulary, technical vocabulary, and low frequency words. However, the two levels that the ESP learners should pay particular attention are academic vocabulary and technical vocabulary. In accordance with Nation and Waring (1997 quoted in Wang, J. et al. 2008), it is generally agreed that the beginners of English learning should focus on the first 2000 most frequently occurring word families of English in General Service List (GSL) (West, 1953), while for intermediate or advanced learners who usually study English for Academic Purposes, the command of these GSL words may no longer be their major concern and priority of their vocabulary acquisition may be shifted to lower frequency vocabulary.

1.3.1 Academic Vocabulary
Academic vocabulary that can be called sub-technical vocabulary (Cowan, 1974) or semi-technical vocabulary (Farrell, 1990), is a very important specialized vocabulary with a high frequency and/or wide range of occurrence for second language learners intending to do academic study in English (Nation, 2001:17 and Farrell, 1990:11). The Academic Word List compiled by Coxhead (2000) and consisted of 570 word families is the great example because they are not in the most frequent 2,000 words of English but they covered 10% of the tokens in 3,500,000 running word academic corpus.
1.3.2 Technical Vocabulary

Technical Vocabulary is vocabulary that is recognizably specific to a particular topic, field, or discipline and useful for specific area of knowledge. It would not be occurred in GSL, function words, and AWL.

1.4 Previous Studies on Corpus Analysis

Previous studies on corpus analysis have produced some benefits of this study. There have been many research studies on corpus both in abroad (Wang, J. et al. 2008 and Verlinde and Selva 2009) and Thailand (Kaewphanngam, C. 2002, and Veerachaisantikul, A. 2007)

Wang, J. et al. (2008) reported a corpus-based lexical study of the most frequently used medical academic vocabulary in medical research articles (RAs). A Medical Academic Word List, a wordlist of the most frequency used medical academic words in medical RAs, was compiled form a corpus containing 1,093,011 running words of medical RAs from online. The established MAWL contains 623 word families, which accounts for 12.24 % of the tokens in the medical RAs under study. The high word frequency and the wide text coverage of medical academic vocabulary throughout medical RAs confirm that medical academic vocabulary plays an important role in medical RAs.

Verlinde and Selva (2009) investigated a selection of words to be contained in a learners’ dictionary of French. This study was a comparison study that used lexicographic approaches: corpus-based and intuition-based. In the study, the corpus was taken from two newspaper issues of Le Monde and Le Soir that were publish in 1998. The corpus had total size of 54,260,926 words. The result of the study indicated that a corpus-based approach provided a practical method for the lexicographer’s personal intuition.

In Thailand, Kaewphanngam, C. (2002) analysed a corpus of psychology texts in order to develop the teaching materials in English for Academic Purposes. A corpus of psychology texts was selected according to comprehensive criteria. The total numbers of running words were 236,086 words. The psychology students' comprehension of technical and sub-technical vocabulary was investigated by a vocabulary test based on the corpus analysis. The subjects were first year and third year psychology students attending the Faculty of Education, Silpakorn University, in the 2002 academic year. Results revealed that most of the frequent words were function words and content words which were highly topic-related words to the subject of psychology. The parts of speech, uses and meanings of the search words could be investigated by applying concordance analysis. The vocabulary test showed that psychology students performed better in sub-technical vocabulary than in technical vocabulary. First year and third year students obtained significantly different scores on the whole test. There was a high correlation coefficient of technical and sub-technical vocabulary.

Veerachaisantikul, A. 2007 studied a business corpus of multi-words expressions in the newspaper. The subjects were news articles in business from the Bangkok Post’s business section collected for 3 months from 1st August to 31st October 2006. The total number of news article were 167 files and the total number of tokens or running words were 172,349 words. The results showed that the most frequency occurring words were function words and business key words, which were related to the subject of business. A concordance program is a convenient and useful tool for the parts of speech, uses, and the meaning of the selected word investigation. The most frequent parts of speech collocated with the selected business key words were “article”, “adjective”, and “noun”. The most frequent used pattern of multi-word expressions in the Business corpus was in the form of “adjective + noun”.

2. Methodology
2.1 Data Collection
The sample used from the English textbooks and course materials collected from the 2008 Academic year of Business Computer curriculum that is composed of only subjects from the Faculty of Management Science together with interviewing 5 lecturers from Business Computer program in Suphanburi Campus.

In the English textbooks and course materials of Business Computer, there were 17 textbooks and course materials as follows:
2. The LINUX Programmer’s Toolbox--John Fusco, 2007
3. A Programmer’s Introduction to PHP 4.0--W.J. Gilmore, 2001
4. Practical PHP and my SQL--Jono Bacon, 2007
5. ASP.NET 3.5 for Dummies--Ken Cox, 2008
10. ASP Configuration--Gary Palmtier, 2001
11. Database Management Systems--Raghu Ramakrishnan and Johannes Gehrke, 1999
12. Database System Concept--Foxit, 2004
14. Database System--Paul Beynon-Davies, 2004

2.2 Data Processing
In the research, data processing incorporated the plain text files of English Business Computer textbooks and course materials to be stored in corpus. Then the concordancing program called WordSmith Tools was used to provide specific text analysis tasks.

2.3 Data Analysis
There were four steps in this study. Firstly, the Lexical words and function words were analysed from wordlists and concordancing line to find out parts of speech based on Biber et al.(1999). Secondly, collocations were analysed in term of cluster. Thirdly, academic vocabulary that were found in Academic Word List complied by Coxhead (2000) were comprised. Lastly, technical vocabulary also were analysed by checking from “Oxford Advanced Learner’s Dictionary (2005)”, “Oxford Dictionary of Computing for Learners of English (1996)”, “The Oxford Interactive Dictionary of Business and Computing for Learners of English(1998)”, “Oxford Dictionary of Business English for Learners of English (2000)” and “IBM Dictionary of Computing Online (2009).”
3. Result.

3.1 Corpus Finding

Figure 1: Statistical Details of the Business Computer Corpus

From the figure above, there were 2,857,687 tokens or running words, 40,116 word types or different words in a corpus since a recurrent word was counted only once. With a type/token ratio of type/token ratio of 1.40 or 1:71.24, each word in the corpus was repeated nearly 71 times on average throughout the corpus.

Within the Business Computer Corpus, the ten most frequency occurring tokens account for 20.9% of the total list. These tokens were the, to, of, a, and, in, is, that, for, and this (see Figure 2 below).

Figure 2: Top Twenty Wordlist in Business Computer Corpus

As can be seen in Figure 2, the top fourteen tokens were function words or closed classes. The word “the” appeared the most frequently with 169,451 times or 5.93%. The first lexical word or open class was “data” which come out at the 15th rank.
3.2 Lexical Features

Table 1: Top 40 Nouns in Business Computer Corpus

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Table</td>
<td>15. User</td>
<td>25. Key</td>
<td>35. Form</td>
</tr>
<tr>
<td>7. Example</td>
<td>17. Figure</td>
<td>27. Application</td>
<td>37. End</td>
</tr>
</tbody>
</table>

From the Table 1, it was interesting that the word “business” is in at the rank of 30th in top noun or at the 97th rank with 3,310 times in the whole corpus. This can be shown that computer is an important tool for business.

3.3 Collocation Selected from Business Computer Corpus

In this section, two words from the top 40 nouns (Data and Business) were selected to find the collocation of grammatical and lexical collocations to show the importance of computer through business.

Table 2: The Most Frequent Lexical and Grammatical Collocation of Data and Business in Business Computer Corpus

<table>
<thead>
<tr>
<th>Type</th>
<th>Lexical Collocation (Freq.)</th>
<th>Grammatical Collocation (Freq.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>relational data model (193)</td>
<td>the data in (135)</td>
</tr>
<tr>
<td>Business</td>
<td>e-business model (226)</td>
<td>type of business (230)</td>
</tr>
</tbody>
</table>

Concordance data from the Business Computer displayed that distinct relationship of the words or phrases which were frequently used with another word or phrase e.g. relational data model (Freq. 193) and the data in (Freq.135). Consider the example of concordances below

Figure 3: The Concordance Data for the Word “Data”
The concordance illustrated how nouns as “data” connected to other words in Business Computer Corpus to create the noun phrases “the data in.” The phrase “the data in” was often followed by noun or group of word such as “a regular heap.”

Whist, the Figure 4 showed the circumstance of the word “business” as “e-business model” and “type of business” which were preceded by verb “store” and determiner “this.”

![Concordance Data for the Word “Business”](image)

**3.4 Academic and Technical Vocabulary**

According to Academic Word List compiled by Coxhead (2000), the word “data”, the first lexical word and the first noun in the whole corpus, was in Sublist 1 of AWL. Let us look the top ten AWL in Business Computer Corpus from the table below.

<table>
<thead>
<tr>
<th>Type</th>
<th>Frequency</th>
<th>Rank</th>
<th>Sublist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>16,767</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>Chapter</td>
<td>5,213</td>
<td>51</td>
<td>2</td>
</tr>
<tr>
<td>Network</td>
<td>4,597</td>
<td>65</td>
<td>5</td>
</tr>
<tr>
<td>Access</td>
<td>3,701</td>
<td>85</td>
<td>4</td>
</tr>
<tr>
<td>Function</td>
<td>3,555</td>
<td>88</td>
<td>1</td>
</tr>
<tr>
<td>Process</td>
<td>3,217</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>Create</td>
<td>2,906</td>
<td>117</td>
<td>1</td>
</tr>
<tr>
<td>Select</td>
<td>2,854</td>
<td>118</td>
<td>2</td>
</tr>
<tr>
<td>Section</td>
<td>2,707</td>
<td>125</td>
<td>1</td>
</tr>
<tr>
<td>Design</td>
<td>2,470</td>
<td>134</td>
<td>2</td>
</tr>
</tbody>
</table>

As can be seen from the Table 3, the sublist1 appeared the most (five word types with 29,152 times): data, function, process, create, and section.

For the technical vocabulary, SQL which is abbreviation in type of initialisms from the full word “Structured Query Language” appeared the fist with 7,544 times at the 30th rank.
Table 4: Top Ten Technical Vocabulary

<table>
<thead>
<tr>
<th>Type</th>
<th>Frequency</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQL</td>
<td>7,544</td>
<td>30</td>
</tr>
<tr>
<td>PHP</td>
<td>4,582</td>
<td>67</td>
</tr>
<tr>
<td>C (C language)</td>
<td>2,939</td>
<td>113</td>
</tr>
<tr>
<td>XML</td>
<td>2,773</td>
<td>120</td>
</tr>
<tr>
<td>DMBS</td>
<td>2,368</td>
<td>143</td>
</tr>
<tr>
<td>Schema</td>
<td>1,812</td>
<td>192</td>
</tr>
<tr>
<td>IP</td>
<td>1,758</td>
<td>197</td>
</tr>
<tr>
<td>Router</td>
<td>1,740</td>
<td>201</td>
</tr>
<tr>
<td>ASP</td>
<td>1,604</td>
<td>224</td>
</tr>
<tr>
<td>HTTP</td>
<td>1,207</td>
<td>313</td>
</tr>
</tbody>
</table>

From above Table, it can be seen that most of top ten technical words are abbreviation. The processes of abbreviation are reduction, which create numerous types of shortened forms in speech or writing of English. Many mechanisms of abbreviation in English are used to create a wide variety of forms (Barnhart, 1995: xiii). The abbreviations are classified as acronyms, initialisms, clippings, apheresises, contraction, and symbols. For the eight abbreviations from the top ten were classified as initialisms.

4. Conclusion

The Business Computer Terms based on Business Computer Corpus with 2,857,687 tokens or running word have been complied for the better learning and application of Business Computer students. Although, this study has not completed and it is still ongoing, the researcher hope the development of the corpus to be the Business Computer lexicography and the availability of textbook for Thai students who study in this field.

References:


Abstract: The need to learn the most important lingua franca, which undoubtedly is English in the contemporary world and to have the desired communicative competence in this language has inspired many in the field of language pedagogy to come up with all kinds of improvements related to linguistic theory as well as material production in this regard. Thus a wide range of pedagogical materials to develop communicative competence of the learner at various stages is currently available on the market. Yet the feeling strongly exits that there is still a gap which needs to be bridged to facilitate the learner to have proficiency in the foreign language at optimal level. The aim of this paper is to suggest a new pedagogical tool in the form of a dictionary of ideas and communication and describe the salient features of such a reference book which is expected to help the learner to develop his communicative competence in the most comprehensive way. The dictionary as such will provide almost all the basic ideas about the social, cultural and material dynamics of life required for learners at tertiary as well as advanced levels via the specific social domains, speech events and speech acts organized in the dictionary and be the most comprehensive pedagogical tool ever prepared for the learner to use efficiently and independently.

1. Introduction
As one of the most fundamental defining characteristics of homo sapiens, competence in communication is vital to leading a practically successful life, especially in the new millennium in which the world seems to have shrunk into what we proclaim in common parlance as “global village” making the dynamics of communication highly complex and really challenging to meet in a realistic way. The situation per se has triggered off an increased interest in investigating the whole process of communication comprehensively at various levels and in varied contexts. However, it is no longer the specialists like linguists, communication engineers, rhetoricians and language teachers, who have looked at communication in their own special perspectives, the common learners have equally demonstrated their appreciation of this phenomenon and evinced renewed interest in acquiring competence in communication at optimal level in almost all social contexts. To acquire competence in communication in a holistic way as such is not restricted to one’s L1; it is required rather more so in learning one’s second/foreign language which in contemporary times happens to be the English Language, one of the most pragmatically important and internationally major languages in the world.

2. Communicative Competence in FL Learning
Emerging essentially as a reaction to Chomskyan theory of linguistic competence, Hymes (1970) postulated a revised notion of language perceived in term of communicative competence, incorporating not just the knowledge of the rules of syntax and phonetics but also the ability to use language in conformity with the social and cultural dimensions of language implicit in communication in general. In contrast to Hymes’ original formulation of the notion of communicative competence as (a) knowledge of whether something is formally
possible, (b) whether something is feasible, (c) whether something is appropriate in relation to its actual context and (d) whether something is actually performed. Canale and Swain (1980) further expanded the sense of communicative competence to include (a) grammatical competence, (b) socio-linguistic competence, (c) discourse competence and (d) strategic competence as the most essential components of this linguistic ability. The conception of language as such provided, of course, not only a new dimension to linguistic theory but also proved seminal for reorienting strategies in the entire field of second/foreign language pedagogy.

Savignon (1983) had serious doubts about the validity of the audio-lingual method in foreign language teaching. Elucidating the basic tenets of the communicative approach to FL teaching, he emphasized that (a) language should be looked at as something creative, (b) the needs and interests of the learner should be given priority, (c) importance should be given to the expression of meaning, (d) correction of errors should be considered secondary, (e) teaching should be learner-centered and (f) learners should be encouraged to take part in different kinds of communicative activities (Cook, 1988). But the communicative approach to second / foreign language teaching gained great momentum when real efforts were made in the areas of course descriptions, curriculum guidelines, adopting communicative methodology and developing communicative testing materials etc. In spite of all this, there are experts who are not fully convinced of the effectiveness of the communicative approach to the language teaching. Yet the need to have the required communicative competence in second / foreign language is still valid and requires further research particularly in the area of producing pedagogically-oriented new reference works considered crucial for foreign learners to develop their communicative competence independently and holistically.

3. Pedagogical Aids to Developing Communicative Competence
In view of the growing demand for achieving maximal ability in second / foreign language communication, a wide variety of pedagogical materials has been developed and produced to serve the needs of various categories of learners across the world. The learner’s dictionaries aim to facilitate the potential users not only in the comprehension but also in the production of L2 discourse by providing extensive verbal illustration in their micro-structures. Yet these dictionaries are restricted to developing the learner’s lexical competence in the second / foreign language. Encyclopedias present specialized factual information of advanced level, not much required for communication about general aspects of life. Encyclopedic dictionaries, although a fine mix of the two, lack in conceptual information needed for the learner. An ESP dictionary is again a tool to be used for experts in a particular field of knowledge, least useful for a common learner of second / foreign language. The use of Almanacs too is restricted to finding out factual information about various branches of knowledge. Wilkin’s (1983) Notional / functional syllabuses having applications in short term ESP programs hardly facilitate the learners in the generative use of language. Browsing through newspapers, magazines, internet or reading specialized books on sociology, psychology, arts and culture for gleaning general knowledge could be off putting for lack of easy access to information. Similarly, watching TV programs to imbibe general knowledge on a wide spectrum of themes can hardly be expected to be much rewarding each time one is exposed to these media.

4. Dictionary of Ideas and Communication
The notion of compiling a Dictionary of Ideas and Communication (hereafter, DICOM) emanates essentially from a realization of the needs of the learner to facilitate him/her in his/her second / foreign language communication in the best possible way. The idea as such
is based on the premise that communication in general is a reflection of one’s linguistic and cognitive development simultaneously. The DICOM aims to blend the two simultaneously, which does not seem to be the case in other reference books followed as aids to developing communication in real life situations. The dictionary under considerations is geared to generating communication activities via the following format embedded in the dictionary.

1. Introduction to the theme
2. Points to ponder
3. Questions on the theme
4. Activities on the theme
5. Model answers to the questions

An “introduction to the theme”, which will be an aspect of social life or any phenomenon which could potentially have any impact on social life in general, will be described concisely in one or two paragraphs, covering all its significant aspects in diachronic as well as cross-cultural contexts. Initial information about the theme as such is expected to make the learner feel confident about communication on the theme at least in its basic form. The “points to ponder” are aimed at developing in the learner an added measure of critical thinking. The learner at this stage may draw on his / her perspective to add to discussion on the theme. If followed in the class, the teacher can ask the learners to work out complete questions as the “points to ponder” are in the form of phrases, rather them complete questions. “Questions on the theme” are supposed to be sufficient enough to lend a broader perspective to the theme, sensitizing the learner to all its subtleties and complexities. The learner’s attempts to answer these questions can be matched with the answers modeled in the dictionary. The teacher can work out various communicative activities based on the theme in the form of dialogues, speeches, discussions and debates in the language class. It will make the task of the teacher relatively demanding, yet it is highly rewarding for the learner to improve his communication and knowledge about the theme. “Model answer(s) to the question” can be one or multiple depending, of course, on the nature of the theme under discussion. In case of multiple answers, special attention is to be given to different levels of learning, for instance, tertiary or advanced. A cross-cultural and ideological perspective is yet another feature which the answers will address in order to make the learner aware of the global perceptions on the theme.

5. Structure of the Proposed DICOM
The structure of any reference book depends essentially on the nature of reference material it chooses to deal with. The proposed DICOM is a new kind of dictionary aiming to provide conceptual knowledge with regard to the social, cultural and material aspects of life and, thus help develop the communicative skills of the learner, learning English as a second/foreign language in their respective national contexts. The conceptual knowledge as such is to be represented into discrete themes like ‘teacher’, ‘eating out’, ‘dress’ belonging in different domains of human life e.g. ‘education’, ‘enjoyment’ and ‘culture’. On the one hand, therefore the DICOM will follow the conceptual method of organizing the contents into different domains, and on the other, pursue the alphabetical order in the sequence of letters, in addition to providing all of this information in the concordance.

6. Functions of the DICOM
As a new kind of pedagogical tool, the DICOM focuses on developing the learner’s conceptual and communicative skills in the following way:

1. make the learner a keen observer of even the trivial things happening around
2. develop the learner’s critical thinking
3. enhance creative abilities of the learner
4. sensitize the learner to global perceptions on issues and problems
5. make him/her aware of inter-cultural perspectives on various matters of life
6. inculcate a humanistic approach to life in general
7. make him/her appreciate the diversity of views
8. facilitate as a reference tool for communication in general
9. improve his/her linguistic skills
10. make him/her feel confident in communication at international level

7. Scope of the DICOM
In view of the presumed academic appeal in the educational context, the DICOM can potentially attract a large variety of users. It can be used in the class as a supplement to language teaching by the teacher. All kinds of learners can follow it independently according to their needs and demands. It can also be highly effective for preparation in all kinds of competitive exams within one’s country, examinations at international levels, and tests like TOEFL and IELTS etc where speaking module stands out quite prominently. Taking into consideration its scope as such a good number of publishers within individual countries as well as abroad are expected to be willing to accept this project for research and publication. In this way, the DICOM could prove yet another entry into ever growing community of reference books and add to the “dictionary boom” in the lexicographical world.

8. Conclusion
The DICOM is, of course, a new idea proposed to meet the potential communicative and conceptual needs of the learner at tertiary and advanced levels in the form of a reference book. Presuming the number and categories of its users, it seems to have a reliable measure of commercial viability. If compiled in the respective L1 contexts, it could be more useful than the one which is global in its readership. However, in any form, it can be a highly useful new pedagogical tool in the field of second/foreign language learning pedagogy.

References


DESIGN DIGITAL MANUSCRIPT: REFINING A DESIGN IDENTITY FROM A VISUAL AND WORD ASSOCIATION

Apisak Sindhuphak
King Mongkut Institute of Technology Ladkrabang, THAILAND
tculture@kmitl.ac.th

Visual and word associations have been an intricate part of a design process. Prior to this view, element of design processes such as choice of materials, the intended functions, and the manufacturing processes, have shown considerable factors in helping to conclude the processes into products or a design solution. However, to justify the status of products during the changes in trends and the shifting of a competitive market, a consideration that connect a symbolic meaning to a design have also been taken into an account. The selection of this area cans response to the changes and to perhaps represent a sole identity in forming its own style from contents that genuinely reflected the intended concepts. This paper is intended to define the significant of “visualization concept” based on a valid connection of words and visual images. Using the methods of collage and photomontage collection, a digital manuscript is build from undergraduate and graduate curriculum in Architectural Education and Design for King Mongkut Institute of Technology Ladkrabang (KMITL). Using “Human and Design” and “Industrial Design VI” as courses focused on current design issues, methods in gathering visual resources of collages or photomontages is selected to find the associations. After an evaluation, work with visual and word association is created into a manuscript, which can then be accessible from either a digital or hard copies for design and educational purposes. This also represents a method of refining and revealing the intended design concepts under the ages of sustainable digital technology.

Key Words: visual and word association; digital manuscript; collage; photomontage; design identity
THAI VOCABULARIES FOR AAC DEVICES

Monthika Boriboon,
Sarinya Chompoobutr
National Electronics, and Computer Technology Center, NSTDA, THAILAND
monthika.boriboon@nectec.or.th, sarinya.chompoobutr@nectec.or.th

Puttachart Potibal
Kasetsart University, THAILAND
puttachart.p@ku.ac.th

This paper is an attempt to demonstrate the process of obtaining Thai vocabularies which will be applicable for the augmentative and alternative communication (AAC) devices for Thais. They are collected from several Thai language sources, namely, Thai Dictionary of Royal Institute of Thailand (1982), vocabularies for preschool and elementary (grade 1-3) compiled by Ministry of Education (1988), and Thai thesaurus ‘Khlang Kham’ of Nawawan Bhandhumedha (2001). To find the common word list, the 2,646 lexical items are checked for their frequency occurrences. The BEST Corpus (5 million words) and the newspaper corpus (10 million words) are used for frequency count. The vocabularies are subcategorized into various groups according to their semantic and syntactic features. The subcategories will be conformed to the concept of Semantic Compaction. They cover more than 80% of daily speaking.

**Key Words**: Thai vocabulary, common wordlist, frequency, corpus
CULTURAL VOCABULARY IN PROENCA’S TAMIL-PORTUGUESE
DICTIONARY OF 1679 AND FABRICIUS’ TAMIL-ENGLISH DICTIONARY
OF 1779 - A COMPARATIVE STUDY

Thomas Malten
Cologne University, GERMANY
th.malten@uni-koeln.de

S. Raja
Annamalai University, INDIA
kiranraja_ling@yahoo.co.in

Early bilingual Tamil-Portuguese and Tamil-English dictionaries are targeted towards European language learners and have more or less elaborate glosses, in the target language, for cultural words. At the end of the 18th century the secularization process and a hundred years of scientific progress/research have influenced dictionary making by inducing a more scientific or historical approach in defining words. At the same time the target audience has shifted from purely missionary to a commercial and administrative one, and English is on the verge of supplanting Portuguese as the trade language in Asia, foreshadowing its world-wide domination in the 20th century. The paper will focus on the structural difference between Proenca’s 1679 and Fabricius’ 1779 Tamil-English dictionaries, viz. sorting sequences, standardization of written Tamil, and treatment of grammatical categories. This can be exemplified by the Tamil words of kinship, religion, the material world and games to mention a few subjects to be elaborated in the paper. The study is based on the ongoing new editions of Proenca’s and Fabricius’ dictionaries undertaken jointly by University of Cologne Germany and Annamalai University, India.
DESIGN FEATURES OF BI-LINGUAL DICTIONARIES IN PUNJABI: A FAR CRY FROM THE MODERN LEXICOGRAPHIC PRACTICE

Masroor Sibtain
BZ University, PAKISTAN
masroorsibtain@yahoo.com

Abstract: Role of bi-lingual dictionary as a reliable resource in learning a second language has assumed enormous significance in the age of globalization. Punjabi, spoken and understood from the length and breadth of Pakistan, is one of major languages. Bi-lingual dictionary takes into account learner’s receptive as well as productive needs as it caters for the user’s information about lexis, their pronunciation, syntax, and meaning but also the instances for usage, the informational gap between L1 and L2 and cultural background. As far as cultural differences are concerned dictionary compiler needs to be in the know of the needs, productive as well as receptive, of the target users. Design features of bi-lingual dictionary need to be in consonance with the modern lexicographical principles and practice in order to meet second language learners’ needs. The paper presents the findings of a study conducted in order to sketch the design features of existing bi-lingual Punjabi dictionaries in Pakistan. Two dictionaries namely, The Vanguard Punjabi-English Dictionary and W.P. Hares English-Punjabi dictionary are taken as a sample of the study. To gain the objective of the study, approach of documentary analysis has been adopted. A check-list, based on the design features in Oxford English into Urdu Dictionary, Qaumi English into Urdu Dictionary and description of structural features by Hartmann (2001), has been prepared. The study shows that the existing bi-lingual Punjabi dictionaries need to be structured in consonance with the modern postulates and practice of lexicography.

Key Words: Bi-lingual dictionary, receptive and productive needs, design features, documentary analysis, lexicography

1. INTRODUCTION

The history of bilingual dictionary is as old as man’s first journey outside his territory to explore new lands inhibited by people speaking different languages and participating in totally diverse cultural and economic activities as this pursuit could be fruitful if the hindrance of communication with others in their language was possible. With the advent and increase in printing, the lexicographic activity manifested expediency in the backdrop of intercontinental travelling. The paper seeks to evaluate the impact of lexicographic development in design features of bilingual dictionary making and its relativity with one of the most frequently spoken languages in Pakistan—the Punjabi language. A dictionary is a most formidable language learning resource; its design features determine the effectiveness of a reference work as they refer to the overall principles which govern the production of efficient reference works. Al-Ajmi (2002) appears to be suggesting that the way of organizing information in a dictionary entry plays a very crucial role in the dictionary use and researchers should divert their attention towards macro as well as the microstructures in bilingual dictionaries in order to enhance the usability of bilingual dictionaries. Landau (2001) pointed out:
A bilingual dictionary consists of a list of words and expressions, in alphabetical order when in printed form, in one language (the source language), for which ideally, exact equivalents are given in another language (the target language). The purpose is to provide help to someone who understands one language but not the other. More, the presumption is that one of the languages is the user’s native language (p. 8).

So a bilingual dictionary represents two languages and it normally serves as a guide in translating from one into the other, or producing texts in language other than the user’s mother tongue (Zgusta: 1971, 213). So the task before the compiler of a bilingual dictionary is not so simple for he is always to keep the user of his dictionary in his mind. He is, on one hand, to provide the users with meanings, sense ordering of equivalents, derivatives, usage, multiword-expressions, and on the other hand he is take in to consideration entry, typeface size, panels of illustrations and the layout of information. As the last two decades have witnessed a remarkable growth in lexicography by virtue of the induction of computer, a bilingual dictionary may rightly be seen as a great resource a learner can wield to learn a foreign language as learners’ encoding as well as decoding needs have witnessed an enormous growth, translational activities are on the rise due to a host of tasks the international communities are involved in since the world has become a global village. Swanson (1975) believed, ‘a bilingual dictionary can be useful and desirable to several kinds of people: students, travelers, translators and linguists. (p 6)’ So learners need reliable information regarding pronunciation, grammar, meaning, appropriateness of usage in a given situation involving communication in a second language. Zgusta (1971) believed:

The basic purpose of a bilingual dictionary is to coordinate with the lexical units of one language those lexical units of another language which are equivalent in their lexical meaning. The first language, to whose lexical units the lexical units of the other language are coordinated is called the source language; the order of the entries in a bilingual dictionary is given by the source language. The other language, whose lexical units are coordinated to the first one, is called the target language. (p. 294)

Stein (2000,24) believes, ‘in general bilingual dictionaries provide translation equivalents by contrast monolingual dictionaries define or paraphrase each word’. Unlike a monolingual dictionary, a bilingual dictionary is smaller which omits less frequent, obsolete and too technical senses (Zgusta,1971). It is neither too short nor too lengthy, and retains richness of information. It touches upon the vocabulary of two languages, and compilers of such dictionaries take into consideration needs of a particular user. The other question encountered by a lexicographer is whether the work is intended for speakers of SL or TL (Harrel, 1975). A complete bilingual dictionary lies in the realm of impossibility as every conceivable need of the prospective user cannot altogether be met due to availability of limited space and a certain desire for compactness. An ideal bilingual would the following characteristics:

I. It would provide for each word or expression in the source language (SL) just the right translation in the target language.

II. It would contain all the words, locutions, circumlocution, and idioms.

III. It would contain information on inflection, derivation, syntax and meaning.

IV. It would contain information on all the level of usage including special warnings about words not to be used in the presence of ladies, children or the superiors.

V. It would contain all the personal names, names of personages past and present, names of places, famous books, plays, names of characters therein any other name learner looks up.
VI. It would contain all the specialized vocabulary items of all the sciences, professions, manufacturing industries and trade.

VII. It would contain correct information about spellings as well as alternate or commonly encountered incorrect spellings.

VIII. It would include all the information to instruct the user to pronounce each word with the same proficiency as does the native speaker.

IX. It ought to be meant for the speakers of both SL and TL (target language) and cater to the needs of both.

X. It should have adaptability to the purpose of machine translation.

XI. It should have compactness.

XII. It would contain pictorial illustrations to items unfamiliar to the speakers of TL.

(Haas, M. R.:1975, 45-46)

Seeing these, it goes without saying that compiling a bilingual dictionary incorporating the aforesaid is quite an uphill task because the compiler has to find equivalent lexical units for the SL which are mostly culture-specific, and to take into account the other aspect of size and cost too. The compilers usually adapt the design features in order to meet the requirements of the expected user. Hatmann (2001) observed:

In dictionaries and other reference works, the information that is presented by the compiler and consulted by the user must be accessible and it is the most fundamental ordering device which achieves such accessibility (p. 58).

So it goes without saying that compiling a bilingual dictionary is a process involving a lot of care and consideration regarding its user, size, language focus and design features.

2. LEXICOGRAPHIC TRADITION IN PUNJABI

Punjabi is quite unique a language as it has an elaborate oral tradition rich in folk tales and mystic poetry. The compilation of dictionaries in Punjabi is a phenomenon which dates back to the English rule; the purpose was to meet the linguistic requirements of officers posted in the Punjabi speaking areas as well as the needs of the missionaries to communicate with the prospective converts speaking the language. The authors were the English and the initial compilations were mostly in the shape of glossaries of names, animals, birds and the literary or difficult words, and the lexicographic tradition before the grammatical studies in Punjabi (Kapoor&Gupta,2005). The dominant oral tradition in Punjabi becomes a problem for the lexicographer as the number of lexical items can not be determined easily, and with less frequent written usage the orthographic mistakes are very common. Seeing that the Vanguard Punjabi-English Dictionary (TVPED) and W. P. Hares’ English-Punjabi Dictionary (WPHEPD) have adopted the Roman script as a convenient means of for their expected users-the Englishmen. After the partition of the Sub-continent and the resultant making of Pakistan and India in 1947 A.D., the above mentioned dictionaries continued to come out by the Pakistani publishers without any consideration of the users’ needs, and the incorporation of modern lexicographic trends and practice. As regards the aim of bilingual dictionary, ‘the lexicographer’s intention to compile the dictionary’ is either to help as an aid in comprehension of texts in the SL or as an aid to the generation of texts in the target language (Zgusta,1971). The intention of lexicographers in case of compiling of TVPED and WPHEPD was to assist the Englishmen in learning Punjabi. TVPED was published in 1841 and its third reprint, claiming to be executed on methodical and scientific way, was brought out in 1972 A.D. WPHEPD, authored by W.P. Hares, was published by the Church Mission Society in 1929 A. D. and a leading publishing house in Lahore, Sang-e-Meel Publishers, Lahore, is bringing out its reprints without any amendments and revision when scores of years have passed and a host of things beggar amendments, revision and addition. As
structure and lay out of a dictionary determines its usability regarding the completion of language learning tasks, Hartmann (2001) appears to be suggesting that the ‘information presented by the compiler and consulted by the user must be accessible’ and this accessibility can be achieved through the ordering and arrangement of information. Design features of bilingual dictionaries in the modern lexicographic practice, the assessment of users’ needs, their accessibility and convenience are the fundamental issues taken into consideration while analyzing the design features of the bilingual dictionaries of Punjabi. The design features of TOEUD, printed in 2003, and QEUD, printed in 2001, were taken as model for preparing a checklist of these features.

1. Objectives
   The study aimed at finding the gaps between the lexicographic practice followed by the compilers of the bilingual dictionaries in Punjabi and the modern practice.

2. Research Methodology
2.1 Population and Sample
   Two bilingual Punjabi dictionaries: the Vanguard Punjabi-English Dictionary (TVPED) and W. P. Hares’ English-Punjabi Dictionary (WPHEPD) were taken as sample for the following reasons:
   I. They are easily accessible as the libraries and book stores have got them.
   II. They are being reprinted by the publishers
   III. There is no problem with them as regards their availability
   IV. They are the sole resource as there is dearth of lexicographic activity in the Punjabi language
   
   Two bilingual dictionaries: the Oxford English-Urdu Dictionary (OEUD) and Qaumi English-Urdu Dictionary (QEUD) were taken for the following reasons:
   I. These dictionaries retain the design features of good bilingual dictionaries
   II. They enjoy enormous popularity among the academics as well as the language learners for encoding as well as decoding.
   III. They are affordable and the price is reasonably less as compared to that of other available bilingual dictionaries.
   IV. They are available at all the leading book shops and libraries.
   V. They are revised and updated from time to time.
   VI. They are in conformity with the tradition of dictionary compilation in Asia

2.2 Checklist
   Two checklists on the macrostructures as well as the microstructures of bilingual dictionaries were prepared; the features discussed by Hartmann (2001) as well as of Oxford English-Urdu Dictionary (OUED) and Qaumi English-Urdu Dictionary (QEUD) were taken as the model. The checklists focussed mainly on the macro as well as the micro structures.

The following abbreviations have been used in the Study:

Abb = Abbreviations
Ack = Acknowledgement
Chem = Chemical elements
Class = Classification of non-verb lexemes
Col = Collocation
Cr = Cross reference
The design features of these dictionaries were analysed as under:

1. **Macrostructure**
   a) Front matter includes the title page, preface, User’s guide, Introduction, labels and abbreviations. As labels are too great in number and writing of all the labels is almost impossible on the limited space of a dictionary, it is the compilers’ own decision as to which labels should be included.

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Tp</th>
<th>Pr</th>
<th>Ug</th>
<th>Ack</th>
<th>Intro</th>
<th>Lab</th>
<th>Abb</th>
</tr>
</thead>
<tbody>
<tr>
<td>TVPED</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>WPHEPD</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
</tr>
</tbody>
</table>
Discussion:

i. TVPED has no usage guide, acknowledgement, introduction, labels, and abbreviation, and with only title page and preface, the user is at a loss regarding different types of information about SL as well as TG.

ii. WPHEPD retains title page, introduction and an incomplete list of abbreviations as in case of the entry *Glorification* gives the equivalent ‘Sitaish’ [praise] denoted by ‘d’ with no information about the symbol.

b) Middle matter includes the panels and plates of illustrations

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Panels and plates of illustrations</th>
</tr>
</thead>
<tbody>
<tr>
<td>TVPED</td>
<td>×</td>
</tr>
<tr>
<td>WPHEPD</td>
<td>×</td>
</tr>
</tbody>
</table>

Table 1 (b) Description of middle matter

Discussion:

Both TVPED and WPHEPD are without any panels and plates of illustrations, and therefore without a very important feature of bilingual dictionary they are of little help to learn certain articles in a second language which are alien to his language and culture.

c) Back matter are the component parts of a dictionary’s macrostructure, they are located between the central word-list section and the end of the work, names of persons, places, weight and measurement, military ranks, chemical elements, alphabetical and numerical symbols and musical notation, quotations and proverbs fall in the category of back matter.

Table 1 (C) Description of back matter

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Whether b/w word-list section or not?</th>
<th>NPr</th>
<th>NPl</th>
<th>W &amp; M</th>
<th>MR</th>
<th>Chem</th>
<th>NSA</th>
<th>MN</th>
<th>Q&amp;P</th>
</tr>
</thead>
<tbody>
<tr>
<td>TVPED</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>WPHEPD</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

Discussion:

Both the dictionaries show a complete absence of back matter. In compiling a bilingual dictionary, cultural data poses a great deal of problem for the foreign language learners. This encyclopaedic data should be given in dictionary to facilitate the learners.

2. Microstructure

Two checklists focusing on phonological, grammatical, semantic and morphological information keeping in mind the design features of OEUD and QEUD, and the microstructures discussed by Hartmann (2001) and Haensch & Omenaca (2004)

I. Phonological Information

Phonological information includes the description of correct pronunciation as well as the pronunciation variants.
II. **Grammatical Information**
Description of parts of speech, singular-plural distinction, gender, countable-uncountable distinction and usage are in the domain of grammatical information.

III. **Semantic Information**
Information regarding headword (lemma), meaning, different senses, lexical relations (hyponym, super ordinate, synonym, antonyms, homonym and homonym), collocation, idioms, verbal or pictorial illustrations and stylistic constraints like formal-informal, written-spoken, slang, colloquial and register are included in semantic information.

IV. **Morphological Information**
From the head word stem different word, morphological information is about the derivational as well as inflectional forms of lexemes.

The Corpus

**TVPED**
Twenty three words have been taken randomly since the dictionary is written in the Roman script 'sh' and 's' sound are represented by 's', and similarly words beginning with 'q' are represented by 'k'. 'w' represents both 'w' and 'v'.

( Arth[definition], Badlaun[exchange], Chaki[flour mill], Dada[father of father], Etibar[trust], Faida[profit], Gabara[balloon], Hukum[order], Ilam[knowledge], Jat[caste], Kanak[wheat], Lalach[avarice], Mafi[redemption], Nath[jewellery related with nose], Ohur pour[endeavours], Palangh[bed], Rada[brick], Sabar[patience], Takkar[confrontation], Uggamn[ to grow], Wahm[caprice], Yakin[conviction].

**WPHEPD**
Twenty six words of English have been taken randomly as corpus for this study, they are as under:

Abandon V [Give up], Bachelor N [Unmarried man], Calf N [Young of cow], Darkness N [No light], Eat V [Take as food, Failure N [Lack of success], Glorification N [Praise], Haste N [Hurry], Irritate V [Annoy], Jeer V [Mock], King N [Royal head], Lady N [Woman], Majestic Adj [Splendid & Kingly], Nothing Pron [Naught], Offence N [Crime], Paramount Adj [significant], Quilt N [Thick covering of the bed], Rain V&N [Fall of water], Screech V [Harsh high pitched cry], Toast V [Making bread crisp], Ugly Adj [Unpleasant], Valueless Adj [Worthless], Wagon N [Four wheeled vehicle for carrying heavy loads], Yoke N&V [Egg protein& harness], Zealous N [Sanguine].

### 2.3 Methodology

TVPED is a Punjabi-English dictionary and WPHEPD is a English-Punjabi dictionary, so the corpus was taken separately from the dictionaries and two check lists were prepared to assess the features in each dictionary. The equivalents for Punjabi lexemes were taken from Tanveer Bhukhari’s Punjabi-Urdu Dictionary and Irshad Ounjabi’s Punjabi-Punjabi Dictionary, the corpus was analysed on the basis of the checklist, the table 2 (a) shows the overall result regarding TVPED and table 2 (b) shows the overall result regarding WPHEPD, ‘√’ shows the presence while ‘×’ shows non-applicable feature. After analyzing the corpus of both the dictionaries their commonalities were highlighted along with giving suggestions as to how they could be made more up to-date as regards their design features.
Table 2 (a)  
Checklist for drawing results of the micro structure and mediostructure of TVPED

<table>
<thead>
<tr>
<th>Head word</th>
<th>Phono-logical information</th>
<th>Grammatical information</th>
<th>Morphological information</th>
<th>Semantic information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro</td>
<td>P</td>
<td>s</td>
<td>D</td>
<td>C/Un</td>
</tr>
<tr>
<td>Arth[ Definition]</td>
<td>x</td>
<td>x</td>
<td>√</td>
<td>x</td>
</tr>
<tr>
<td>Badla un [Exchange]</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>–</td>
</tr>
<tr>
<td>Chaki [Flour mill]</td>
<td>x</td>
<td>x</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Dada[Father of father ]</td>
<td>x</td>
<td>x</td>
<td>√</td>
<td>x</td>
</tr>
<tr>
<td>Etibar [trust]</td>
<td>x</td>
<td>x</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Faida [Benefit]</td>
<td>x</td>
<td>x</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Gabara [Balo on]</td>
<td>x</td>
<td>x</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Hukam [order ]</td>
<td>x</td>
<td>x</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Ilam[Knowledge]</td>
<td>x</td>
<td>x</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Jat [Caste ]</td>
<td>x</td>
<td>x</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Kanak</td>
<td>x</td>
<td>x</td>
<td>√</td>
<td>x</td>
</tr>
</tbody>
</table>

8
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>
Table 2 (b)
Checklist for drawing results of the micro structure and mediostructure of WRHEPD

<table>
<thead>
<tr>
<th>Head word</th>
<th>Phonological information</th>
<th>Grammatical information</th>
<th>Morphological information</th>
<th>Semantic information</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pros</td>
<td>Pos</td>
<td>D/ C/ Use</td>
<td>H</td>
</tr>
<tr>
<td>Abandon</td>
<td>x</td>
<td>√</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Bachelor</td>
<td>x</td>
<td>√</td>
<td>√</td>
<td>x</td>
</tr>
<tr>
<td>Calf</td>
<td>x</td>
<td>√</td>
<td>√</td>
<td>x</td>
</tr>
<tr>
<td>Darkness</td>
<td>x</td>
<td>√</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Eat</td>
<td>x</td>
<td>√</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Failure</td>
<td>x</td>
<td>√</td>
<td>√</td>
<td>x</td>
</tr>
<tr>
<td>Glorification</td>
<td>x</td>
<td>√</td>
<td>√</td>
<td>x</td>
</tr>
<tr>
<td>Haste</td>
<td>x</td>
<td>√</td>
<td>–</td>
<td>x</td>
</tr>
<tr>
<td>Irritation</td>
<td>x</td>
<td>√</td>
<td>–</td>
<td>x</td>
</tr>
<tr>
<td>Jeer</td>
<td>x</td>
<td>√</td>
<td>√</td>
<td>x</td>
</tr>
<tr>
<td>King</td>
<td>x</td>
<td>√</td>
<td>√</td>
<td>x</td>
</tr>
<tr>
<td>Lady</td>
<td>x</td>
<td>√</td>
<td>√</td>
<td>x</td>
</tr>
<tr>
<td>Majestic</td>
<td>x</td>
<td>√</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Nothing</td>
<td>x</td>
<td>√</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Offence</td>
<td>x</td>
<td>√</td>
<td>–</td>
<td>x</td>
</tr>
<tr>
<td>Paramount</td>
<td>x</td>
<td>√</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Quilt</td>
<td>x</td>
<td>√</td>
<td>√</td>
<td>x</td>
</tr>
<tr>
<td>Rain</td>
<td>x</td>
<td>√</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Screech</td>
<td>x</td>
<td>√</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
3. FINDINGS
A bilingual dictionary exposes TL to the learner and TL is invariably different from SL. Each
glanguage is distinct as regards its phonological characteristics. So it is highly important for a
bilingual dictionary to provide phonological information in canonical form. Table 3 (a) shows that TVEPD and WPHPD show complete disregard toward this needs of users. The
user often finds himself guessing the pronunciation of TL.

### Table 3 (a)
**Overall Percentage of Phonological Information:**

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Pro</th>
</tr>
</thead>
<tbody>
<tr>
<td>TVPED</td>
<td>0%</td>
</tr>
<tr>
<td>WPHPD</td>
<td>0%</td>
</tr>
</tbody>
</table>

Grammatical information includes both morphological and syntactic characteristics. Not only
does a learner need knowledge about lemma but rather about its inflected forms, derivatives.
The information regarding part of speech, declension, singular and plural and usage are also
among the needs of dictionary user. Zgusta (1971:340) described:

“The character and amount of the grammatical information is correlated with the
intention of the dictionary in an obvious way: a dictionary for the comprehension of the
source language can afford (if it is strictly limited in its intentions) to concentrate only upon
those grammatical indications by means of which the multiple meaning of the entry-word is
disambiguated; the intention to describe the source language brings with it necessity of
indicating all such grammatical properties of the entry-word, irrespective of whether or not
they disambiguate its meaning; the dictionary which intends to help to generate texts in the
target language must give rich instruction on how to use the equivalent.”

TVPED rarely gives information about parts of speech and the instances of usage and
examples are few and far between as shown in table 3 (b); whereas WPHPD is deficient in
information about countable-uncountable distinction of nouns, and the examples of lexical
items as to how they can be used.

### Table 3(b)
**Overall Percentage of Grammatical Information:**

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Ps</th>
<th>D</th>
<th>C/Un</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>TVPED</td>
<td>4.45%</td>
<td>81.81%</td>
<td>66.66%</td>
<td>36.8%</td>
</tr>
<tr>
<td>WPHPD</td>
<td>100%</td>
<td>57.14%</td>
<td>0%</td>
<td>3.8%</td>
</tr>
</tbody>
</table>

Headword or lemma, the entry marked in bold, and in canonical form and abbreviations in
the form of clipping, contractions, acronyms, blend and alphabetism are desirable features of
a bilingual dictionary in the present day practice. Both TVPED and WPHPD have head
words organized alphabetically, yet both are deficient as far as inflections, abbreviations and derivatives are concerned. Irregular forms of word pose a great deal of trouble for foreign language learners, and subsequently it decreases the usefulness of dictionary.

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>HW</th>
<th>Inf</th>
<th>Der</th>
<th>Et</th>
</tr>
</thead>
<tbody>
<tr>
<td>TVPED</td>
<td>86.36%</td>
<td>40.9%</td>
<td>45.45%</td>
<td>36.36%</td>
</tr>
<tr>
<td>WPHEPD</td>
<td>100%</td>
<td>0%</td>
<td>3.8%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 3 (c)

Overall Percentage of Morphological Information:

Definition and equivalents, etymology, lexical relations-synonyms, antonyms and homonyms etc.-collocation, idiomatic expressions pertain to the domain of semantics. About semantic information in dictionary, Apresjan (2000:7) seems to be suggesting that lexical correspondence between the lexical units of two different languages are rare except for technical terms and similar words, and each equivalent describes only one aspect of the word. There is a definite need of hierarchical arrangement of different senses in the microstructures. Both TVPED and WPHEPD give meaning equivalents but lack in information regarding lexical relations among different equivalents, their habitual co-occurrence, stylistic information, diction of written and spoken and information about colloquial speech, slang and register. Elaboration both verbal and pictorial is a great tool wielded by a bilingual lexicographer in order to sensitize the user of dictionary with concepts and things alien to his language and cultural background. Table 3 (d) shows both TVPED and WPHEPD are not good resources as far as semantic information is concerned, meaning equivalents are not perfect match for translation of a word in SL or vice versa. So by giving meaning equivalents without explaining different senses involved, these dictionaries are of little practical use in translation and speech production. Moreover, in-text cross reference to related notions is a desired feature in modern lexicographic practice in order to avoid repetition as Hartmann (2001) identified it as mediostructure or the various means of achieving cross-reference. Mediostructure is absent in both the dictionaries.

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Me</th>
<th>Lex</th>
<th>Ft</th>
<th>Col</th>
<th>Idm</th>
<th>Vil</th>
<th>Pil</th>
<th>Cr</th>
</tr>
</thead>
<tbody>
<tr>
<td>TVPED</td>
<td>100%</td>
<td>4.45%</td>
<td>9.09%</td>
<td>27.27%</td>
<td>31.81%</td>
<td>45.45%</td>
<td>0%</td>
<td>13.63%</td>
</tr>
<tr>
<td>WPHEPD</td>
<td>100%</td>
<td>23.07%</td>
<td>0%</td>
<td>0%</td>
<td>11.53%</td>
<td>3.8%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 3 (d)

Overall Percentage of Semantic Information:

4. CONCLUSION
The study investigated the design features of two bilingual Punjabi dictionaries namely the Vanguard Punjabi-English Dictionary and W. P. Hares’ English-Punjabi Dictionary. Design features in two Urdu bilingual dictionaries and Harmann’(2001) description of such features. Gap between the practice of bilingual lexicography, at present, and the one seen in these dictionaries of Punjabi is wide. The checklist and findings throw light on the following aspects:

1. Phonological information has not been given any representation.
2. Grammatical information regarding, part of speech, declension ,singular-plural distinction, and usage has not been given completely.
3. The number of lexical entities is very limited in WPHEPD, and TWPED contains a lot of words which are no more in currency.
4. Meaning, collocation, idioms, proverbs are in great dearth in both the dictionaries.
5. Examples of the lexical entities items are not given.
6. Information on lexical relations like synonyms, antonyms, hyponyms, homonyms, homographs and homophones is totally absent.
7. Roman script is not used for Punjabi; the users with the knowledge of Persio-Arabic script do not feel at home with using these dictionaries.
8. Stylistic information, which is highly desirable for foreign language learners, is totally absent.
9. Inflection of head words have not been given in both the dictionaries.
10. There is no description of the frequency counts, there are equivalents of words given without any sequence based on the sense and closeness with the entry.

5. RECOMMENDATION
The findings provide insights into the design features of these dictionaries, the following recommendations are based on these findings:

1- Correct pronunciation should be given in these dictionaries.
2- Grammatical information neglected in the dictionaries should be added in them.
3- Sense relations should also be given to make the users better able to associate meaning with codes.
4- Stylistics information should also be given where necessary.
5- Elaboration of meaning through examples of usage should be included with each other entry.
6- Morphological information should be complete in all respects.
7- Pictorial illustration should be given for entries having different cultural background.
8- Culture-bound words should be illustrated in detail.
9- Macro structure should be included in the dictionaries.
10- A general restructuring and revamping is needed in order to make these dictionaries more up-to-date, so a panel of linguists as well as lexicographers should take upon the task of revising these dictionaries which are a great contribution of the Englishmen in the sub-continent.

REFERENCES


Swanson, D. C. (1975) “The selection of entries for a bilingual dictionary”. In Householder, F. W. & Saporata, S. (eds.) Problems in Lexicography. Research Center for language and Semiotic Studies Indiana University, the U. S. A.

Abstract: The present study will examine the results of a small-scale survey on dictionary use of Japanese high school students. The author asked a former colleague to help conduct a survey on dictionary use at her school. The objectives of the survey were (1) to look into how English-Japanese (E-J) dictionaries were used by high school students, (2) to learn how much students know about using a dictionary effectively, and (3) to offer suggestions for the improvement of E-J dictionaries. From the students’ answers, the author obtained the following results. It seems that the choice of a dictionary was greatly influenced by outside factors such as a recommendation by the school or an older sibling. The students sometimes had difficulty finding the correct meanings of words, and even more so for idioms. It seems that the teachers’ expectations of E-J dictionaries were different from those of the students. The teachers hoped that E-J dictionaries would have more headwords and example sentences, while the students thought they would be more user-friendly in terms of looking up the meanings of words and idioms. In conclusion, the author thinks that Japanese teachers of English should be aware of the problems their students have with dictionaries and try to spare some time for increased training in dictionary use. As for the students, I would strongly urge them to read the instructions and use their dictionaries as often as possible.

Key Words: English-Japanese (E-J) dictionary, dictionary use, dictionary training

1. INTRODUCTION

English-Japanese (E-J) dictionaries are important tools for learning English. They contain a variety of information about English words and phrases, and users of E-J dictionaries are expected to communicate actively with other people when using them. This author thinks that for E-J dictionaries to be truly useful and effective, we also need to know how users use them and what opinions they have about them. For that purpose, the author asked Japanese high school students to fill out a questionnaire about their dictionary use in 2004. The present study puts forth the results, looks into some of the problems E-J dictionaries have, and elaborates on the author’s ideas for improving them.
2. THE QUESTIONNAIRE

2.1 The Goals

The goals of the questionnaire are as follows:

(1) To look into how E-J dictionaries are used by high school students.
(2) To learn how much students know about using a dictionary effectively.
(3) To make some suggestions for the improvement of E-J dictionaries based on the results of this research.

2.2 About The School and The Students

(1) The name of the school: Sano High School (in Osaka)
(2) The categories of the students
   - General Course: 79 1st-year students in 2 classes.
   - International Liberal Arts Course: 77 1st-year students in 2 classes.
   A total of 156 students participated in this study.
(3) The dates the questionnaire was conducted
   - December 16th, 17th and 20th in 2004.
(4) The reasons why this author chose Sano High School are:
   (a) It is known for its higher-than-average academic level, so it is assumed that the students use their E-J dictionary frequently.
   (b) The school has two courses, General Course and International Liberal Arts Course.
      The author thought it was possible to compare the students’ dictionary use between the two courses.
   (c) The school is located in the same city as where the author lives, which makes it convenient for contacting her former colleague for discussion and consultation.

3. THE RESULTS

In this chapter, the results of the questionnaire are shown.

(Q-1) Do you have an E-J dictionary?
   a. Yes.  b. No.

Table 1 Percentage of students possessing an E-J dictionary

<table>
<thead>
<tr>
<th></th>
<th>1-1</th>
<th>1-2</th>
<th>1-7</th>
<th>1-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>97.4</td>
<td>100</td>
<td>97.3</td>
<td>97.5</td>
</tr>
<tr>
<td>b</td>
<td>2.6</td>
<td>0</td>
<td>2.7</td>
<td>0</td>
</tr>
<tr>
<td>NA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.5</td>
</tr>
</tbody>
</table>

(unit: %)

Almost all the students had an E-J dictionary. It is assumed that it’s because they understood its necessity learning a foreign language.
(Q-2) How many dictionaries do you have?
   a. 1   b. 2   c. 3   d. 4   e. 5

Table 2 Number of dictionaries owned by the students

<table>
<thead>
<tr>
<th></th>
<th>1-1</th>
<th>1-2</th>
<th>1-7</th>
<th>1-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>42.1</td>
<td>45</td>
<td>47.1</td>
<td>35</td>
</tr>
<tr>
<td>b</td>
<td>39.5</td>
<td>40</td>
<td>41.7</td>
<td>37.5</td>
</tr>
<tr>
<td>c</td>
<td>7.9</td>
<td>10</td>
<td>5.6</td>
<td>20</td>
</tr>
<tr>
<td>d</td>
<td>7.9</td>
<td>2.5</td>
<td>5.6</td>
<td>2.5</td>
</tr>
<tr>
<td>e</td>
<td>2.6</td>
<td>0</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>0</td>
<td>2.5</td>
<td>2.5</td>
<td></td>
</tr>
</tbody>
</table>

(Unit: %)

More than 70% of the students had one or two E-J dictionaries. The author thinks that the number of dictionaries is not of much importance as long as the dictionary is suitable for the needs and level of the students. When students have more than one dictionary, how they use them differently would be a topic for future research.

(Q-3) Please specify the names of your dictionaries
   a. G   b. WIS   c. YP   d. LEX   e. KNC   f. PROG   g. LUM
   h. NG   i. AF   j. AG   k. LH   l. NP   m. E-G   n. WP
   o. F   p. BG   q. LPROG   r. SA   s. Others   t. I don’t remember

Table 3 Name of students’ dictionaries

<table>
<thead>
<tr>
<th></th>
<th>1-1</th>
<th>1-2</th>
<th>1-7</th>
<th>1-8</th>
<th>Numb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>50.9</td>
<td>48.5</td>
<td>61</td>
<td>44.4</td>
<td>120</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>0</td>
<td>1.9</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>e</td>
<td>0</td>
<td>3.2</td>
<td>3.7</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>f</td>
<td>0</td>
<td>0</td>
<td>1.9</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>g</td>
<td>0</td>
<td>0</td>
<td>3.7</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>h</td>
<td>0</td>
<td>3.2</td>
<td>0</td>
<td>1.7</td>
<td>3</td>
</tr>
<tr>
<td>i</td>
<td>1.8</td>
<td>12.9</td>
<td>1.9</td>
<td>6.3</td>
<td>14</td>
</tr>
<tr>
<td>j</td>
<td>0</td>
<td>1.6</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>k</td>
<td>1.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>l</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>m</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>n</td>
<td>0</td>
<td>1.6</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>o</td>
<td>19.2</td>
<td>8.1</td>
<td>3.7</td>
<td>9.5</td>
<td>24</td>
</tr>
<tr>
<td>p</td>
<td>1.8</td>
<td>3.2</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>q</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>r</td>
<td>0</td>
<td>1.6</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>s</td>
<td>7</td>
<td>4.8</td>
<td>1.9</td>
<td>9.5</td>
<td>14</td>
</tr>
<tr>
<td>t</td>
<td>15.7</td>
<td>11.3</td>
<td>18.4</td>
<td>25.4</td>
<td>42</td>
</tr>
<tr>
<td>NA</td>
<td>1.8</td>
<td>0</td>
<td>1.9</td>
<td>3.2</td>
<td>4</td>
</tr>
</tbody>
</table>

(Unit: % for the classes) Total 236
About half of the students owned dictionary A. Dictionaries A, I and O are the recommended dictionaries at the school. It is obvious that their choice of dictionaries is greatly influenced by the school, or more accurately, by the English teachers working there. As for the answer, “Others,” ten students replied that they used an electronic dictionary, but they did not reveal the name of the E-J dictionary used by the software.

**Q-4** When did you buy your dictionary?

Table 4 Time at which the students bought their dictionaries

<table>
<thead>
<tr>
<th>Date</th>
<th>March, 2004</th>
<th>April, 2004</th>
<th>May, 2004</th>
<th>April-March, 2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer</td>
<td>8</td>
<td>94</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

(Unit: number of the students, This is a part of the total.)

**Q-5** Why did you choose the particular dictionary? (Multiple choice, select all that apply)

- a. Recommended by the school
- b. Large number of many headwords
- c. Detailed explanations of grammar and usage
- d. Included information on related subjects and culture
- e. Numerous examples
- f. Easy-to-understand definitions
- g. Many pictures and illustrations
- h. Use of large fonts
- i. Easy-to-read
- j. Multi-colored
- k. Portable, handy
- l. Easy-to-look at in general
- m. Others
- n. No answers

Table 5 Reasons for choosing dictionaries

<table>
<thead>
<tr>
<th>Reason</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>i</th>
<th>j</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answers</td>
<td>114</td>
<td>11</td>
<td>3</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>21</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Reason</th>
<th>m</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answers</td>
<td>9</td>
<td>30</td>
</tr>
</tbody>
</table>

(Unit: The number of the students)

More than 70% of the students (114 students) chose answer (a). This clearly shows the strong influence of recommendations made by the school. In regard to the other answers, which indicate features of the dictionaries themselves, the students tended to choose superficial features such as “being handy,” and “easy to look at,” as their
primary reasons for choosing a particular dictionary. Features related to the content of dictionaries, such as “the number of headwords,” and “have many example phrases and sentences,” were secondary. The results indicate that although the content is the most important feature of a dictionary, its physical properties are also important and should not be neglected. Some students gave answers such as, “The dictionary was recommended to me by my family members or relatives,” “My brothers (or sisters) used the same one,” “I heard it was very popular,” etc. Opinions from other users are also an important factor in choosing a dictionary.

(Q-6) What items of a word do you look up in your dictionary? (multiple choice, select up to three)

a. Spelling  
b. Pronunciation  
c. Part of speech  
d. Conjugation  
e. Meaning of words  
f. Grammar, usage  
g. Meaning of idioms  
h. Derivatives  
i. Synonyms, antonyms  
j. Useful information such as culture, corpus panel  
k. Etymology  
l. Others  
m. No answers

Table 6 Items looked up in dictionary

<table>
<thead>
<tr>
<th>Item</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>i</th>
<th>j</th>
<th>K</th>
<th>l</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answers</td>
<td>89</td>
<td>26</td>
<td>10</td>
<td>32</td>
<td>152</td>
<td>24</td>
<td>81</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

(Unit: the number of answers)

The results clearly show that the students used their dictionaries mainly to look up the meanings of words and phrases (233 out of 421 replies), as well as to check the spelling, conjugation and pronunciation (147 out of 421 replies). The results show the importance of how E-J dictionaries define the meanings of words and phrases, and how lexicographers describe them.

(Q-7) How do you find the meaning of words in your dictionary?

a. Always read only the first meaning  
b. Always read the first and second meanings  
c. Read the first meaning, but sometimes read the second one  
d. Read through all the meanings and choose the most appropriate one according to context  
e. Others  
f. No answers
From the results, it is clear that the students understand the importance of choosing the right meaning according to the context of a given sentence. There was one noteworthy reply under “others,” where a student stated that she used example sentences to determine the most appropriate meaning. Words often collocate with some words more than others and the meaning is decided by the context and the other words it collocates with. Therefore, reading example sentences to decide the meaning of a word is a useful method and should be strongly recommended to all students. From the students’ results, this author thinks that lexicographers should try to list high-frequency phrases and collocations more effectively in their dictionaries.

(Q-8) Do you think the definitions in your dictionary are easy to understand? And if so, why?

- a. very easy to understand
- b. easy to understand
- c. average
- d. difficult
- e. very difficult

Table 8 Difficulty in understanding the definitions

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>NA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4.5</td>
<td>48.1</td>
<td>39.7</td>
<td>6.4</td>
<td>0.6</td>
<td>0.6</td>
</tr>
</tbody>
</table>

About 90% of the students answered positively. As to the reasons for the negative answers, some students said that they sometimes found the descriptions in parentheses or explanations difficult to understand. The students didn’t clearly specify what they meant, but the author guesses they have difficulty understanding usage of words in parentheses. If this is the case, E-J dictionaries should have clear and easy-to-understand descriptions of word usage. As for the students themselves, the first thing they should be encouraged to do is read the instructions and get used to the terms used throughout the dictionary. Future studies should deal with these matters in more detail.

(Q-9) What items of words do you feel are difficult to find when you use your dictionary?

- a. Spelling
- b. Pronunciation
- c. Part of speech
- d. Conjugation
e. Meaning of words f. Grammar, usage
g. Meaning of idioms h. Derivatives
i. Synonyms, antonyms j. Useful information such as culture, corpus panel
k. Etymology l. Others
m. No answers

Table 9 Items that are difficult to find

<table>
<thead>
<tr>
<th>Items</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>i</th>
<th>j</th>
<th>k</th>
<th>l</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td>10</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>47</td>
<td>14</td>
<td>87</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

(Unit: The number of the students)

As we find in question #6, about 50% of the students used their dictionaries to look up meanings of words and idioms, but at the same time many students had some difficulty in doing it, especially looking up the meanings of idioms. The author guesses that one reason for this might be that they don’t know which word in the idiom they should look up. There is a difference in the level of details of the descriptions about it among E-J dictionaries and this author wonders how many students understand those descriptions in their dictionaries.

(Q-10) Please specify the items of a word that you would like to see in future editions or the items that should be included in more detail.

Many students said that finding definitions of idioms in their dictionaries was difficult and they would like to see more idioms and their meanings in future editions and wished them to be easier to look up. They also hoped that E-J dictionaries would contain more example phrases and sentences. Only a few students said that they wanted the pronunciation of headwords to be written in katakana.

(Q-11) In your opinion, what would make E-J dictionaries better?

Many of their opinions expressed were concerned with the dictionary’s physical appearance such as the size of fonts, the use of color ink etc. A few students said that they wanted the paper of their dictionary to be a little thicker so that they could write on the pages with a marker, and some students said that they wanted a lighter, more portable dictionary. Not all of the comments were focused on appearance, however. Some of the students said they wanted the explanations to be in easier or more colloquial Japanese, and a few noted that there was a growing trend towards using electronic dictionaries.

3. DISCUSSION

In this section, the three major points of the questionnaire will be discussed.
3.1 How The Students Choose Their Dictionaries
The first point is the students’ reasons for choosing their dictionaries. As we can see from question #5, their choice of E-J dictionaries was influenced by external factors such as a recommendation by the school or the popularity of a dictionary among its users. Also, some of them were given a dictionary by one of their siblings. Only a small number of students chose their dictionaries based on factors directly related to its contents, such as information on grammar, usage of words, or the number of example sentences. For students to choose their E-J dictionaries more independently, their teachers should try to explain some of the features in the recommended dictionaries first, and then give some advice on comparing E-J dictionaries. Such advice should include the pros and cons of both physical and substantial features, like the layout of the content and the order of definitions of headwords and phras-es. Students should try to follow that advice and take some time in choosing a dictionary that is most suitable to them.

3.2 The Students’ Difficulty
The second point is that the students sometimes had difficulty in looking up the meanings of words and phrases. We should think of this issue from the standpoint of teachers, students and lexicographers.

For English teachers, it is necessary to spare some time for dictionary training. During the course of this study, the author asked the English teachers some questions about their dictionary-use instructions. The results showed that only two out of nine teachers gave instructions on dictionary use. They said that they taught how to use an E-J dictionary in four or five parts, each part lasting about five minutes. They also said that their guidance was focused mainly on how to find headwords and to recognize their parts of speech. The lack of dictionary-use instructions is also reported in a research paper by Herbst and Stein(1987: 119-121). They summarized the dictionary-use instructions in schools in Bavaria, Germany as follows:

(1)When the monolingual English dictionary is first introduced in class, this is accompanied by a systematic initial treatment of dictionary-use skills. However, on average, no more than two to three lessons are devoted to the dictionary at this stage.

(2)Some teachers practice the use of the dictionary regularly after this phase, but this seems to be exceptional. Often the dictionary is discussed in connection with texts or student errors which could have been avoided by consulting a dictionary. According to students’ statements, however, many teachers simply refer them to the dictionary.
at such times rather than explaining in detail how the information could actually have been retrieved.

(3) As far as the major types of dictionary use are concerned, it is the value of the dictionary as an instrument in encoding. More importance is attached to meanings and idioms than to information on syntax. and there too it seems that example sentences and prepositions in bold type are given more attention than verb patterns or other codes (such as C or U).

(4) Most teachers seem rather skeptical about their students’ ability to use the dictionary appropriately. It is interesting to note, however, that many teachers do not blame this supposed inability on the quality of the dictionaries but attribute it to laziness or a general lack of interest on the part of the students. Both their results and this research results emphasize the importance of dictionary-use instruction.

Also Scholfield (1999: 29) clearly states the importance of dictionary training as follows:

The message for the teacher is again clear. Learners need experience of using monolingual dictionaries in class. They need exercises in reading definitions and example and exploiting the information appropriately, to develop the appropriate strategies.

Therefore, this author strongly suggests that English teachers encourage their students to read the “How to use the dictionary” section in the front of a dictionary. And when Béjoint (2000: 145) did a study on the use of dictionaries among 122 university students in France, one of his findings was that the students were not aware of the wealth of information dictionaries contain and that the forward and instructions were hardly ever read (only about 10 percent read them). Continuous and practical dictionary training should be introduced into daily classes.

Some researchers point out the lack of dictionary-use skills of their users, especially student users. Herbst and Stein (ibid: 115) comment on the problem:

Research in various countries into the use of monolingual English learners’ dictionaries has shown quite clearly that the majority of school pupils and university
students are not able to make full use of the different types of information which those dictionaries provide. (・・・) The problem appears to be, rather, that successful use of EFL dictionaries such as the Oxford Advanced Learner’s Dictionary (OALDCE) or the Longman Dictionary of Contemporary English (LDOCE) presupposes a specific competence because of the wide range of information they contain, and that students lack this competence.

After introducing the results of their research on dictionary-use in Bavaria, they point out the important place dictionaries have in language acquisition as follows (ibid: 125)

In all learning there is a very subtle and gradual process of acquisition. (・・・) The dictionary is an institution which enables language learners to check their own knowledge, and to eliminate weakness in spelling, pronunciation, grammar and meaning. (・・・) The more effective their handing of this tool, the better their command of the language and the greater their self-confidence. (・・・) The dictionary is a tool that will wean learners away from the classroom teacher and guide them towards further independent study.

They state that dictionaries also help learners acquire new knowledge. It can be said that dictionaries are tools not only for gaining knowledge but also for giving courage to learners to be independent in their studies.

Regarding lexicographers, this author would like to ask them to be aware of the problems Japanese high school students face when they use their E-J dictionaries. They said that they had difficulty understanding definitions of headwords, descriptions in parentheses and finding idioms. Bogaards (1996: 280-281) lists three important criteria when we use a dictionary; findability, comprehensibility and usability. When students read “How to use the dictionary” section carefully, the problem of comprehensibility is expected to be solved somewhat because descriptions in parentheses are mainly used as labels in the dictionary. But even in such cases, lexicographers should use labels which are easy to understand for users. The students also had difficulty finding idioms. The explanation for where to find idioms varies from dictionary to dictionary. For example, one dictionary states that idioms are listed under their main component words and cross-referenced if necessary, and that generally, idioms which include a noun are listed under that noun. Another dictionary states that idioms are listed under their most characteristic word and that idioms which contain a noun such as “a verb + a noun” or “a preposition + a noun”
are listed under the noun, while phrasal verbs are listed under the verbs. This author doubts that Japanese high school students clearly understand the main component words or most characteristic words of idioms. Also, some dictionaries do not give any examples regarding this problem. Concerning this problem, Iwasaki(2002:27) lists the ability to discern collocations as one of the most important things needed for dictionary search. He (ibid: 73) also says that including as many collocations as possible is one of his criteria in choosing an E-J dictionary. There is a growing interest in idioms and phraseological phrases in English these days. These expressions are useful for more effective communication in that learners can use them as set phrases. In other words, they have practical and pedagogical importance. Yagi (2006: 214-216) and Yagi (2007: 20-21) explain what phraseological phrases are and stress their importance in English lexicography and education, and introduce some examples such as “forget it” and “pretty much.” This author supports their claims and expects lexicographers to compile balanced learner’s E-J dictionaries which are useful for both receptive and productive purposes.

3.3 The Difference In Opinions Between The Students And The Teachers
It seems that there is a difference in what the students and the teachers ask for in E-J dictionaries. The students wanted E-J dictionaries to be easier to look up definitions of headwords and idioms, and easier to understand the explanations provided. They also answered that they wanted more interesting example phrases and sentences. In other words, they hoped that E-J dictionaries would improve qualitatively. On the other hand, the teachers attached importance to the number of headwords, example phrases and sentences when they recommended E-J dictionaries. Quantitative aspects were more important to them. As students’ English abilities progress, they might need an E-J dictionary with more headwords and examples, but learners should use one that they think is most suitable for them to start. Additionally, the author believes that they should give priority to the findability of words and comprehensibility of E-J dictionaries when they choose one. In that sense, teachers should try to observe dictionary use of their students during classes, find out if they are having any problems, and give appropriate advice as required.

4. CONCLUSION
E-J dictionaries have a wealth of information on English words and culture and they are indispensable tools for learning English. We are expected to use them effectively. For E-J dictionaries to be more user-friendly, this author believes that cooperation among English teachers, students and lexicographers is strongly required. English teachers are expected to have in-depth knowledge of the features in E-J dictionaries and give their students
useful advice for selecting the best one. Students are advised to choose an appropriate
dictionary, read the instructions carefully and use it as often as possible. For
lexicographers, it is important to receive feedback on their dictionaries and strive to make
them more user-friendly. This small-scale study clearly shows the importance of
cooperation between the three partners to make E-J dictionaries better tools for English
learners.

ACKNOWLEDGMENTS
I would like to express my deepest gratitude to Professor Katsumasa Yagi, who gave me
valuable advice for this paper. My appreciation also goes to the students and Ms. Kumie
Yamamoto at Sano High School, and Sylvain Primeau, who proofread the paper and made
comments on my English. Finally I would like to express my thanks to the Selection
Committee of Asialex 2009 who gave me a chance to present my research.

REFERENCES
Tokyo: Sanseido
Language Learner. Tubingen: Niemeyer: 115, 119-121, 125.
Abstract: Bilingual dictionaries are an essential, perhaps even indispensable tool in every day activities or language education. While such dictionaries are available to and from numerous widely used languages, less represented language pairs have rarely a reliable dictionary with good coverage. The need for bilingual dictionaries for these less common language pairs is increasing, but qualified human resources are scarce. Considering that in these conditions manual compilation is highly costly, alternative methods are imperative. This paper presents an alternative method for dictionary construction that is implementable with virtually any language pair. In the first step we automatically build the core of the dictionary. We perform a fully automated, robust, pivot language based bilingual dictionary generation method that uses the WordNet of the intermediate language to build a new bilingual dictionary. We propose the usage of WordNet in order to increase accuracy; we also introduce a bidirectional selection method with a flexible threshold for high recall. With the Japanese-Hungarian language pair the evaluations showed 79% accuracy and 51% weighted recall, performing best among pivot language based dictionaries. In the second step we attempt to correct the spurious entries that were generated during the automatic acquisition. This is performed by developing an online community-based interface which provides an environment for any unqualified speaker of the two languages to help and cooperate as human correctors in editing the dictionary. User reward systems and intelligent evaluation algorithms assure that the initial accuracy and recall values are in constant increase.

Key Words: bilingual dictionary generation; Hungarian; Japanese; online cooperative tool

1. INTRODUCTION
Compiling dictionaries for daily usage or for educational purpose can very costly, especially in the case of uncommon language pairs. The need for bilingual dictionaries for these less common language pairs is increasing, but qualified human resources are scarce. Considering that in these conditions manual compilation is highly costly, alternative methods are imperative.

Our paper targets bilingual dictionary generation, together with a cooperative tool designed for correcting and editing with the use of unqualified speakers of both languages. Our goal is to provide a low-cost, robust and accurate dictionary generation method. Low cost and robustness are essential in order to be re-implementable with any arbitrary language pair. For improved precision, we propose the usage of WordNet, while for good recall we introduce a bidirectional selection method with local thresholds. Since the first, English WordNet, numerous languages adopted the WordNet structure, thus creating a potential large multilingual network. We also believe that besides high precision, high recall is also crucial in order to facilitate post-editing which has to be performed by human correctors with the use of our collaborative dictionary editor.
Our paper is structured as follows: first we overview the related work, after which we present the details of our proposal, exemplified with the Japanese-Hungarian language pair. We evaluate the dictionary, performing also a comparative evaluation with two other pivot-language based methods. Next, we present iChi, the implementation of our own pivot method, an easy-to-use, customizable tool that generates a bilingual dictionary. Finally we present our conclusions.

2. RELATED WORK

2.1. Bilingual Dictionary Generation

Pivot language based bilingual dictionary generation methods rely on the idea that the lookup of a word in an uncommon language through a third, intermediated language can be automated. Bilingual dictionaries to a third, intermediate language are used to link the source and target words. The pivot language translations of the source and target head words are compared, the suitability of the source-target word pair being estimated based on the extent of the common elements.

Tanaka and Umemura’s method uses bidirectional source-pivot and pivot-target dictionaries (harmonized dictionaries). Correct translation pairs are selected by means of inverse consultation, counting the number of pivot language definitions of the source word, through which the target language definitions can be identified (Tanaka and Umemura, 1994). In case of Sjöbergh’s English pivoted Swedish-Japanese dictionary, each Japanese-to-English description is compared with each Swedish-to-English description (2005). Scoring is based on word overlap, weighted with inverse document frequency. These two approaches are the best performing ones that are robust enough to be applicable with other language pairs as well. We will use these two methods as baselines in our evaluations.

There are numerous refinements of the above methods, such as of Shirai and Yamamoto (2001), Paik et al. (2001), Bond and Ogura (2007) but for various methodological reasons they cannot be implemented with any arbitrary language pair.

There are two known problems of conventional pivot methods. First, a global threshold is used to determine correct translation pairs. However, the scores highly depend on the entry itself or the number of translations in the intermediate language, therefore there is a variance in what that score represents. Second, current methods perform a strictly lexical overlap of the source-intermediate and target-intermediate entries. Even if the translations from the source and target languages are semantically transferred to the intermediate language, lexically it is rarely the case. However, due to the different word-usage or paraphrases, even semantically identical or very similar words can have different definitions in different dictionaries. As a result, because of the lexical characteristic of their overlap, current methods cannot identify the differences between totally different definitions resulted by unrelated concepts, and differences in only nuances resulted by lexicographers describing the same concept, but with different words.

2.2. Collaborative Tools

Web-based collaborative tools are a relatively new method of sharing and improving information, with a rapidly growing tendency. Collaborative tools rely on the idea that a large number of collaborators can successfully create, manage and self-sustain a system. Owing to its simplicity it is implementable in virtually any environment, regardless of the field of application.
In dictionary editing, especially in the case of bilingual dictionaries there isn’t any acknowledged, pure collaborative tool that serves creation/editing. Most of the tools are based on an initially manually created dictionary, with the collaborators having only the possibility to add new entries, with the new data being first validated by a superior entity, or administrator of the system (Breen, 1995). This is especially needed in case of systems where the number of users is small enough to have difficulties overcoming noise generated by certain users. There are limited number of cases of administrator-free online collaborative experiments (Sztaki English-Hungarian dictionary: http://szotar.sztaki.hu/angol-magyar, experiment unpublished), but these did not prove do be self-sustainable, the system not being able to correct its own noise.

3. BILINGUAL DICTIONARY GENERATION

3.1. Translation Resources

As an example of a less-common language pair, we have chosen Japanese and Hungarian. For translation candidate generation, we have chosen two freely available dictionaries with English as the pivot language. The Japanese-English dictionary had 197282, while the Hungarian-English contained 189331 1-to-1 entry pairs. The Japanese-English dictionary had part-of-speech information as well, but to ensure robustness, our method does not use this information.

To select from the translation candidates, we mainly use WordNet (Miller et al., 1990). From WordNet we consider four types of information: sense categorization, synonymy, antonymy and semantic categories provided by the tree structure of nouns and verbs.

3.2. Dictionary Generation Method Overview

To improve recall, we introduce bidirectional selection. We can group the translations that share the same source or target entry, and set local thresholds for each head word. For example, for a source language head word entry_source there could be multiple target language candidates: entry_target1, ... entry_targetn. If the top scoring entry_targetk candidates are selected, we ensure that at least one translation will be available for entry_source, maintaining a high recall. Since we can group the entries in the source language and target language as well, we perform this selection twice, once in each direction. Local thresholds depend on the top scoring entry_target, being set to maxscore ∙ c. Constant c varies between 0 and 1, allowing a small window for not maximum, but high scoring candidates.

3.2.1. Step 1: translation candidate generation

Using the source-pivot and pivot-target dictionaries, we connect the source and target entries that share at least one common translation in the pivot language. We consider each source-target pair a translation candidate. We generated 436966 such translation candidates for Japanese-Hungarian.

3.2.2. Step 2: translation pair selection

We examine the translation candidates one by one, looking up the source-pivot and target-pivot dictionaries, comparing the translations in the pivot language. First, we perform a strictly lexical match based only on the dictionaries. Next, using information extracted from WordNet we attempt to identify the correct translation pairs.

(a) Lexically unambiguous translation pairs

Some of the translation candidates have exactly the same translations into the pivot language; we consider these pairs as being correct by default. Also among the translation candidates we identified a number of source entries that had only one target translation;
and a number of target entries that had only one source translation. Being the sole candidates for the given entries, we consider these pairs too as being correct. 37391 Japanese-Hungarian translation pairs were retrieved with this method (type A pairs).

(b) Using sense description

For most polysemous words WordNet has detailed descriptions with synonyms for each sense. We use these synonyms of WordNet’s sense descriptions to disambiguate the meanings of the common translations. For a given translation candidate \((s,t)\) we look up the source-pivot and target-pivot translations \((s \rightarrow I = \{s \rightarrow i_1, \ldots, s \rightarrow i_n\}\) and \((t \rightarrow I = \{t \rightarrow i_1, \ldots, t \rightarrow i_m\}\). We select the elements that are common in the two definitions and we look up their respective senses from WordNet. We identify the word’s senses \((\text{sns})\) comparing each synonym in the WordNet’s synonym description of the word in question with each word from the dictionary definition. As a result, we arrive at a certain set of senses from the source to pivot definitions \((\text{sns}(s \rightarrow i))\) and a certain set of senses from the target to pivot definitions \((\text{sns}(t \rightarrow i))\), the score being their Jaccard coefficient.

\[
\text{score}_{\text{B}}(s,t) = \max_{i_1 \in s \rightarrow I, t \in t \rightarrow I} \frac{|\text{sns}(s \rightarrow i) \cap \text{sns}(t \rightarrow i)|}{|\text{sns}(s \rightarrow i) \cup \text{sns}(t \rightarrow i)|}
\]

Since we don’t use the entry’s part-of-speech, all translation candidates are verified based on all four parts-of-speech available from WordNet. Since synonymy information is available for nouns (N), verbs (V), adjectives (A) and adverbs (R), four separate scores are calculated for each part-of-speech.

Scores that pass a global threshold are considered correct. 33971 Japanese-Hungarian candidates (type B translations) were selected, with these two languages the global threshold was set to 0.1. Even this low value ensures that at least one of ten meanings is shared by the two entries of the pair, thus being suitable as translation pair.

(c) Using synonymy, antonymy and semantic categories

We extend the source-to-pivot and target-to-pivot definitions with information from WordNet, thus the similarity of the two expanded pivot language descriptions gives a better indication on the suitability of the translation candidate. Using the three relations, the Jaccard coefficient of the common and total number of translations describes the appropriateness of the translation candidate.

\[
\text{score}_{\text{C,D,E}}(s,t) = \frac{|\text{ext}(s \rightarrow i) \cap \text{ext}(t \rightarrow i)|}{|\text{ext}(s \rightarrow i) \cup \text{ext}(t \rightarrow i)|}
\]

Since the same word’s translations into the pivot language also share the same semantic value, the extension with synonyms \((\text{ext}(l \rightarrow i) = \{l \rightarrow i\} \cup \text{synonym}(l \rightarrow i))\), where \(l \in \{s,t\}\) the expanded translation set should also share more common elements.

In case of antonymy, we expand the initial definitions with the antonyms of the antonyms \((\text{ext}(l \rightarrow i) = \{l \rightarrow i\} \cup \text{antonym}(\text{antonym}(l \rightarrow i)))\), where \(l \in \{s,t\}\). This extension is different from the synonymy extension, in most cases the resulting set of words being considerably larger. Together with synonymy, antonymy is also available for nouns, verbs, adjectives and adverbs, four separate scores are calculated for each part-of-speech.
Semantic categories are provided by the tree structure (hypernym/hyponymy) of nouns and verbs of WordNet. We transpose each entry from the pivot translations to its semantic category \( \exists \text{semcategory}(l \rightarrow i) = (l \rightarrow i) \cup \semcategory(l \rightarrow i), \) where \( l \in \{s,t\} \). Assuming that the correct translation pairs share a high percentage of semantic categories, the translations of semantically similar or identical entries should share a high number of common semantic categories.

The scores based on these relations highly depend on the number of pivot language translations; therefore we use the bidirectional selection method with local thresholds for each source and target head word. The thresholds were maxscore-0.9 for synonymy and antonymy; and maxscore-0.8 for the semantic. Using synonymy, 196775 candidate pairs (type C), with antonymy 99614 pairs (type D); while with semantic categories 195480 pairs (type E) were selected.

(d) Combined semantic information
A combined selection method is also performed using the separate lists generated by the synonymy, antonymy and semantic categories selection methods (type F). A multiplication factor \((m\text{factor})\) that varies between 0 and 1, awards the candidates that were selected both times during the double directional selection; and punishing when selection was made only in a single direction. \( c_1, c_2 \) and \( c_3 \) are also language dependent constants. In case of Japanese and Hungarian, these method scored best with the constants set to 1, 0.5 and 0.8, respectively.

\[
\text{score}_F(s,t) = \prod_{rel} \left( (c_1 + \max(\text{score}_{rel}(s,t))) \cdot (c_2 + c_3 \cdot m\text{factor}_{rel}(s,t)) \right)
\]

(3)

During pre-evaluation type A and type B translations received a precision score of above 75%, while type C, type D and type E scored low. However, type F translations scored close to 80%, therefore from the six translation methods presented above we chose only three (type A, B and F) to construct the dictionary, while the remaining three methods (type C, D and E) are used only indirectly for type F selection (Table 2). With the described selection methods 187761 translation pairs, with 48973 Japanese and 44664 Hungarian unique entries was generated.

3.3. Evaluation
For comparative purposes we evaluated our dictionary together with the two best performing robust method. In order to do so, we constructed two other Hungarian-Japanese dictionaries using the methods proposed by Tanaka & Umemura and Sjöbergh, using the same source dictionaries.

3.3.1. Recall evaluation
We use a Japanese frequency dictionary \((F_p)\) that we generated from the Japanese EDR corpus (Isahara, 2007) to weight each Japanese entry. Setting the standard to the frequency dictionary (its recall value being 100), we automatically search for each entry \( w \) from the frequency dictionary verifying whether it is included or not in the bilingual dictionary \((W_p)\). If it is recalled, we weight it with its frequency from the frequency dictionary.

\[
\text{recall}_w = \frac{\sum w \in W_p \text{frequency}(w)}{\sum w \in F_p \text{frequency}(w)} \cdot 100
\]

(4)
Our method’s dictionary (51.68) outscores every other automatically generated method’s dictionary (37.03, 30.76) with a significant advantage. Moreover, it maintained the score of the initial translation candidates, therefore managing to maximize the recall value, owing to the bidirectional selection method with local thresholds.

### 3.3.2. 1-to-1 precision evaluation

We evaluated 2000 randomly selected translation pairs, from each of the three dictionaries, manually scoring them as correct (‘the translation conveys the same meaning, or the meanings are slightly different, but in a certain context the translation is possible’), undecided (‘the translation pair’s semantic value is similar, but a translation based on them would be faulty’) or wrong (‘the translation pair’s two entries convey a different meaning’). Our method performed best with an average of 79.15%, outscoring Tanaka method’s 62.50% and Sjöbergh method’s 54.05% (Table 1).

<table>
<thead>
<tr>
<th>Method</th>
<th>Evaluation score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>correct</td>
</tr>
<tr>
<td>our method</td>
<td>79.15%</td>
</tr>
<tr>
<td>Sjöbergh method</td>
<td>54.05%</td>
</tr>
<tr>
<td>Tanaka method</td>
<td>62.50%</td>
</tr>
</tbody>
</table>

### 3.3.3. 1-to-multiple evaluation

With 1-to-multiple evaluation we quantify the true reliability of the dictionary: when looking up the meanings or translations of a certain keyword, the user, whether he’s a human or a machine, expects all translations to be accurate. We evaluated 200 randomly selected Japanese entries from the initial translation candidates, together with all of their Hungarian translations, scoring them as correct (‘all translations are correct’), acceptable (‘the good translations are predominant, but there are up to 2 erroneous translations’), wrong (‘the number or wrong translations exceeds 2’) or missing (‘the translation is missing’). Our method scored 71.45%, outperforming Sjöbergh method’s 61.65% and Tanaka method’s 46.95% (Table 2).

<table>
<thead>
<tr>
<th>Method</th>
<th>Evaluation score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>correct</td>
</tr>
<tr>
<td>our method</td>
<td>71.45</td>
</tr>
<tr>
<td>Sjöbergh method</td>
<td>61.65</td>
</tr>
<tr>
<td>Tanaka method</td>
<td>46.95</td>
</tr>
</tbody>
</table>

### 3.4. Discussion

We managed to maximize the recall of our initial translation candidates, but in many cases certain translation pairs still could not be generated because the link from the source language to the target language through the pivot language simply doesn’t exist. The main reasons are: the entry is missing from at least one of the dictionaries; translations in the pivot language are expressions or explanations; or there is no direct translation or link between the source and target entries. The entries that could not be recalled are mostly expressions, rare entries, words specific to a language, certain part-of-speech.

With words whose senses are clear and mostly preserved even through the pivot language, most of the correct senses were identified and correctly translated. Nouns, adjectives and adverbs had a relatively high degree of accuracy. However, verbs proved to be
the most difficult part-of-speech to handle. Because semantically they are more flexible than other part-of-speech, and the meaning range is also highly flexible across languages, the identification of the correct translation is increasingly difficult.

The source dictionaries themselves sometimes lacked accuracy, containing a great number of irrelevant or low usage translations, shadowing the main, important senses of some words. In other cases the resource dictionaries don’t contain translations of all meanings; homonyms are present as pivot entries with different meaning, sometimes creating unique (type A), but faulty links.

4. ICHI: BILINGUAL DICTIONARY GENERATOR

ichi is the implementation of our bilingual dictionary generation method (http://mj-nlp.homeip.net/ichi). Programmed in Java, it is a platform-independent tool with a graphical interface (Image 1). Besides using MySql, it consists of iChi.jar (java executable), iChi.cfg (configuration file), iChi.log (log file) and iChip.jar (parameter estimation tool).

4.1. Main Functions

(1) Resources (①)
The location of the pivot language’s WordNet and the dictionary locations need to be specified. All locations are stored in the configuration file. The dictionaries are simple text files, with one pair in each line:

source entry 1@pivot entry 1
source entry 2@pivot entry 2
...

(2) Parameter settings (②)
iChip.jar estimates language dependent parameters needed for the selection methods. Its argument is a text file that contains marked correct translation pairs:

$+source entry 1@correct target entry 1
$-source entry 2@incorrect target entry 2
...


(3) **Save settings**

The generated source-target dictionary is saved into a text file that uses the same format described in 4.1.(1). The output can be customized by choosing the desired selection methods. The default value is a dictionary with selection types A, B and F; selection types C, D and E are used only indirectly for type F.

(4) **Tasks**

The tasks are run sequentially, every step being saved in the internal database, along with being logged on the screen (⑤) and into the log file.

5. COMMUNITY-BASED DICTIONARY EVALUATOR

5.1. **Purpose and Main Functions**

Our dictionary generation method’s initial target was machine translation resource generation; with the community-based dictionary evaluator first and foremost we intended to evaluate our Hungarian-Japanese dictionary. We also attempted to determine whether such a dictionary is useful or not for regular users as well. As a result, we implemented a very basic framework that serves as an evaluation system designed to confirm our internal evaluations (http://mj-nlp.homeip.net/mjszotar).

Qualified evaluators for a rare language pair, such as Hungarian and Japanese are scarce, therefore we targeted both languages speaking, but unqualified active users to evaluate the dictionary. Besides evaluation, for passive users the system’s major functions are browsing (magyar→japán: ‘Hungarian to Japanese’ ①; japán→magyar: ‘Japanese to Hungarian’ ②); search (‘keresés’: Japanese or Hungarian; full or partial matches ③); addition (‘adatbevitel’ ⑤) (Image 2). Users are also able to evaluate the translation pairs resulting by the browsing or search functions. The purpose and usage of the dictionary are explained in the main page (‘kezdőoldal’).

Since this system is evaluation oriented, we attempt to determine the general accuracy of the dictionary, rather than evaluating and thus correcting each entry in part. Thus purposes random evaluation was made available for evaluation validation under evaluation.
(‘értékelés’ (4)), where with each access 50 randomly selected translation pairs are prepared for manual verification. As explained in the user guide (9), all translations are to be marked as correct (√), undecided (?) or wrong (×). Potential comments regarding the translation pairs, personal criteria, etc. can be also transmitted through the evaluation process (7). Evaluation is completed after filling out the name of the evaluator and pushing the save (‘ment’ (10)) button. Previous evaluation results are shown next to the translation pair (8).

Image 2: Community-Based Dictionary Evaluator: Random Translation Pair Verification Page

### 5.2. Results and Evaluation

Starting from the first year of operation (2007.11.27) to the current date (2008.5.30) we registered 106553 unique visits with 28770 searches, 128 new entries and a total of 2829 evaluations, proving its usefulness for human users as well.

For precision evaluation we considered only the 1143 entries that were evaluated through the random evaluation page (4). Other entries, evaluated through the browsing (1,2) or search (3) functions or the dictionary were selected by choice, not by random, thus not being statistically correct. The precision value resulted by the online evaluation (51.09%) is considerably lower than what we calculated during the 1-to-1 precision evaluation (Table 3).

<table>
<thead>
<tr>
<th>Evaluation period</th>
<th>Online evaluation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>correct</td>
<td>undecided</td>
</tr>
<tr>
<td>2007.11.27. – 2008.3.4.</td>
<td>52.13% (256)</td>
<td>32.17% (158)</td>
</tr>
<tr>
<td>2007.11.27. – 2008.9.30.</td>
<td>55.89% (555)</td>
<td>25.78% (256)</td>
</tr>
<tr>
<td>2007.11.27. – 2009.5.30.</td>
<td>51.09%</td>
<td>33.68%</td>
</tr>
</tbody>
</table>
5.3. Discussion
The community-based dictionary evaluator had two major problems: the online evaluation did not achieve its goal; the evaluation system failed to keep its active users.

There are two main reasons for the large differences between our internal and the online evaluation. First, the data can not be considered statistically correct. Although the online system provides a fixed number of randomly selected entry pairs, the users mostly evaluate only a certain number of them, the evaluation becoming one by choice rather by random. The second reason is the questionable reliability of the evaluators. When investigating the labels set by our human evaluators, we could found a large number of translation pairs that are obviously correct, but they were labelled as incorrect (ex: 変える (kaeru) – kicserél: 'to change'; 年越し (toshikoshi) – szilveszter: 'new year’s eve’). The evaluators’ reliability, combined with the low number of evaluations (2829) and high number of entries (187761) resulted in the system failing as a self-sustainable system, the reliable active users not being able to overcome the noise created by faulty evaluations.

The number of evaluations (2829) registered over the 18 months is low, one reason being the low number of speakers of both languages. The number of returning active users was also low, the dictionary receiving the majority of the evaluations in the first half of the 18 month period, thus the system failing to keep its active users.

However, considering the accesses count and the consistency of returning passive visitors, the evaluation system did succeed in proving that an automatically generated dictionary can be useful for regular users as well.

6. COMMUNITY-BASED DICTIONARY EDITOR
6.1. Purpose and Main Functions
To overcome the inefficiencies of the community-based dictionary evaluator, we developed a prototype system. Whereas the purpose of the community-based dictionary evaluation system was mainly evaluation, the dictionary editor serves as a collaborative tool for improving the dictionary’s content. We propose a number of changes that are designed to abolish the weak points of our previous evaluation system.

6.6.1. Registration needed for active users
Active users, who desire to cooperate in the editing of the dictionary, are required to register. The reason for this is two layered: first, most potentially mischievous users will abandon the idea of completing the process of registration; second, this way it is easier to control any low-rated collaborator. For basic functions, such as browsing or search, no registration is required.

6.6.2. Collaborator rating
To overcome the un-sustainability of the previous system, noise generated within the system needs to be eliminated by the system itself. In order to do so, we introduce a collaborator rating system that rates the collaborator’s evaluation accuracy. The rating is used as a weight in all evaluations of the collaborator: high-rated collaborators adding a bigger weight to each evaluation compared to lower-rated collaborators. The rating is the combination of two coefficients: accuracy coefficient $c_1$ reflects the accuracy of the collaborator’s evaluations, while tendency coefficient $c_2$ shows the tendency of evaluating with the majority.
The accuracy coefficient is separately calculated for each evaluated entry. The score is the combination of the downscaled percentage ratio between the majority ($score_{majority}$) and minority ($score_{minority}$), versus total score of already existing evaluations (total accuracy coefficients of the collaborators to the given entry), and the number of voters for ($n^+$) and against ($n^-$) the entry. If the entry is evaluated for the first time, $c_1$ is 1.

$$c_1 = \begin{cases} 
1, & \text{if } n^+ + n^- < 2 \\
\pm \frac{10}{n^2} \sum_{i} score_{majority} - \sum_{i} score_{minority}, & \text{if } n^+ + n^- \geq 2
\end{cases}$$

For example, users $u_1$, $u_2$ and $u_3$ voted the entry ($n^+$) as correct, while user $u_4$ voted as wrong ($n^-$). The total accuracy coefficients for these users are 100, 170, 180 and 210, thus the total score for the entry is 350 ($score_{majority}$), while against it is 210 ($score_{minority}$). As a result, the precision index for the entry itself is $68.18\%$ for and $31.81\%$ against, resulting in a correct translation pair. The total reward for the supposedly correct evaluations is the downscaled ratio of the percentage difference: $(68.18 - 31.81) \times 10 = 3.63$. Users $u_1$, $u_2$ and $u_3$ share the reward, scoring $3.63 / 3 = +1.21$ for this entry ($n^+ = 3$), while $u_4$ is penalized by $- (3.63 / 1) = -3.63$, being the sole voter against the entry ($n^- = 1$).

The tendency coefficient is used to identify collaborators who attempt to artificially inflate their rating. Voting constantly with the majority steadily increases the accuracy index, which might be a tool for mischievous behaviour. $c_2$ is the weighted ratio ($a=5$) of the votes with for and against the majority. Evaluations without other collaborators are not counted. For example, voting always with the majority results a tendency coefficient value of 1, while keeping a well balanced evaluation tendency of 300 with the majority and 200 against it results in a value of $(5 \times (300 - 200) + 500 + 1) = 5$.

$$c_2 = a \frac{\sum_{i} c_{i, majority} - \sum_{i} c_{i, minority}}{\sum_{i} c_{i, majority} + \sum_{i} c_{i, minority}} + 1$$

At registration, an initial value accuracy coefficient value (100) is provided for the collaborator. For each login, a certain amount is added to it (1 point per day), encouraging activity. It is also possible to reduce the influence of low rated collaborators by not considering their votes if their accuracy or tendency coefficients are below some pre-defined thresholds.

### 6.6.3. Improved usability

Besides the functions already existing in the evaluation system (browsing, search and addition) we added new functions that are designed to increase the usability of the system, and thus creating a self-sustainable environment. Periodical and overall ranking lists of the collaborators with high activity and high overall rating are available under top collaborators (’toplisták’ २). A discussion forum (’fórum’ ३) for debating is also available (Image 3). Accessing the forum is possible for unregistered, passive users as well, for writing rights registration is required. Together with registration (३) a personal page
(‘adatlap’ (4)) becomes available, detailing personal ratings and collaboration details (evaluated or newly added entries) of the collaborator. The evaluation results also are more recognizable, showing both the precision index (5) and the total number of evaluators (6). A combination of high precision index with high number of evaluators should indicate a correct translation pair. A link to the forum thread about the entry in question is available, together with evaluation details, such as collaborators and evaluation timestamps (8). The evaluation method was simplified, reducing the criteria to correct (✓) and wrong (✗) (7), available only for registered users. Evaluation was extended with two new functions: evaluation of recent evaluations; evaluations of entries performed by a certain collaborator (4). Artificially inflated ratings or newly added contributions can be overcome this way.

Image 3: Community based dictionary editor: random translation pair verification page

7. CONCLUSIONS
We proposed a new pivot language based method to create bilingual dictionaries, which is robust enough to be implementable with any language pair. Opposed to conventional methods we use WordNet as a main resource of the pivot language to select the suitable translation pairs. As a result, we eliminate most of the weaknesses caused by the structural differences of dictionaries, while profiting from the semantic relations provided by WordNet. Finally we presented our online collaborative prototype system designed for correcting and editing a bilingual dictionary with the use of unqualified speakers of both languages.

8. REFERENCES
PAKISTAN NATIONAL LANGUAGE AUTHORITY AND ITS CONTRIBUTION IN LEXICOGRAPHY

Yasmeen Zafar
University of Management and Technology, PAKISTAN
Yasmeenzafar6@yahoo.com

It is not uncommon in the history of many countries to set up academies of societies to deal with the matters of linguistic diversities of regional languages or the promotion of national language in the country. The establishment of nation language authority in Pakistan is to be looked at from this perspective. The aim of this paper is to highlight the contribution made by the Urdu language authority particularly with reference to compiling various kinds of dictionaries in the country.
AN EMPIRICAL STUDY OF ‘ENGLISH TO ENGLISH & URDU’ SEMI-BILINGUAL DECODING DICTIONARIES USED BY ADVANCED LEARNERS IN PAKISTAN

Rana Faqir Muhammad Aslam
Bahauddin Zakariya University, PAKISTAN
ranabzu@gmail.com

Abstract: In Pakistan, learners at the advanced level use various types of dictionaries for decoding purposes. Out of which, the monolingual ‘L2-L2’ and the bilingual ‘L2-L1’ dictionaries pose certain problems for the learners due to cultural differences between the speakers of English and Urdu languages. In this scenario, a perfect semi-bilingual ‘L2-L2-L1’ (English to English & Urdu) dictionary is needed to bridge the cultural gap. This paper presents the results of the research into dictionary use by the advanced learners conducted in Pakistan on a sample of 30 students from the Department of English Bahauddin Zakariya University, Multan. The present research is empirical in nature and the data have been collected through surveys using the tool of questionnaire. The collected data have been analysed applying the qualitative methods of data analysis. Along with the focus on the use of dictionary for decoding purposes, the paper also presents analysis of the design features of semi-bilingual ‘L2-L2-L1’ dictionaries available in Pakistan. The study comes up with the results indicating that the ‘L2-L2-L1’ dictionaries presently available in Pakistan provide insufficient information and do not fulfill the needs of the advanced learners comprehensively. In the final part of this paper it is recommended that there is a dire need to develop a model which could provide a guideline for compiling a perfect ‘English to English & Urdu’ semi-bilingual dictionary in Pakistan. The dictionary compiled following this model is expected to remove the problems posed by both the monolingual and bilingual dictionaries.

Key Words: Semi-bilingual dictionary, language learning, decoding purposes, cultural gap.

1. INTRODUCTION

In Pakistan, mainly two approaches are followed for teaching and learning of English as a Foreign Language. One of them is the monolingual approach which is exercised by the use of Direct Method for teaching English, and the other is the bilingual approach practiced by the use of Grammar Translation Method. Learners of English consult either L2-L2 (English to English) monolingual dictionary or L2-L1 (English to Urdu) bilingual dictionary for decoding purposes depending on the particular method followed by them. Both the monolingual and bilingual dictionaries have certain drawbacks which pose problems for the learners. The monolingual dictionaries provide information without keeping the cultural context of a particular non-native learner of English. As a result the learners feel alienated from their own culture. On the other hand, in bilingual dictionary, the focus is the indigenous culture without giving much consideration to the native English culture. Both these situations create a gap for the learner and he/she does feel at ease while using these either L2-L2 or L2-L1 dictionaries. In this situation, a third approach is developing as semi-bilingual approach which attempts to fill the existing gap created by the monolingual and the bilingual dictionaries. As the semi-bilingual
approach is a middle way round and in this scenario, so a perfect semi-bilingual ‘English to English and Urdu’ dictionary is needed to bridge the gap perfectly.

The present paper consists of two parts. The first part of this paper is concerned with the advanced learners’ choice for a particular dictionary and the ratings given by them to the different types of dictionaries used by them for decoding purposes. In the second part the design features of the L2-L2-L1 i.e. ‘English to English and Urdu’ semi-bilingual dictionaries available in Pakistan are discussed.

1.2 Research Questions

The research reported in this paper tries to find the answers to the following questions:

1. Which type dictionary is favoured by advanced learners for decoding purposes? Why?
2. Which dictionary types are of greatest benefit to learners at the advanced level?
3. What are the merits and demerits of the existing semi-bilingual dictionaries?
4. How can the design features of the existing dictionaries be improved to develop a perfect semi-bilingual dictionary of international standard in Pakistan?

2. REVIEW OF THE RELATED LITERATURE

Dictionaries play a significant role in learning a foreign language and they are frequently used by the learners for both decoding and encoding purposes. These are considered to be a basic tool in the process of foreign language learning, and it appears that ‘the conviction of the usefulness of dictionaries is common among lexicographers, as well as language learners themselves’ (Lew, 2004).

Most of the previous studies (e.g. Béjoint 1981; Tomaszczyk 1979; Kharma 1985) into dictionary use have been concerned with either monolingual or bilingual dictionary ratings. However, there are a few studies (e.g. Lew, 2004; Laufer and Melamed (1994) which have been conducted into the dictionary use concerning three different types of dictionaries i.e. monolingual, bilingual and bilingualised or semi-bilingual.

The research into dictionary use directly and indirectly started as early as in the 1960’s. The first significant study in this context was done by Barnhart (1962), which is regarded an important landmark in this field. It was based on teachers’ opinion about what they considered their students did as dictionary users. However, it did not involve dictionary users themselves. The research which directly involved actual dictionary users was conducted by Tomaszczyk (1979), who was the first researcher to investigate the dictionary needs of non-native speakers of English. He found that bilingual dictionaries were favoured exclusively by the participants.

There has been found a significant growth in research on dictionary use since the 1980s, out of which I would like to mention a few studies. The study conducted by Ard (1982) involved students’ video recorded recollections of their use of bilingual dictionaries and indicated that the use of bilingual dictionary, with few exceptions, causes certain lexical errors. Hartmann (1983) studies the use of bilingual dictionaries by the learner of German in South West England. He found that the users did not have adequate training in dictionary use.

Hatherall (1984) observed the dictionary users in action. He found the protocol
method to be the best method of data collection in dictionary use. The research on dictionary in Pakistani context has its beginning in the 1980s. The pioneer work in this context was carried on by Iqbal (1987). In his study, he evaluated the reference skills and the language needs of Pakistani advanced learners. He proposed to provide separate guides for dictionary use and also suggested the design features for a dictionary to be designed for Pakistani learners of English.

Diab (1990) conducted research into the dictionary use by the Jordanian nurses. His research is unique of its type as it has been one of the pioneer works in dictionary use in the perspective of ESP (English for Specific Purposes). Then a series of publications appeared in Lexicographica reviewed by Dolezal and McCreary (1999).

Laufer and Melamed (1994) studied the differences in the effectiveness of three types of dictionaries; monolingual, bilingual and bilingualised on the comprehension and production of 15 low frequency words in English by EFL learners. They found that the bilingualised dictionary was significantly more effective than the other two.

Tono (2001) conducted a series of researches on dictionary use in the context of foreign language learning. One of the main themes of his research was to explore the cognitive aspect of dictionary reference skills.

In 2004, Robert Lew conducted a study on a broad sample of non-native English language learners as dictionary users. This study was predominately exploratory in nature which was conducted on Polish learners of English. It was carried on to find out the patterns that might throw some light on the receptive dictionary use by language learners.

The story has not stopped there. A number of researches have been appearing the International Journal of Lexicography. The review of all these researches at this place is not a possibility.

3. RESEARCH METHODOLOGY

The present paper is the outcome of the pilot study of my Ph. D project on semi-bilingual dictionaries. It is based on an empirical study conducted on a sample of 30 students from the department of English, Bahauddin Zakariya University, Multan Pakistan. All the participants are the students of masters in English language and linguistics. The sample is made on random basis. I have used the tool of questionnaire to collect data from the participants. The participants were asked to give their ratings for three different types of dictionaries, i.e. L2-L2, L2-L1 and L2-L2-L1 dictionaries. The data collected through questionnaire have been analysed qualitatively, however, quantitatively analysis have also been given where necessary. In order to provide the comparative differences more clearly, the results are shown in graphs as well.

Since no single method is without its problems, [and] restricting dictionary use research to just one specific method or technique is not the best policy to follow (Lew, 2002), so in order to compensate the limitations of questionnaire and to get more authentic and reliable results, I have conducted the focus group interview. This focus group comprises of five students from the same department. All of them are the students of Ph.D. (Linguistics). The interview conducted from them was unstructured in its design and the participants were given free hand to talk on the use of dictionaries for decoding purposes. The open discussion in this regard has been analysed qualitatively as the qualitative analyses are considered to be the most favoured one in social science.
The paper also presents the analysis of the design features of semi-bilingual dictionaries currently available in Pakistan. For this purpose, I have selected the most popularly known and easily available dictionaries in Pakistan. These include the following dictionaries:

1. Kitabistan’s New Millenium Practical English to English and Urdu Dictionary
2. Ferozson’s English to English and Urdu Dictionary
3. Gem Practical English to English and Urdu Dictionary
4. Rabia 21st Century Practical English to English and Urdu Dictionary

The choice of the above given dictionaries is based on my seventeen years personal experience as a teacher and more than that as a learner of English. This is also supported by this research as the participants of this research have also favoured these dictionaries and many of them have these in their personal possession. The analyses of design features include both the micro and macro structure of the dictionary.

4. ANALYSIS

The analysis are provided in two main parts: one is concerned with the use and ratings of monolingual, bilingual and semi-bilingual dictionaries as perceived by the advanced learners; the other is concerned with the comparison and evaluation of the four semi-bilingual dictionaries most commonly used in Pakistan.

4.1 Monolingual vs Bilingual vs Semi-bilingual Dictionaries

The first part of the questionnaire used for this research required the participants to compare three types of dictionaries, i.e. English-English, English-Urdu and English-English-Urdu. The results show that the learners favour the semi-bilingual dictionaries almost exclusively.

Out of the thirty participants, twenty three own semi-bilingual i.e. English to English & Urdu dictionary, while fifteen possess monolingual i.e. English to English dictionary and eleven participants own bilingual i.e. English to Urdu dictionary. The participants have been using semi-bilingual dictionary for a longer period of time in comparison to the use of bilingual or monolingual dictionaries.

The questionnaire contained a question related to the frequency of use. They were asked to provide information about how frequently they used a particular dictionary for decoding English text. The comparison of three dictionaries indicates that the participants use semi-bilingual dictionary more frequently than they use the monolingual or bilingual dictionaries. The comparison is shown in Fig.1 below:
The graph shows that 10% of the participants use English to English dictionary daily, 10% of them use it twice a week, 16.7% use it once a week, while 30% of them use it less frequently and 33.3% of them do not use monolingual dictionary at all. As far as the use of English to Urdu dictionary is concerned, 6.7% of the participants use it on daily basis, 13.3% use it twice a week, 23.3% use it once a week and 26.7% use it less frequently while 30% of them do not use bilingual dictionary at all.

The case is different as far as the frequency of use of semi-bilingual dictionary is concerned. Here the participants have opined that they use this dictionary more frequently than they do any other dictionary. 26.6% of them use it on daily basis, 30% use it twice a week, 16.7% use it once a week and 13.3% use it less frequently, while only 13.3% of them do not use English to English & Urdu dictionary at all.

When asked to compare and rate the usefulness of the three dictionaries, the majority of them gave their opinion in favour of semi-bilingual dictionary with the monolingual dictionary being the second in the comparison. The results are shown in the following figure:

---

**Fig. 1. The frequency of use of Eng-Eng, Eng-Urd and Eng-Eng-Urd dictionaries**

The graph shows that 10% of the participants use English to English dictionary daily, 10% of them use it twice a week, 16.7% use it once a week, while 30% of them use it less frequently and 33.3% of them do not use monolingual dictionary at all. As far as the use of English to Urdu dictionary is concerned, 6.7% of the participants use it on daily basis, 13.3% use it twice a week, 23.3% use it once a week and 26.7% use it less frequently while 30% of them do not use bilingual dictionary at all.

The case is different as far as the frequency of use of semi-bilingual dictionary is concerned. Here the participants have opined that they use this dictionary more frequently than they do any other dictionary. 26.6% of them use it on daily basis, 30% use it twice a week, 16.7% use it once a week and 13.3% use it less frequently, while only 13.3% of them do not use English to English & Urdu dictionary at all.

When asked to compare and rate the usefulness of the three dictionaries, the majority of them gave their opinion in favour of semi-bilingual dictionary with the monolingual dictionary being the second in the comparison. The results are shown in the following figure:
The usefulness of Eng-Eng, Eng-Urd and Eng-Eng-Urd dictionaries

The graph shows that 16.7% and 13.3% of the participants have found English to English dictionary to very useful and useful respectively, while 16.7% have found it just OK, and in the opinion of 26.7% participants monolingual dictionary is less useful or absolutely useless. On the other hand, the opinion about the English to Urdu dictionary is that the 6.7% of the participants believe that it is very useful and 20% think that it is useful, 13.3% find it to be OK, while 33.3% find it less useful and 26.7% consider it to be useless.

Majority of the users have given their opinion in favour of semi-bilingual dictionary as 33.3% of them find it very useful, 30% find it useful and 13.3% find it just OK, while 13.3% think that it is less useful and only 10% believe it to be useless.

The results obtained through the questionnaire are further strengthened by the unstructured interview conducted with a focus group. Since the participants of the focus group are doing their Ph.D. they have more experience as compared with the participants of questionnaire. Further, four out of five of them are also the practicing teachers of English. They gave their opinion on the basis of their experience as the learners and teachers of English as a foreign language.

The participants believe that the monolingual dictionaries alienate the learners from their own language and, sometimes, they feel so fed up that they do not want to use a dictionary at all and prefer simply to guess the meaning. It was opined by one of the participant that the compilers of monolingual dictionaries do not have a particular non-native learner in mind and the meanings given in the dictionaries are not clear for the users in his/her own context, for example the word ‘bed’ is used in context in Pakistani and English cultures, and English does not have an equivalent for Urdu ‘Charpai’.

4.2 The analysis of the design features of Semi-bilingual dictionaries

Twenty-six of the participants of this research have reported that they use English to English and Urdu dictionary. They were asked to provide the name of the dictionary used by them. The obtained results are given in the following table:
Table: 1. Dictionary used

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>4</td>
<td>13.3</td>
<td>13.3</td>
</tr>
<tr>
<td>Ferozson</td>
<td>8</td>
<td>26.7</td>
<td>40.0</td>
</tr>
<tr>
<td>Gem</td>
<td>3</td>
<td>10.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Kitabistan</td>
<td>12</td>
<td>40.0</td>
<td>90.0</td>
</tr>
<tr>
<td>Rabia</td>
<td>3</td>
<td>10.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

As the table shows, the maximum number (N=12) of participants are using Kitabistan’s Dictionary. This makes 40% of the total. Then there is Ferozson’s dictionary which is used by 26.7% of the participants which is a good number (N=8) of users. The other two dictionaries used by the participants are (a) Gem dictionary and (b) Rabia dictionary. These are used by a small number (N=3) of participants which makes 10% each of the total number of participants.

The users’ opinion about how they find their dictionary was also sought in the questionnaire. Keeping this in view and on the basis of personal observation of the design features, the analysis of the selected semi-bilingual dictionaries is given in the following:

4.2.1 Kitabistan’s New Millennium Practical English to English and Urdu Dictionary

Kitabistan’s dictionary has proved to be the most favoured dictionary among the participants. Its macro and micro structures are discussed below:

**Macrostructure:**

This is the only dictionary out of the four, which provides the publishing date. In its front matter, the dictionary contains the following:

a) Preface

b) Foreword

c) Contents page

d) List of abbreviations used in the dictionary

e) Get to know your dictionary (contains general information how to use it)

f) Key to pronunciation

There are certain drawbacks observed in this part of the dictionary. Although it provides a list of abbreviations used, but it is not a comprehensive list since some of the important abbreviations used in the microstructure are missing in this list, for example, the dictionary uses *n* to indicate the *noun*, and *pr* to indicate the *preposition*, while both of these abbreviations are missing in the provided list.

Similarly the Key to pronunciation does not follow any regular pattern. It ignores the IPA (International Phonetic Alphabet). It is a combination of Fowler, Chamber and
Daniel Jones as claimed by the dictionary. Furthermore, the provided phonemes are confusing and incomplete. The users also find it difficult to get clear guidance of pronunciation.

The dictionary contains the following in its back matter:

- a) Notable names and characters
- b) Names of countries with adjectives formed from them
- c) Christian names and their meanings
- d) Abbreviations
- e) Stems and affixes
- f) Signs and symbols
- g) Metrical chart
- h) World time differences
- i) Names of continents

All the above given information is provided in English language and these are not translated into the target user’s native tongue. This creates problems for the learner since they expect from a semi-bilingual dictionary that it should provide information in Urdu language as well.

**Microstructure**

The microstructure of the dictionary contains the following information:

- a) Lemma (the headword given in bold print)
- b) The pronunciation of the word is provided in brackets ( ) after the lemma
- c) Part of speech
- d) Meaning in English with some paraphrasing
- e) Meaning in Urdu

The dictionary ignores the aspects related to etymology (historical origin and derivation) of the word. It also does not provide citation i.e. the examples of the use of word. The meanings of the words related to lemma are not treated as separate head words, and are given under the same head word, although these are also printed in bold.

The participants rate it as a good dictionary. It is regarded as a reliable dictionary which contains thorough and objective definitions. Most of the users were recommended this dictionary by their teachers.

**4.2.2 Ferozson’s English to English and Urdu Dictionary**

Ferozson’s dictionary is the second most favoured dictionary among the users. Its macro and micro structures are analysed in the following:
**Macrostructure:**
This dictionary has the weakest macrostructure. It does not contain any front matter at all. So there is no pronunciation key and no list of abbreviations used in the dictionary. Dictionary does not provide its publishing date. The back matter of the dictionary contains the following:
   a) Abbreviations
   b) The metric system

**Microstructure:**
In its microstructure, the dictionary provides the following information:
   a) Lemma
   b) Pronunciation of the word is provided in brackets ( ).
   c) Part of speech
   d) Meanings in English
   e) Meanings in Urdu

The dictionary ignores the etymology and does not give examples of the use of words. The pronunciation does not follow a regular pattern and because of the non-availability of key to pronunciation, the users find it very difficult to use it for seeking pronunciation. Same is the case with the use of abbreviations. As far as the meanings are concerned, this dictionary provides one word equivalents and does not bother in paraphrasing. The words related to lemma are also treated as head words, and are given in bold print.

The participants rate this dictionary to be OK. They do not regard it to be either excellent or awful. The reason for the liking of this dictionary is the long standing repute of the publisher. According to the participants, this dictionary is mostly recommended by the parents.

**4.2.3 Rabia 21st Century Practical English to English and Urdu Dictionary**

Although it is not much used by the participants, the macrostructure of the dictionary is found to be the best of all. The detailed analysis is provided in the following:

**Macrostructure:**
The front matter of the dictionary contains the following in its macrostructure:
   a) Preface
   b) How to look up a word
   c) Contents
   d) Sample page
   e) Alphabetical list of abbreviations
   f) Pronunciation key
The back matter of the dictionary contains the following information:

a) Better English; this part contains detailed description of punctuation marks and their usage.
b) Common abbreviations
c) Weights and measures
d) The metric system
e) World time differences
f) Continents list
g) Notable mountain peaks of the world

**Microstructure:**

The dictionary contains the following in its microstructure:

a) Lemma (in bold print)
b) The pronunciation of the word is provided in brackets ( ) after the lemma
c) Part of speech
d) Meanings in English with some paraphrasing
e) Meanings in Urdu

The dictionary does not give information related to etymological aspects of the word. It also does not provide the examples of the use of word. The meanings of the words related to lemma are treated as separate head words.

This dictionary is regarded to be very good by its users. This dictionary provides maximum information in its macrostructure. However, the main flaw of this structure is that it gives whole of the information in English language which is the cause of difficulty for the learners. The guide to use this dictionary is a good addition. However, there are certain problems with the key to pronunciation. It is devised without keeping in view the IPA, so it may not be systematic and the users feel problems in seeking pronunciation. Most of the users were recommended this dictionary by the book sellers.

**4.2.4 Gem Practical English to English & Urdu Dictionary (21st Century edition)**

This dictionary is also not much used by the participants as only 10% of them have reported that they use this dictionary. The detailed analysis is provided in the following:

**Macrostructure:**

The front matter of the dictionary contains the following information:

a) Preface
b) Alphabetical list of abbreviations used

The back matter of the dictionary contains the following information:

a) Appendix I (The World states, capitals and main cities)

b) Appendix II (Signs and symbols used in Astrological, Science, Commerce, Medicine, mathematics, and Typography with proof reading abbreviations)

c) Appendix III (The planets)

d) Appendix IV (Rules for the use of capital letters)

e) World Time

f) Tables of weights and measures

g) The metric system

**Microstructure:**

The dictionary contains the following in its microstructure:

c) Lemma (in bold print)

d) The pronunciation of the word in brackets ( )

e) Part of speech

f) Meanings in English with some paraphrasing

g) Meanings in Urdu

h) Examples of use (in some cases)

Like the other three dictionaries, this dictionary, too, does not give information related to etymological aspects of the word. Although it provides examples of the use of word, but it is limited to certain words only. The meanings of the words related to lemma are treated as separate head words.

As far as the users are concerned, they have graded this dictionary to be a good one. Although, this dictionary provides maximum information in its back matter of the macrostructure, yet it has a major flaw that it does not give any key to the pronunciation symbols used in its microstructure. So the users face problems in consulting it for pronunciation. Most of the users were recommended this dictionary by the book sellers.

5. CONCLUSION

In the light of the research carried on, it has been found that the problems caused by the L2-L2 (English-English) and L2-L1(English-Urdu) dictionaries can be eradicated by the use of L2-L2-L1 (English-English-Urdu) dictionaries. In this way semi-bilingual dictionary becomes the need of the hour in an EFL situation like Pakistan.

On the other hand, the study of existing semi-bilingual dictionaries have suggested that they are not compiled according to the international standard and do not fulfill the decoding needs of the advanced learners, it is recommended that a model should be
developed which could guide the lexicographers to compile a perfect semi-bilingual dictionary. The proposed dictionary will solve the existing problems and will fulfill all the needs of the learners at advanced level in Pakistan. It is worth mentioning here, that I am working on the proposed project which is also the part of my Ph.D. research. The proposed dictionary will be useful not only for the advanced learners inside Pakistan, but can be used by Urdu speaking community outside Pakistan as well.

References


KAMUS BESAR BAHASA INDONESIA (INDONESIAN BIG MONOLINGUAL DICTIONARY): A USER ORIENTED STUDY

Dora Amalia
Departement of National Education, INDONESIA
yayak_amalia@yahoo.co.id

Kamus Besar Bahasa Indonesia (KBBI) is the most comprehensive monolingual dictionary which was compiled by a lexicographic team of Language Center and recently launched to public. It is a revised edition of the previous one and has been modified macrostructurally and microstructurally as well. During compilation process, the team has created an on-line discussion group to receive feedback from all users. The users’ input can be categorised into four main aspects, namely the users’ needs, their reference skills, their expectations, and their usage situation. These aspects have a definite influence in compiling strategies including the design, structure and contents of the dictionary. The team continues the same method until now, besides the dictionary review held by the publisher after the launching, to evaluate the final product by its capacity to satisfy the user need. From the very starting point as well as the end of lexicographic project, the user is the central focus. This paper appeals to any lexicographic endeavour to turn from the tradition of focusing only on the transfer of data, to concentrate on the importance of the user in modern-day lexicography.
TRANSLATION IN DEVELOPING AN ENGLISH-FILIPINO DICTIONARY OF PHILOSOPHY

Constantino T. Ballena
De La Salle University-Dasmariñas, PHILIPPINES
cballena@mail.dasma.dlsu.edu.ph

This lexicographic project highlights the role of translation in developing an English-Filipino dictionary of Philosophy. The translation theories of Larson (1984), and Newmark (1982 and 1988) served as the major theoretical foundation of the paper. Additionally, the lexicographic Principles of Newell (1995), Zgusta (1971), and Bartholomew and Shoenhals (1983) guided the researcher in the scientific preparation of the dictionary. The Language Planning Theory of Haugen (1966, 1969) in Fishman (1974) and Eastman (1983) stressing the quadruplet process of Language Planning: Selection, Codification, Elaboration, and Implementation provided a sound foundation of the study. The primordial task of the researcher was to develop an English-Filipino dictionary of philosophy employing a translation model that is anchored on the theories of Larson (1984) and Newmark (1982 and 1988). The researcher culled as many philosophic terms with their definitions as he could from philosophy books and dictionaries. These were translated from English, the source language text, into Filipino, the target language text. Translation, thus, occupies a pivotal role in developing an English-Filipino dictionary of Philosophy. The bilingual dictionary contains 1,842 entries or dictionary articles which have the following components: lexeme, etymology, pronunciation, word class, Filipino equivalent, explanatory definitions in Filipino, source of the definitions, examples, and synonyms. The development of English-Filipino dictionary of philosophy is the researcher’s invaluable contribution to the progress of philosophy in the Philippines and the gargantuan task of intellectualizing the Filipino language. Specifically, it will benefit philosophy students and professors, researchers and textbook writers in philosophy.

Key Words: Lexicography, Lexicographic Project, Translation, Bilingual Dictionary, Dictionary of Philosophy
USING THAI ORTHOGRAPHIC CONVENTIONS IN ENGLISH-THAI DICTIONARY: HELPFUL OR HARMFUL?

Amon Saengmanee
Mahidol University, THAILAND
jiwabaharu@yahoo.com

Though intelligible pronunciation is the most essential skill among all language skills, it has long been recognized that the major problem of Thai learners of English is pronunciation. According to the curriculum, Thai university students have no less than ten-year experience in their English classes. Unfortunately, after such a long period, most students have failed to learn to pronounce English properly. Among various language learning kits, it is widely recognized that a dictionary is highly recommended as one of the most important tools that help improve the learners’ pronunciation. This paper is an attempt to reflect the clear images of the unhelpful writing systems using in giving information about English pronunciation appeared in English–Thai dictionaries in the market. An analysis of more than 20 pronunciation transcribing systems reveals that almost all transcription are written by using Thai orthographic conventions. It is found that only one of them utilizes the practical IPA and among those using Thai orthographic conventions, only one single dictionary seem to be qualified in their system of choice. Many of them do emphasize the importance of pronunciation but provide the users with the unqualified and at the same time seem to be harmful like putting the Thai tones on the transcription. The recommendation is that the phonetic transcription provided by IPA should be actively promoted to be more widely used in the future.
This dictionary was created for both service industry professionals and students alike. Our aim in creating this dictionary was to provide an easy-to-use reference with practical vocabulary and accurate renditions of it into each target language: Thai, Chinese, Korean, and Japanese. We chose to use the native scripts for each respective language and supplemented them with their native phonetic alphabet readings (where applicable). We then went a step further and provided Romanized readings as well. This was done to help users that have little or no previous ability in any of the four languages. This dictionary is filled with useful illustrations and photographs of actual hotel/resort facilities (many taken at the Garden Cliff Resort and Spa Pattaya). Bundled with the dictionary is a supplemental audio CD containing recordings of native speakers pronouncing approximately 500 terms in each of the four languages. The artwork, photographs, and audio used are both original and recent, having been specifically sourced for this project. This dictionary was specifically designed to provide any user with a valuable source of reference for hotel-related vocabulary, in multiple languages.

Key Words: Dictionary, reference, vocabulary, multilingual, Thai, Chinese, Japanese, Korean, hotels, resorts, service industry, industry jargon, hotel terminology
THAI AND ENGLISH ELECTRONIC DICTIONARY
FOR MOBILE PHONES

Chakrit Saengkaew
Mahasarakham University, THAILAND
chakrit@msu.ac.th

Abstract: The world today can communicate in a wide dimension. There is a connection among groups of people of different races. One obstacle of the communication is the problem about language. Even though there are many courses in English language in Thailand, there are still a lot of Thai people who cannot communicate in English. The developer realized this importance and decided to develop a tool which can help people (Thais and foreigners) connect. Mobile phones are widely-used devices and they have a lot of potentials in becoming an electronic dictionary (both Thai-English and English-Thai) so that mobile phone users can look for words and translations. This way, Thais and foreigners can study, look up, understand, widen opportunities and reduce the gaps during communication. Programs and databases on mobile phones do not need to be accessed online because users can look up Thai and English words on their own mobile phones.

Key Words: Python DBMS, XML, electronic dictionary, mobile, PyS60

1. INTRODUCTION

English is considered the universal language for communication between people in the world who use different languages. In Thailand, there is a communication between Thais and foreigners all the time. During the moment when the word is not recognized, we tend to use a dictionary [1] to look up the definitions or the foreign words. It takes time to look up the words online [2] because there is a restriction about the computer size and the internet connectivity. There are electronic dictionaries available but they are quite expensive. The developer decided to build a dictionary on a mobile phone because mobile phones are devices widely used. The dictionary which the researcher developed can translate Thai into English and English into Thai. Thai is a unique language in that there are 44 consonant letter and 32 vowels along with 4 tone marks. Some consonants can work as final consonants and vowels. Some vowel forms require other vowel forms so that they become words. Therefore, dictionaries on mobile phones are important for Thais and foreigners during communication.

This research aimed at designing and building dictionaries on mobile phones. The program has 2 main components that are database for vocabulary and program to look up words. The database consisted of vocabulary, translations. The data were stored in the form of Table (Python DBMS). The program to look up words interacts with users and is used to search for words in terms of SQL on mobile phone display. This article will explain the procedures and summarizes in the last section.

2. BACKGROUND

Yuen Poovarawan [3] did research on analysis of Thai words and built a dictionary by choosing randomly word and sentence examples extracted from books, newspaper, journals, magazines, letters, official letters and general reading books, excluding books about literature
and academic texts translated from other languages. Therefore, words from foreign countries are not included in LEXiTRON [4], an online dictionary developed by Nation Electronics and Computer Technology Center or NECTEC, and Nation Science and Technology Development Agency or NSTDA. This dictionary contains frequently used words in documents. The current database consists of 79,000 English-Thai translations and 51,000 Thai-English translations [5]. However, LEXiTRON is limited to those who use can access the internet, resulting in limitations for mobile phone users who cannot access the internet. This project was initiated to make LEXiTRON available for mobile phones. At the same time, the accuracy of the data is still the same as the database. The contents are the same. Only the format is changed. The overall picture of the research is shown in Figure 1.

![Diagram of Thai and English electronic dictionary for mobile phones](image1)

**Figure 1** shows the overall idea of Thai and English electronic dictionary for mobile phones.

### 3. METHODOLOGY

Converting Thai and English vocabulary database required programming and working on microcomputers via Python Language. The database of LEXiTRON [4] which is called LEXiTRON DB was distributed in the form of XML and then converted to Python DBMS format. The advantage of this was that the file size was smaller because only the necessary information was extracted for development. The developers used Python (PY60 V1.4) [6] because Python is stable and has a lot of libraries. It is also platform-independent and freeware (Python, RDBMS). The procedures are shown in Figure 2.

![Diagram of procedure of converting database on microcomputer](image2)

**Figure 2** shows the procedure of converting database on microcomputer.

The vocabulary database was converted from XML format which is specific like

```xml
<Doc>
  <esearch>yaffle</esearch>
  <eentry>yaffle</eentry>
</Doc>
```
The XML path (XPath) is very important for conversion. However, XML format needs XML element, therefore, this kind of data was also stored in the file, resulting in larger file size. This is the limitation of mobile phones in terms of memory. The developer decided to code a program to extract only the important parts of the data in the database to build a new database (Python DBMS), collecting only vocabulary and translations to reduce the size of the dictionary. The procedure is shown in Figure 3. The principle of program is to Match on regular expression by instruction like

```
"\<esearch\>w+\</esearch\>"
```

in order to compare the words in the XML-formatted database. The example looked up the word “doc” which divides the data into groups and is used along with other keywords. The important data are organized in the table of database in the format of Python as shown in Figure 4.

A sample dictionary database after converting XML DB to Python DBMS (Table 1, Table 2).
4. **PROGRAM TO LOOK UP WORDS**

The program will show the display and wait for users to put the words to look up. Users can type via keyboard. They can choose between Thai and English or they can use standard keys on mobile phones. Users press “enter” and the program will look up the words then show the data about the words on the screen of the mobile phones.

Figure 3 shows the procedure of converting XML DB to Python DBMS
The dictionary program depends on the program to look up words in the database which was converted from XML. The word searching feature is based on SQL which uses e32db.Dbms module in Python-mobile phone programming language because this module is suitable for the database which need little revision. The module will look up words from the TE table when users want to translate from Thai into English and will look up words from the ET table when users want to translate from English into Thai. The developer use Thai Unicode encoding so that it can be used with the Thai font on mobile phones (see Figure 5).

Table 1. Examples from DBMS database of TE (Thai-English) to translate Thai into English

<table>
<thead>
<tr>
<th>tsearch</th>
<th>tentry</th>
<th>eentry</th>
<th>tcat</th>
<th>tsyn</th>
<th>tsample</th>
<th>id</th>
</tr>
</thead>
<tbody>
<tr>
<td>ดังกล่าวข้างต้น</td>
<td>วัวป่าขนยาวในทิเบต</td>
<td>abovementioned</td>
<td>PRON</td>
<td>ที่ว่างวงหน้าการ</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>การงดออกเสียง</td>
<td>การงดออกเสียง</td>
<td>abstention</td>
<td>N</td>
<td>การไม่แสดงความเห็นตามสิทธิ์</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Examples from DBMS database of ET (English-Thai) to translate English into Thai

<table>
<thead>
<tr>
<th>eSearch</th>
<th>entry</th>
<th>Tentry</th>
<th>ecat</th>
<th>id</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>a</td>
<td>หนึ่ง (คำสำคัญที่เหมาะสมกับคำนี้ไม่ชี้เฉพาะ)</td>
<td>DET</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>อักษรตัวแรกของภาษาอังกฤษ</td>
<td>N</td>
<td>1</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
<td>อักษรตัวแรกในภาษาอังกฤษ</td>
<td>N</td>
<td>2</td>
</tr>
<tr>
<td>a posteriori</td>
<td>a posteriori</td>
<td>จากผลไปสู่เหตุ</td>
<td>ADJ</td>
<td>3</td>
</tr>
</tbody>
</table>

Sample code to store data on DBMS database

```python
import e32db, time

dbms = e32db.Dbms()
dbms.open(u'E:\sample.db')

query = u"CREATE TABLE et (ESEARCH VARCHAR, EENTRY VARCHAR, TENTRY VARCHAR, ECAT VARCHAR, ID INTEGER, ETHAI VARCHAR, ESYN VARCHAR, EANT VARCHAR)"

rtn = dbms.execute(query)
print rtn

insert = u"INSERT INTO et VALUES ('a', 'a', '%s', 'DET', 0, '', '')" %
(thaithaicode("หนึ่ง (คำสำคัญที่เหมาะสมกับคำนี้ไม่ชี้เฉพาะ)")

dbms.begin()
count = dbms.execute(insert)
dbms.commit()

Sample code to retrieve data from DBMS database to display on the screen.

dbv = e32db.Db_view()
dbv.prepare(dbms, u"SELECT TENTRY FROM et WHERE esearch = '%%vocab%'")
dbv.first_line()
dbv.get_line()
th = dbv.col(1)
appuinfw.note(th)
dbms.close()
```
5. RESULTS

In the test, the word “a” was to be looked up on the mobile phone display. After “enter” was pressed, the program would show the translations on display that is “หนึ่ง (คำนำหน้าค้า
คำนำหรือแสดงว่าคำนำหน้าฯ ไม่ขัดขวาง)” as shown in Figures 6 and 7.

Figure 6 shows the display waiting for word input.
Figure 7 shows the display with word searching results.

6. CONCLUSION AND FURTHER STUDY

The words which were not available in the dictionary were normally proper nouns, for example, names of persons or places, new words and words from foreign countries. In the future, the program will be developed to include new words.

The translation for sentence can be done if the sentence undergoes Thai word segmentation by using dictionary [7] because this dictionary is to look up only words in the dictionary. In the future, there might be a program which can perform Thai word segmentation by rules [8] to check the rules about Thai language in terms of mixing letters, spacing, and beginning a new paragraph. This research can be further developed by giving sounds for the word which is looked up as well.

7. REFERENCES

MULTI-WORD COMBINATIONS AND LEXICOGRAPHY: AN ENGLISH-SPANISH DICTIONARY-MAKING EXPERIENCE

Eduardo de Gregorio-Godeo  
Universidad de Castilla-La Mancha, SPAIN  
Eduardo.Gregorio@uclm.es

Silvia Molina-Plaza  
Universidad Politécnica de Madrid, SPAIN  
Silvia.Molina@upm.es

Abstract: By exploring multi-word combination analysis in bilingual dictionaries, this contribution will present an ongoing research and editorial project aiming to produce an English-Spanish dictionary of multi-word combinations for Spanish students of EFL. After discussing the fundamental role of corpus linguistics for lexicographic analyses, and the treatment of multi-word combinations in different types of dictionaries, the lexicographic process guiding the elaboration of the dictionary will be described. With a special focus on the use of the BNC and CREA (Contemporary Spanish Reference Corpus) corpora, an account will be accordingly made of the principles determining the inclusion of entries and their presentation within the dictionary. Given that this paper is based on research aiming to build an English-Spanish dictionary of multi-word expressions, emphasis will be laid on a usage context where the source language is English and the target language is Spanish, thereby examining the educational implications of this lexicographic project for Spanish-speaking students of EFL.

Key Words: multi-word combinations, collocations, idioms, lexicography, English-Spanish bilingual dictionaries.

1. INTRODUCTION

A language ‘corpus’ is “a collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about language” (Crystal 1997: 95). By exploring the phenomenon of ‘multi-word combinations’, this paper delves into a field of linguistic enquiry where corpus linguistics is a fundamental tool today, to wit, lexicographic analyses (cf. Hartman 2001; McEnery and Wilson 2001; Hunston 2002; Ilson 2002; Leech 2002). Through a focus on word-combination analysis in bilingual dictionaries, this contribution presents an ongoing research and editorial project aiming to produce an English-Spanish dictionary of word combinations. Thus, after discussing the fundamental role of corpus linguistics for lexicographic analyses, and the treatment of word combinations as a phenomenon in different types of dictionaries, the lexicographic process guiding the elaboration of the dictionary is described. An account is accordingly made of the principles determining the inclusion of entries and their presentation within the dictionary. Given that this paper is based on current research aiming to build an English-Spanish dictionary of multi-word expressions, emphasis will be laid at all times on a usage context being one where the source language is English and the target language is Spanish. Some final remarks will make regarding the educational implications of this lexicographic project for Spanish-speaking students of EFL.

2. CORPUS-BASED LANGUAGE ANALYSIS AND LEXICOGRAPHY

Corpus linguistics has become a vibrant field of linguistic enquiry. As stressed by McEnery and Wilson, “from being a marginalized approach used largely in English linguistics and specifically in studies of English grammar, corpus linguistics has started to widen its scope [….] corpus linguistics is a lively subject, with corpus-based approaches
being taken to many research questions in linguistics” (2001: 1). As McEnnery and Wilson’s (2001: 104-128) thorough overview reveals, corpus linguistics is fruitfully being employed in numerous language-related disciplines including, but not being limited to, speech research, lexical studies, grammar, semantics, pragmatics and discourse analysis, sociolinguistics, stylistics, language teaching, historical linguistics, dialectology or psycholinguistics.

In particular, corpora have become a fundamental tool for lexicography, a domain of linguistic enquiry which deals with “the writing of dictionaries, or the scholarly study of this activity” (Trask 1997: 131) in general, and, in particular, with “the principles that underlie the process of compiling and editing a dictionary” (Jackson 1988: 247). As Trask puts it somewhere else, over the last two decades dictionaries have been revolutionised by the introduction of corpus-based techniques “and modern dictionaries are now usually based upon huge corpora of English, from which words, forms, spellings, meanings and grammatical behaviour are extracted, thus allowing lexicographers to appeal directly to the observed facts of language” (1999: 166). Since the first attempts to explore the potential of language corpora for identifying the frequency of collocations (Kjellmer 1987), modern dictionaries have come to build their information about multi-word expressions on the basis of language corpora, to the extent of there having appeared specific dictionaries of word combinations.

3. MULTI-WORD COMBINATIONS AND LEXICOGRAPHIC ANALYSIS: COLLOCATIONS AND IDIOMS

‘Collocation’ and ‘idiom’ are familiar terms for users of English monolingual and bilingual dictionaries having achieved a certain linguistic proficiency. Such linguistic phenomena are often conceived of as ‘multi-word combinations’ (Ilson 2002: 333) whose meaning is more than the sum of the meaning of their components. As Benson, Benson and Ilson maintain in their approach to this lexical phenomenon, ‘word combinations’ come into being when “certain words regularly combine with certain other words or grammatical constructions” (1997: ix). To a great extent, when exploring the word combinations of a language, both collocations and idiomatic expressions are at some point examined.

The notion of ‘collocation’ refers “to the tendency for certain words to occur together. The term itself comes from the verb collocate, meaning ‘to go together’” (Finch 2000: 152). It is common practice for monolingual dictionaries to include information about units ‘above’ the word level, such as “units including more than one complete word, i.e. compounds and idioms like blackbird, bank on, give up, night owl, hammer and tongs, at all, kick the bucket” (Ilson 2002: 333). As substantiated by Ilson’s investigation of lexicographic practices, most dictionaries generally incorporate syntagmatic information about the use of items in forming sentences, which may include notes on complementation and, which is the case in point here, information on “collocation with specific words or types of words (e.g. fond of vs fondness for; [or] the association of capsize with boats or ships)” (2002: 335).

According to Bussmann, collocations are characteristic word combinations “which have developed an idiomatic semantic relation based on their frequent co-occurrence” (1996: 81; emphasis added). Consequently, collocations are semantically and syntactically close to such lexicogrammatical units as ‘idioms’, that is, “a sequence of words which is semantically and often syntactically restricted, so that they function as a single unit. From a semantic viewpoint, the meanings of the individual words cannot be summed to produce the meaning of the idiomatic expression as a whole” (Crystal 2003: 225-226).
Idiomatic expressions are semantically related to collocations to such an extent that, as Crystal underlines, “an alternative terminology refers to idioms as ‘habitual collocations’” (ibid., 226). In point of fact, collocations and idioms may be taken to partake of the same stock of prefabricated units which – following the Russian tradition of phraseology – various authors have labelled as ‘word combinations’ – and also as ‘phraseological units’ or ‘phrasal lexemes’ – to delineate the “ready-made memorized combinations in written and spoken language” (Cowie 1998: 1), which comprise both “word-like” units, which function syntactically at or below the level of the simple sentence, and ‘sentence-like’ units, which function pragmatically as sayings, catchphrases, and conversational formulae” (Cowie, 1998: 4). Studies on the phraseology of English like Altenberg’s (1998: 120 and passim) likewise acknowledge idioms and collocations to be types of word-combinations. Indeed, Cowie himself highlights that “collocations of words in familiar literal senses are at one end of a broad spectrum of word combinations in English. At the other are idioms: combinations whose constant re-use in a fixed form has led to a radical change of meaning” (1988: 131).

4. THE CASE OF ENGLISH-SPANISH WORD-COMBINATION LEXICOGRAPHIC ANALYSES
As Cowie insisted at the end of the last decade, there is now a “wider recognition of the crucial part that ready-made memorized word combinations “play in first- and second-language acquisition and adult language production […] native-like proficiency in a language depends crucially on a stock of prefabricated units – ‘prefabs’ – varying in complexity and internal stability (1998:1). Although there exist English-Spanish dictionaries of phrasal verbs as a fundamental subcategory of idiomatic expressions like Khalaili and Marina’s (1984), comprehensive bilingual dictionaries of idiomatic expressions and collocations are almost non-existent. Thus, exceptions like Carbonell Basset’s (1971, 1995, 1996) phraseological dictionaries tend to lay a strong emphasis on proverbs and other idiomatic expressions, so that, in addition to disregarding a great deal of English idioms which are not proverbs, they do not take into account collocations on the whole. A thorough lexicographic examination of English-Spanish word combinations needs undertaking. Such a dictionary would be particularly useful for learners of English as a Foreign Language whose mother language is Spanish.

5. AN ENGLISH-SPANISH DICTIONARY OF MULTI-WORD COMBINATIONS: LEXICOGRAPHIC PROCESS
Bearing in mind the above-mentioned lack of lexicographic resources providing translations of English word combinations into Spanish, we proceed to describe the approach followed for the elaboration of an English-Spanish dictionary of such phraseological units. Admittedly, this procedural account may serve as a reference point for further lexicographic work exploring English multi-word combinations from the viewpoint of other target languages. The following subsections present the lexicographic process undertaken for the research and editing process leading to an English-Spanish dictionary of word combinations.

The project is consistent with the basic principles of contemporary lexicographic practices, thereby comprising all four stages of successful lexicographic work as echoed by Hartman (2001: 14-20), namely ‘planning’, ‘fieldwork’, ‘description’ and ‘presentation.’ The last stage includes final revision as well.
5.1. Preliminary Plan
Hartman stresses that “to be successful, lexicography as dictionary making requires careful planning and implementation of the compilation process on the basis of market research and the specification of the potential users’ reference needs to be met” (2001: 20). As discussed above, an examination of the current market of bilingual English-Spanish dictionaries evidences a significant lack of specific dictionaries dealing with word combinations. Therefore, a dictionary of this kind including collocations and idioms seemed to be a must, and, given the specificity of word combinations in every language, it seemed that the project would be beneficial for learners of English as a foreign language. This dictionary might accordingly be included within the category of dictionaries for the foreign learner, which are “essentially general-purpose dictionaries, but tailored to the needs of a specific group of users” (Jackson 1988: 174). After having been carrying out some preliminary work through some local university- and regionally-funded research, the positive attitude of the Spanish Department of Education in granting funding on a national competitive basis confirmed the expected positive results of the project.

5.2. Fieldwork
Further to the initial planning stage of the dictionary-making process, the first step in the research involved producing a thorough database of word combinations in English that might be used as a basis or ‘macrostructure’ for the alphabetically-ordered dictionary ‘entries’ in English including Spanish equivalents, usage examples and other information within the English-Spanish dictionary of word combinations. The macrostructure of lexical entries has thus been completed on the basis of existing English monolingual dictionaries of multi-word expressions including The BBI Dictionary of English Word Combinations (Benson, Benson and Ilson 1997) or the Oxford Collocations Dictionary (2004), plus various other current English-Spanish dictionaries and general-use monolingual English dictionaries. Specialised dictionaries of English idioms, such as the Collins COBUILD Dictionary of Idioms (1995), have been consulted as well to build a macrostructure of word combinations in English consisting of collocations and idioms. Given that, rather than come up with a new repertoire of English multi-word expressions defined in English, the main focus of the project was to provide Spanish equivalents for word-combination compilations in English, fieldwork has mainly relied on ‘secondary sources’.

The BNC (2001) has been adopted as a key source of information for the usage examples of the entries in the dictionary. The BNC has been chosen on grounds of its relevance for English-related lexicographic analyses all over the world nowadays. In any case, when usage examples are not found on the BNC, other English-language corpora are consulted, by way of example, the Collins COBUILD English Collocations on CD-ROM (1995). The Internet is drawn upon as a last resource for usage contexts when no examples are found within existing language corpora.

Again, equivalents in Spanish for the word combinations in the macrostructure of the dictionary are provided on the basis of existing bilingual English-Spanish general-use dictionaries. A wide range of such bilingual dictionaries have been consulted to this regard. All in all, the appropriateness of Spanish equivalents has been consistently tested by consulting the Corpus de Referencia del Español Actual (CREA) as a wide and thorough corpus of contemporary Spanish, available online on the Royal Spanish Academy (RAE) website.
5.3. Description
5.3.1. Word combinations included in the dictionary
For the purposes of this dictionary, we have followed Benson, Benson and Ilson’s (1986: 252-254) overall taxonomy of ‘lexical combinations’, namely: (i) ‘free combinations’ of words, which are those whose “components are the freest in regard to combining with other lexical items”; (ii) ‘idioms’, which are “relatively frozen expressions whose meanings do not reflect the meanings of their component parts”; and (iii) ‘collocations’, which are “fixed”/“recurrent” word combinations, that is, “loosely fixed combinations” between free word combinations and idioms. Nonetheless, free combinations of words have not been considered in the dictionary, since their meaning may be found by examining the meaning of their individual constituent words in general dictionaries.

5.3.1.1. Collocations
Benson, Benson and Ilson (1997) make a distinction between ‘grammatical collocations’ and ‘lexical collocations’. A ‘grammatical collocation’ is “a phrase consisting of a dominant word (noun, adjective, verb) and a preposition or grammatical structure such as an infinitive or clause” (Benson, Benson and Ilson 1997: xv); for instance, decide on (decidirse por) in decide on a boat. In contrast, ‘lexical collocations’ “do not contain prepositions, infinitive or clauses. Typical lexical collocations consist of nouns, adjectives, verbs and adverbs” (ibid, xxx); for example, warmest regards (saludos afectuosos) in I send warmest regards. The dictionary is mainly concerned with lexical collocations, which Spanish-speaking users of English might convert into wrong collocations very easily (e.g. *deserted children/abandoned children [nínos abandonados]). Moreover, although there is a large number of English collocations following the Verb + Noun structure (abandon hope), this lexicographic repertoire does not include all possible free word combinations. In their discussion of English word combinations, Benson, Benson and Ilson (ibid, xxx) make this point clear by considering the construction condemn murder (condenar delitos), which is a free word combination in English, the verb condemn combining with an unlimited number of nouns (e.g. condemn abortion, abuse of power) and murder similarly combining with hundreds of verbs (e.g. accept, acclaim, advocate); however, commit murder (cometer un crimen) is a collocation. Following Benson, Benson and Ilson’s (ibid, xxx-xxxiii) taxonomy of collocations, the main types of lexical collocations included in the dictionary are as follows:

(i) Verb + Noun / Pronoun / Prepositional Phrase
Most of these collocations consist of a verb of action or activation plus a noun or a pronoun (e.g. come to a conclusion, make an appointment, set a record [llegar a una conclusión, concertar una cita, fijar un récor]). In many cases, the collocation incorporates an idea of eradication (e.g. override a veto [anular el veto]).

(ii) Adjective + Noun
Typical examples include strong/weak tea [té fuerte/flojo], which may not be replaced by might/feeble tea. Sometimes it is possible for various adjectives to collocate with a single noun. In those cases, only the most frequent lexical collocations are included. Furthermore, it is necessary to bear in mind that in English many nouns have an adjectival function when they are placed before another noun (e.g. placement test [test de nivel]). Such collocations appear in the dictionary entry of the second noun. Nevertheless, if the meaning of the second noun within the word combination is different from its meaning when used independently, the word combination is included in the dictionary on the basis
of the second word (e.g. acceptance speech [discurso de ingreso en una institución pública]).

(iii) Noun + Verb
These are collocations where the verb reflects characteristic actions of the noun, be it a person or a thing (e.g. bombs explode/go off [las bombas explotan]). Combinations which are easily predictable are not considered in the dictionary (e.g. dancers dance, teachers teach, writers write [los bailarines bailan, los profesores enseñan, los escritores escriben]).

(iv) Lexical collocations indicating the ‘unit’ commonly associated to a noun
Typically, their structure in English is noun, of noun. Such collocations project the meaning that an individual belongs to a larger group (e.g. a pack of wolves, a swarm of bees [una banda de ladrones, un enjambre de abejas]), or the specific and concrete character of a unit with regard to larger elements (e.g. a bit of advice, an act of violence [un consejo, un acto violento]).

(v) Adverb + Adjective
Some examples of this type of collocation are utterly abhorrent; patently/totally/utterly absurd [totalmente aborrecible; completamente, absolutamente absurdo].

(vi) Verb + Adverb
Here typical examples could be to appreciate deeply, greatly, keenly, sincerely, very much; argue bitterly, heatedly, passionately, strenuously, vehemently [agradecer sumamente, muchoísimo; discutir apasionadamente, acaloradamente, vigorosamente].

5.3.1.2. Idiomatic expressions
The dictionary takes account of a wide range of idiomatic expressions of contemporary British and American English. Idiomatic expressions fall within different categories such as: (i) traditional idioms (e.g. Spill the beans [irse de la lengua]); (ii) new phrases (e.g. it’s all gone pear-shaped [ha salido fatal]); (iii) metaphorical phrases (e.g. face the music [afrentar las consecuencias]); (iv) two-word phrases (e.g. wild card [comodín]); and (v) various other similes (e.g. like to peas in a pod [ser como dos gotas de agua]). The dictionary does not incorporate ‘phrasal verbs’ as a characteristic type of multi-word combination because there are various dictionaries of phrasal verbs, both monolingual and bilingual, in the market.

5.3.2. Structure of entries
The dictionary is articulated in accordance with the typical procedure of alphabetical order, which “is based on the written form of the lexically relevant units rather than on their meaning” and adopts a ‘semasiological’ approach in entries, that is, “going from name to notion” (Ilson 2002: 291). The basic principles guiding the organisation of entries within the dictionary are as follows:

(i) Key headwords – compounds included – are alphabetically ordered in the dictionary. Single-word compounds precede those written as two words. Homographs follow this order: adjective, adverb, noun, verb. For instance, tease (n.) precedes tease (v.). Determiners, prepositions and pronouns are not often headwords.

```
 tease I n.
 [person who teases] a terrible tease un bromista.
tease II v.
```
1. to tease a person cruelly burlarse o reírse de una persona. By then she had acquired a distinctive Geordie accent and she was upset when her friends at school teased her about her rounded vowels and up and down, sing-song voice.

2. to tease an animal provocar a un animal. He would meddle with fishing nets, pull up anchors and tow boats, tease dogs and tow swimmers.

3. to tease a fabric; wool cardar un tejido; lana. As a test of his strength Utnapishtim challenged him to stay awake for six days and seven nights - But while Gilgamesh sat there resting on his haunches, a mist of sleep like soft wool teased from the fleece drifted over him (…).

(ii) Entries contain at least one context of usage. The key headword is written in small capital letters. English collocations are written in bold and Spanish equivalents appear in italics. Regular font face has been adopted for usage examples, the word combination in question being underlined. Here is an example:

machete n.
to brandish, wield a machete blandir un machete. We cant take time to talk to the stylist before they wield the machete.

(iii) Cross-references within the dictionary are highlighted in yellow. For example, in the following entry penny is highlighted in yellow because the same idiom may be found in the entry for penny within the dictionary:

ten n.
1. to be two/ ten a penny ser baratísimo/ no valer nada. Uncritical testimonials to the postmodern’s attractions are ten a penny, and conservative denunciations thereof not much scarcer.
2. ten to one (inform.) diez a uno/ te apuesto lo que quieras. The great bulk of those who thronged Emmett place last night wanted to see the film, and they outnumbered protestors by a minimum of ten to one.

(iv) When a word collocates with others, the dictionary pinpoints series of collocations in an alphabetical order (e.g. ABILITY: to demonstrate, display, exhibit, show). This helps not only to save space in the dictionary, but also to display synonyms and near-synonyms. So, in the series for ability, demonstrate ability and display ability are treated as synonyms. However, collocations which are not synonyms are separated by semicolon (;) (e.g. domestic; physical; sexual ABUSE). Synonyms are thus grouped together and separated by commas within the series of collocations. As shown in the example below, in to operate, run, use, work; shut down a machine, to operate, run, use and work are treated as synonyms, and are likewise separated from shut down (which is not a synonym) by a semicolon:

machine n.
1. to operate, run, use, work; shut down a machine usar, utilizar; apagar una máquina. A knowledge of how to set up and operate a machine could reduce the likelihood of a breakdown so minimising the time spent on repairs.

(v) Collocations whose meaning may be difficult to grasp by the student are often defined between brackets on the left of the collocation:

market n.
[...]

7
5. a bear ("falling"); bull ("rising") market mercado/bolsa a la baja; en alza. Spicer points out that the sector has thrived in a bull market but the advertising boom may be coming to an end.

Information about register, dialects and other social factors is also indicated between brackets, e.g. formal (form.), American English (IAm), British English (IBr); etc:

**marine n.**

1. a mercantile (IBr), merchant (IAm) marine marino mercante. It is true that certain measures of nationalisation had been undertaken in the first months of Soviet government - for example, the Merchant Marine had been taken over in January 1918 (…).

(vi) As the following example of entry from the dictionary instantiates, idiomatic expressions are written in blue at the end of entries:

**mat n.**

1. to weave a mat sacudir una alfombrilla/ un felpudo. The Lele are subsistence cultivators, growing maize, ground-nuts, and raffia palms from whose fronds they weave mats which are used as a special currency - a rare case of money really growing on trees.
2. a bath; exercise; prayer; welcome mat alfombrilla de baño; para hacer ejercicio/ colchoneta; para arrodillarse y rezar; de bienvenida. For elderly people, grab rails at strategic places (including the bathroom/toilet) and use of a non-slip bath mat are devices likely to prevent accidents.
3. to go to the mat (for sb) (IAm) hacer todo lo que puedes para solucionar un problema difícil, para ayudar a algn que tiene menos poder que tú.

(vii) Examples have been taken from the BNC and, to a lesser extent, from the Collins COBUILD English Collocations on CD-ROM (1995). Although this is the case on very few occasions, the dictionary includes sometimes collocations without usage examples as no examples have been found in the language corpora utilized. An equivalent is provided in Spanish at all times, idiomaticity has been consistently tested by consulting the Corpus de Referencia del Español Actual (CREA):

**air conditioner n.**

[…]

3. to run; turn on an air conditioner; turn off an air conditioner poner/encender; apagar el aire acondicionado.

(viii) Alternative translations of a collocation or idiom are separated by a slash (/):

**mad adj.**

[…]

5. mad cow disease enfermedad de las vacas locas/ encelofatía espongiforme bovina. Excuses abound: world markets have collapsed, diet-conscious Europeans are eating less red meat, some people in Britain fear it will give them mad cow disease.

(ix) Usage notes are highlighted in light blue, and range from pragmatic information to differences between British and American English, through other grammatical questions. They are also used to make reference to ‘false friends’, namely “terms in two languages which are phonologically and graphologically similar (cognates), but have subtly different meanings” (Malmkjær 2002: 82). Here is an example:
tape n.

9. audiotape; videotape cinta de audio; de video. The screen replaces the car windscreen, and on it is shown a road scene that has been recorded on video tape, and transferred to disk.

Nota de uso: se pueden encontrar los compuestos juntos (videotape) o separados (video tape).

(x) The use of a long underscore (______) in an idiomatic phrase indicates that various nouns, adjectives or verbs may be inserted in the gap:

man n.

33. our man in_____: nuestro hombre en ______. I don't know why our man in Madrid came to see me; just idle curiosity, perhaps – so few people have ever met me, so many seem anxious to do so.

5.4. Presentation and Revision

At this stage of the editing project, the dictionary on the whole is close to completion. The final stage of the project still needs to be undertaken. This will entail formatting, printing and proofreading the whole dictionary several times. A thorough revision of the dictionary will accordingly have to be carried out prior to its final publication.

6. CONCLUSION: EDUCATIONAL IMPLICATIONS

Ordinary English-Spanish/Spanish-English dictionaries admittedly include information on idioms and, to a lesser extent, collocations. Representative cases are the Gran diccionario español-inglés, english-spanish (García-Pelayo y Gross 1993); the Diccionario Oxford avanzado para estudiantes de inglés, español-inglés inglés-español (1996); El Diccionario Oxford español-inglés, inglés-español (2001); and the Collins Dictionary: Español-Inglés, English-Spanish (2003). However, such information is not always presented to the language learner in a student-friendly way. This project intends to compensate for the shortage of bilingual dictionaries providing instant and user-friendly access to Spanish-speaking students of EFL. Authors like McCarthy stress that “the relationship of collocation is fundamental in the study of vocabulary” (1990: 12), and thereby is a major area of concern for learners of EFL. In a similar fashion, “idioms are a great source of difficulty for foreign learners” (Taylor 1990: 49), which converts bilingual dictionaries of multi-word expressions into an invaluable resource for both Spanish learners of EFL. Dictionaries like this may thus help to facilitate and increase Spanish speakers’ cultural awareness of English, since “all fluent and appropriate language use requires collocational knowledge” (Nation 2001: 318).

NOTES

(1) This paper is part of a research project titled “Aplicación electrónica para traducción de colocaciones y frases idiomáticas inglés-español” and funded by Castilla-La Mancha regional government in Spain (JCCM, PAI07-0018).

REFERENCES

CREA (Corpus de Referencia del Español Actual). Real Academia Española [http://corpus.rae.es/creanet.html].
STANDARDIZING THE PRESENTATION OF A PEDAGOGICAL
DICTIONARY-CUM-GRAMMAR WITH REFERENCE TO
TAMIL-ENGLISH (FROM THE POINT OF VIEW
OF LANGUAGE LEARNING)

K.Karunakaran
Annamalai University, INDIA
karuna@umich.edu

There are a few Pedagogical Dictionaries in Tamil, based on modern applied linguistics methodology. But none of them are adequate for Language Learning-Teaching purposes from the pedagogical point of view. Hence this serious and sincere attempt. Pedagogy is an important domain, that helps-motivates-attracts a learner very much. Pedagogical Dictionary is an useful reference-resourceful treasure for a language learner. A Pedagogical Dictionary-cum-Grammar is an innovative reference work that makes a learner to identify a lexical item along with its grammatical relevance and significance in its use. In a way it is a dictionary that takes into account, briefly, the two sides (aspects) of language viz., "language structure" (LU)-lexical and grammatical on the onehand, and that of "Language use (LU)on the otherhand. This kind of dictionary-cum-grammarmakes a learner-especially a second language learner, to identify a lexical item not only with its grammatical significance, but also with its socio-contextual relevance and social meaning(like the culture bound usages, contextual usages etc.). It tries to correlate the "structure" with "contextual meaning" in a number of cases in actual use. It is also possible to cross-check the lexical item or form with the contextual use. Hence, the purpose and aim of this kind of research study is to standardize a set of vocabulary and usages (which are basic-frequent-socially and contextually significant and so on) from the points of view of the use of modern Tamil (as reflected in modern spoken varieties, and modern written Tamil) in modern communication, like language teaching, teaching through the modern written Tamil, modern massmedia, etc. This paper tries to present the different processes followed in the modernization, and thereby standardization of a "Tamil-English Pedagogical Dictionary-cum-Grammar" with suitable format as well as pertinent structural and contextual conditions and
explanations. For example, the functional significance of the following nouns and utterances (some of the noun sub-classes as well as other word classes) are very important from the point of view of structure and context in Tamil use.

As Tamil is a highly diglossic language, it becomes quite relevant to include the equivalent spoken forms and their behaviour in various linguistic environments in language structure and use. Hence this kind of Pedagogical Dictionary-cum-Grammar tries to present the spoken equivalents in 'phonetic writing' along with a series of conversion rules for the easy conversion from the written variety to the Spoken one. This paper also discusses the usefulness and impotance of the practical utility value of such a dictionary with suitable illustrations and arguments.
BILINGUALIZED TECHNICAL DICTIONARIES: THE NEEDS OF THAI UNIVERSITY STUDENTS

Jirapa Vitayapirak
King Mongkut’s Institute of Technology Ladkrabang, THAILAND
kvjirapa@kmitl.ac.th

Abstract: This paper presents the lexical needs of Thai engineering students. The first part is devoted to a review of the major research into dictionary users and uses, focusing on the EFL and ESP situation at tertiary educational level. The second part reports on an investigation into aspects of dictionary use among engineering students. The subjects were a population of 350 undergraduates, Faculty of Engineering, King Mongkut’s Institute of Technology Ladkrabang (KMITL), Thailand. SPSS for MS WINDOWS was used to analyse the data from questionnaires. The results show that the students’ reliance mainly on bilingual pocket English-Thai dictionaries. The main purpose in using dictionaries was to decode text to access meaning. They use dictionaries in the context of translating English texts into Thai (94.2%) and reading textbooks (93.5%). The top three purposes of dictionary use were for (1) decoding meaning (99.7%), (2) checking spelling (73.2%), and (3) pronunciation (53.3%). The participants’ preference among the dictionary samples tested was for the bilingualized version.

Key Words: bilingualized, technical dictionaries, dictionary use, Thai, dictionary needs, EFL, ESP

1. INTRODUCTION

In order to compile a technical dictionary to support language learning, one must first appreciate both what a technical dictionary is, and what kinds of information it is designed to convey. However, the needs of ESP learners in particular areas have been little analysed or reported in the literature of dictionary-making. Most studies focus on general monolingual dictionaries. It is rare among ESP students to find a lexicographical product that has been based on research results gained from what actually goes on in the teaching and learning of ESP. The need for empirical research on ESP users’ needs seems obvious.

2. RESEARCH ON THE USE OF MONOLINGUAL, BILINGUAL, AND BILINGUALIZED DICTIONARIES

Research on foreign language learners and their dictionaries stretch from the pioneer works of the 1970s to recent investigations of the 2000s. There has been a substantial amount of research on the study of dictionary making and the question of the interaction between learners and dictionaries. There are two main groups of relevant research on two headings: (1) Works on Monolingual/Bilingual Learners’ Dictionary (e.g. Hartmann 1983; Tomaszczyk 1979; Baxter 1980; Bensoussan and et al 1981; Atkins and Knowles 1990), (2) Works on Bilingualized Learners’ Dictionary (e.g. Hartmann 1994b; Laufer and Melamed 1994; Laufer 1995; Laufer and Kimmel 1997; Laufer and Levitzky-Aviad 2006; and Tsakona 2007). Despite the different parameters of language context, users, and approaches, their works converged in some of their findings.

The first major conclusion is that most of EFL students prefer bilingual dictionaries rather than monolingual dictionaries (Tomaszczyk 1979), Baxter (1980), Bensoussan and et.al. (1981), and Hartmann (1983), Tickoo (1989). This finding was confirmed by one of the most comprehensive studies of this kind, empirical project on the use of
dictionaries by over 1,100 learners of English in seven different European countries (Atkins and Knowles 1990). Their finding was that 75% of the students use a bilingual dictionary and only 25% a monolingual. It seems that L2 learners, even those who have achieved a good level of L2 proficiency and have been trained in academic skills, including dictionary use, still reach out for a bilingual dictionary. In Piotrowski’s (1989) survey of studies of dictionary use, he concluded that “no matter what their level of competence foreign learners and users use their bilingual dictionaries as long as they use dictionaries at all”. Why do the EFL students prefer bilingual to monolingual dictionaries? The main reasons are that the monolingual dictionaries are too difficult for the learner to use. Though the monolingual ones introduce the user right into lexical system of L2, the bilingual ones are ideal for quick consultation (Kharma 1985). Tickoo (1989) commented that certainly many students are not aware of the riches that monolingual dictionaries contain.

The second common finding was that most of the students were unable to use monolingual dictionaries effectively (Kharma 1985, Hert and Stein 1987). The EFL students are rarely trained in dictionary use (Griffin 1985, Iqbal 1987). This was demonstrated in Tickoo’s study (1989) of dictionary among Asian students. He wanted to find out whether tertiary-level students develop the essential reference skills after they enter the university. He used entries from two English dictionaries and gave a few dictionary-based tasks to 60 undergraduates. Most of them were unable to take advantage of the additional features that make the Monolingual Learner’s Dictionary (MLD) a more powerful learning tool. They know little about its grammar codes or about how to use them as an aid to their own use of words, word-forms and word-meanings.

Actually EFL learners should have dictionaries that combine the pedagogic features that they need to learn, with whatever user-friendly features it is possible to give them. Moreover, in most countries of the world, language learning incorporates some degree of use of the mother tongue. When left to the local publishing industry, this usually finds expression in simple bilingual non-pedagogical dictionaries. Bilingual dictionaries can only encourage use of the indirect or translation method in foreign language teaching, which is obviously not the aim of these education systems. Therefore, these monolingual dictionaries are not compatible with the students’ requirements. They are imposed upon the students for want for more appropriate dictionaries. If this is the learner reality, then a hybrid dictionary which contains the two types of information (monolingual and bilingual) seems to be the most appropriate product of lexicographers’ effort. Since a good product is expected to satisfy the needs and preferences of its consumers. This realization has generated bilingualized versions of general English dictionaries over the last decade. They are an emerging genre in EFL pedagogical lexicography, still at an exploratory stage.

In comparison with all the research into bilingual and monolingual dictionary use, there is little in the use of bilingualized dictionaries. Among the learner’s dictionaries on the market, the ‘bilingualized learner’s dictionary’ is a separate development of monolingual, bilingual and pedagogical lexicography. The bilingualized (BLD) learner’s dictionary is a type of dictionary based on a monolingual (MLD) dictionary, whose entries have been translated in full or in part into another language. Since a bilingualized learner’s dictionary is based on a MLD, its entries are much richer in grammar than those of the L2>L1 section of a bilingual dictionary (Marello 1998). Translations in most BLDs are provided, as L1 shortcut definitions of the L2 entries. The reasons for creating the bilingualized learner’s dictionary is the need to integrate the dictionary into language learning process, to provide more information for both encoding and decoding, as well as
the need to take into account problems of L1/L2 interference (Nakamoto 1995). The most important task of the bilingualized dictionary is to overcome its current limitation to passive decoding activities, and move on to innovative formats that can help promote active encoding, and thus interlingual communication (Hartmann and James 1998). There is an overwhelming evidence that if learners are going to make mental associations between their mother tongue and their target language, it is more productive to provide a tool which offers appropriate help than something which potentially affords less than adequate support (Hartmann 1994). With its translation equivalents of headword and/or examples, the BLD makes the EFL dictionary more user-friendly for the student who is still pretty dependent on his/her mother tongue. It means that the bilingualized dictionary is a more effective resource than the monolingual and bilingual dictionaries for some study activities (Cowie 1999). Clearly the bilingualized dictionary is suited for passive decoding (reading) (Hartmann 1992). A close examination shows that BLDs can differ considerably from each other, even if they are called bilingualized dictionaries. For instance, the Hornby-based BLDs differ from each other, despite their common parent dictionary.

3. DICTIONARY NEEDS ANALYSIS OF THAI ENGINEERING STUDENTS

There are two basic motives for undertaking this study of dictionary needs of Thai engineering students. Firstly, all of the users’ needs research has been conducted outside the context of Thailand, e.g. Poland, Germany, France, Japan, Jordan, or China, etc. There has never been any research on dictionary use at any level in Thai educational system, and the lexicographical input to bilingual dictionaries had never been very sufficiently discussed. The Thai language system has its own features derived mainly from Pali and Sanskrit and the learning context in Thailand differs from other countries, so the problems in learning English are different from those of say Polish, French, or Chinese. It is obvious that the needs of Thai students will be different from those of other language communities around the world. Therefore, the previous studies are not necessarily applicable to Thai learning contexts.

Secondly, the immediate concern is to make technical dictionaries for language learners that are more sophisticated in content and format in order to make them better tools for language learning and teaching. The existing technical dictionaries are no more than straightforward technical glossaries which merely aim to provide equivalents of technical terminology. The major previous research into dictionary users and use concerns the general dictionaries not ESP ones. Our understanding of the ESP students’ reference skills and their interaction with dictionaries is quite limited. We thus need to uncover the needs of the intended user (Thai) and target situation in relation to the information categories to be corporated in the ESP dictionary, to gain insights into how to improve the scope of ESP dictionaries and to enhance language learning.

3.1 Methodology and Data Analysis

This inquiry into the lexicographical needs of Thai engineering students is intended to relate pedagogical lexicography to the context of ESP. Series of specific questions were formulated, to obtain background information in this area. They included (1) the students’ experience in dictionary use at school level, (2) dictionary use at university level, (3) preferences among three types of dictionary: monolingual, bilingual, or bilingualized. The subjects were a population of 350 undergraduates (third and fourth year).

3.2 Selected Findings and Discussions

3.2.1. Student Survey: Reference Skills/Dictionary Use Prior to University
When asked about at which stage they started using dictionaries and when they used dictionaries most often, more than half of the students (66.5%) used dictionaries at the elementary level of education, but they made very little use of them (1.5%). The remaining 32% of the students indicated that they started using dictionary at the preparatory stage. They use dictionaries with greater frequency at the preparatory stage (21.1%) than at the elementary level. The amount of use increased steadily when they were at Secondary and Tertiary level (31.7% and 44.3% respectively) as shown in Graph 1 below:

![Graph 1: Dictionary Use at School](image)

The graph effectively shows how slowly the students develop the habit of using their English dictionaries. This goes with the fact that the need to read in English develops only gradually through the secondary years. Students were also asked how they started using dictionaries and under what influences/factors. The results show that the English teacher influenced students’ decisions (73.5%), personal initiative (16.9%), and other advisers (9.5%). Other researchers have also reported that the learner’s choice of dictionary is often recommended by teacher. But when a monolingual learner’s dictionary is first acquired, usually on the advice of a teacher or taken on teacher’s authority (Stake 1990), a wide gap often exists between a student’s positive perception of the dictionary and his or her capacity to make full and proper use of it. We can see that teachers’ role is important here and they require good judgment and sharper criteria for judging the usefulness of dictionaries. Generally, bilingual dictionaries were recommended for elementary levels and monolingual dictionaries for more advanced learners. Teachers should be able to take advantage of the dictionary features that make them an effective learning tool.

### 3.2.2 The Type and Size of Dictionaries

Information was also obtained as the types and sizes of dictionaries used before joining the university. Here in Table 1 is a summary of responses:
This table clearly indicates the students’ reliance on bilingual pocket English-Thai dictionaries (77.2%) during their school education, not the pocket monolingual ones (6.5%). It is interesting to note that Thai-English dictionaries (i.e., ones for encoding) were hardly used (1.8%). The main purpose in which students used dictionaries at the pre-university level was reported to be overwhelming concerned with decoding i.e., ‘Meaning’ (87.4%). The purposes noted are for Pronunciation (6.2%), Spelling (3.7%) and Grammar (0.6%) respectively.

<table>
<thead>
<tr>
<th>Dictionary Size</th>
<th>Dictionary Type</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pocket</td>
<td>English-Thai</td>
<td>77.2</td>
</tr>
<tr>
<td>Pocket</td>
<td>English-English</td>
<td>6.5</td>
</tr>
<tr>
<td>Large</td>
<td>English-Thai</td>
<td>6.2</td>
</tr>
<tr>
<td>Large</td>
<td>English-English</td>
<td>4.3</td>
</tr>
<tr>
<td>Large</td>
<td>Thai-English</td>
<td>3.1</td>
</tr>
<tr>
<td>Pocket</td>
<td>Thai-English</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Table 1: Types and Sizes of Dictionaries Used at School

3.2.3 Dictionaries use in university contexts

For the ESP lexicographer, it is helpful to know in which communicative situations the ESP learners are engaged while consulting a dictionary.

(1) Strategies for Dictionary Use
The students were asked about their reading strategies. The results showed that while reading, 66.5% of them tried to guess the meanings of unfamiliar words or phrases first, before deciding to consult a dictionary. The second strategy reported by about two thirds of them (62.2%) was checking on how to use an English word that they know already. 51.4% said they looked up the meanings of all unfamiliar words or phrases. Between them these findings show that many students attempt some sort of guessing while reading before consulting a dictionary. However, when asked about their use of dictionaries, the vast majority of them (94.5%) say they use dictionaries frequently. It means that they find dictionaries useful after all.

(2) The Situations of Dictionary Use
When asked about the activities in which they used English-Thai dictionaries 67.9% of the students said they used them in translating and 23.4% in Reading, 7.1% for Writing. For the students speaking (0.9%) is almost as unimportant as listening (0.6%). The results were shown in Table 3 below.
Table 3: Situations of Dictionary Use

The further questions, the majority of students said they needed most help in understanding textbooks. They use dictionaries in the context of translating English texts into Thai (94.2%) and reading textbooks (93.5%). This is because textbooks constitute a major source of reading material in the academic life of the undergraduate. It is in fact a big problem that Thai students are required to consult or buy English textbooks and usually designed for native speakers of English. English textbooks are extremely used as source material when students are asked to prepare technical reports or oral presentations while they are studying.

Table 4: Situations in which Students Need to Use a Dictionary

Apart from the major situations of dictionary use shows that the meaning of unfamiliar words is very common reasons for consulting the dictionary. The vast majority (83.7%) of students use it for this purpose showing students use dictionaries more for decoding than encoding. The decoding involves mainly language of the reading, not of listening.

(3) The Linguistic Purposes of Dictionary Use

The populations were asked about what linguistic purposes made them go to dictionaries. The results were shown in Table 5 below:

Table 5: The Purposes of Using Dictionaries
The top three purposes are for (1) decoding meaning (99.7%), (2) checking spelling (73.2%), and pronunciation (53.3%). This hierarchy agrees with Barnhart’s (1962), Béjoint’s (1981), Hartmann’s (1983) surveys: all found that looking up meaning came first for their students. Despite this common finding, the three investigations had also focused on different kinds of dictionary. The first (Barnhart 1962) referred to monolingual dictionaries in the users’ native language, the second (Béjoint 1981) deals with monolingual dictionaries not in the users’ native language (how French students of English used their monolingual general English dictionaries), and the third (Hartmann 1983) discusses bilingual dictionaries as used for reading foreign-language text.

<table>
<thead>
<tr>
<th>Barnhart 1962 (MLD in L1)</th>
<th>Béjoint 1981 (MLD not in L1)</th>
<th>Hartmann 1983b (EFL students)</th>
<th>This Study (EFL/ESP students)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meaning</td>
<td>Meaning 87%</td>
<td>Meaning</td>
<td>Meaning 99.7 %</td>
</tr>
<tr>
<td>Spelling</td>
<td>Grammar 53%</td>
<td>Grammar</td>
<td>Spelling 73.2 %</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>Synonyms 52%</td>
<td>Use in context</td>
<td>Pronunciation 53.3 %</td>
</tr>
<tr>
<td>Synonyms</td>
<td>Spelling/pronunciation 25%</td>
<td>Spelling</td>
<td>Example of Usage 52.4 %</td>
</tr>
<tr>
<td>Usage notes</td>
<td>Language Variety 19%</td>
<td>Synonyms</td>
<td>Synonyms 51.7 %</td>
</tr>
<tr>
<td>Etymology</td>
<td>Etymology 5%</td>
<td>Pronunciation</td>
<td>Syllabification 41.2 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Etymology</td>
<td>Grammar 36.3 %</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Etymology 23.7 %</td>
</tr>
</tbody>
</table>

Table 6: Users’ Requirements as regards Dictionaries

All of these findings together, with the works of Iqbal (1987) (99.43%) and Diab (1990) (92.8%) underscore the fact that dictionaries are mainly used for decoding meaning.

a) Spelling Issues: The fact that the Thai language has a completely different phonological and orthographical system from English challenges students’ spelling. A majority of students (73.2%) reported that they use dictionaries to check spelling, since they are an obvious source of information. It is noteworthy that spelling has also been ranked highly in relation to native language dictionaries. Other studies (Hartmann 1983b, Béjoint 1981, Kipfer 1985) also confirm that spelling forms a major reason for consulting the dictionary, and the users of various backgrounds do need help in this respect. For ESP lexicographers this means that provision of spelling guidance should be made, e.g. provision of alternative ways of spelling (e.g. program-programme, enrol-enroll).

b) Grammar Needs: In general, not many students seem to look up grammatical information in dictionaries. The findings of the KMITL survey revealed that only 36.3% of students look up grammatical information in the dictionaries. The students may not be aware of the fact that dictionaries provide a wealth of grammatical advice to help them; or else they may not understand the abbreviation or symbols used in dictionaries. They are probably used to traditional bilingual dictionaries which do not provide much grammatical information. Yet although looking up grammar came in second last in this study, in Hartmann (1983b) and Béjoint’s study (1981), it came in second top after the search for meaning. Such conflicting results suggested that each group of users had its own needs and background.
c) Pronunciation needs: As non-native speakers studying English as a foreign language, two-thirds of the students (65.5%) reported consulting dictionaries for pronunciation. Pronunciation seemed important for them as it came in the third position among eight purposes of dictionary use. Hartmann (1983) and Béjoint (1981) have also reported on the use of dictionaries for pronunciation, but it seemed relatively unimportant for their students. However, it is hard to compare the findings, since other factors (such as the balance of writing/reading, listening/speaking) are involved in each situation. Yet the importance of pronunciation for Thai students should not be underestimated. The results revealed that 41.2% of students claimed to be using dictionaries for syllabification. It is a reasonable decision to hyphenate the phonetic transcription to enable the learners to pronounce the term correctly.

d) Etymology: Only 23.7% of students indicated using dictionaries for word origin. Other researchers have found that etymology did not capture the interest of dictionary users (Barnhart 1962, Béjoint 1981, and Hartmann 1983). On the contrary, Diab (1990:164) found that 58% of his Jordanian students did use dictionaries for this purpose. This is because his population is nursing students who need to know about the medical terms and most of the medical terms are derived from Greek and Latin. So etymology in ESP dictionaries is helpful for them. However, etymology came last of the list in this study, so it seems like the students do not need it much in their engineering studies.

3. Preferences among Three Types of Dictionary Entry: Monolingual, Bilingual, or Bilingualized

When asked about how dictionary entries may be improved in the future, the students and lecturers were given three kinds of entries i.e. “A” is a monolingual entry, “B” is a bilingual and “C” is a moderately bilingualized one. The findings in Table 7 below show that they clearly preferred type C, i.e. the Bilingualized version of dictionary (48.9%).

<table>
<thead>
<tr>
<th>Types</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (Monolingual)</td>
<td>6.8</td>
</tr>
<tr>
<td>B (Bilingual)</td>
<td>21.2</td>
</tr>
<tr>
<td>C (Bilingualized)</td>
<td>48.9</td>
</tr>
<tr>
<td>Missing</td>
<td>23.1</td>
</tr>
</tbody>
</table>

Table 7: Dictionary Entry Type Preferences

The main reason for this preference was that the use of Thai and English definitions side by side makes it easy to understand the concepts completely. They don’t have to use two dictionaries, i.e. monolingual and bilingual technical dictionaries. The bilingualized dictionary is ideal for quick consultation. This confirms the value of bilingualized dictionaries reported earlier. They can better support the language learning process, with information for both decoding and encoding activities. They can help students overcome the common tendency to passive decoding activities, because of the way older dictionaries were designed. The bilingualized dictionary seems to offer most to EFL learners.

In other comments, the students complained that the existing English–Thai technical dictionaries sometimes let them down when they cannot find out what they need. The definitions of technical terms, are sometimes ambiguous or too simple or too broad to understand. Moreover, due to the lack of examples in those dictionaries, it is difficult to understand their contextual meaning. They recommended that an entry should include both definitions and examples, including illustrations which show the specific context for
the important terms. Examples of usage are especially important since they show accurate grammatical use of a term in contexts. All such information provides a lead for the lexicographer who is considering producing a dictionary. A good deal more content verbal is needed, as well as illustrations, to stimulate and inform student readers.

4. CONCLUSION

The study presented in this paper examined the lexical needs of Thai engineering students. Both ELT methodology and that of learners’ dictionaries foreground the need to evaluate the learning situation, as input to the development of teaching and reference materials. This research confirmed earlier reported findings that such students are heavily dependent on bilingual dictionaries for reading activities. But their evaluation and appreciation of more bilingualised dictionary mentioned indicated that this was the way to go, and set the target of the dictionary development research. This study thus charts an approach to ESP dictionary-making. Its research is a contribution to the bilingualised technical dictionary of the future.

5. REFERENCES


CELTIC WORDS IN ENGLISH DICTIONARIES AND THEIR APPEARANCE IN THE BNC AND WORDBANKS ONLINE

Mitsuhiko Ito
Toyohashi University of Technology, JAPAN
mitsu@hse.tut.ac.jp

Abstract: In previous research, the current researcher analyzed words of Celtic origin from several English dictionaries and a few of English etymological dictionaries and found that there are about 300 words of Celtic origin listed in present English dictionaries. He has studied which words and how many of these 300 words native speakers of English know. The research method used in this previous research included giving matching tests of words and definitions as well as having subjects write appropriate words for definitions. The subjects were all adult volunteers. Given this previous research, the main purposes of the present study are (1) to survey which of these 300 words mentioned above appear in the British National Corpus (BNC) and the Collins Wordbanks Online English Corpus (WB), and (2) to survey if well-known words of Celtic origin by native speakers are high frequency words in the BNC and WB. There were two main results deduced from the present study. One is that not all of the 300 words appear in the BNC or WB; there are some that only appear in one of the two, some in both, and some that don’t appear in either. The other result is that well-known words in the research do not necessarily appear at the top of high-frequency words as listed in the two corpora.

Key Words: Celtic words; English; Dictionaries; Corpus; native speakers of English

1. INTRODUCTION

In previous research, the current researcher analyzed words of Celtic origin from several English dictionaries and a few of English etymological dictionaries and found that there are about 313 words of Celtic origin listed in present English dictionaries (Ito, 2005). He has studied which words and how many of these 313 words native speakers of English know. The research method used in this previous research included giving matching tests of words and definitions as well as having subjects write appropriate words for definitions (Ito, 2006). One question arose after these studies: Would these results show similarities when compared to the frequencies of these words found in two large corpora?

2. PURPOSES OF THE PRESENT STUDY

The main purposes of the present study are (1) to survey which of these 313 words mentioned above appear in the British National Corpus (BNC) and the Collins Wordbanks Online English Corpus (WB), and (2) to survey if well-known words of Celtic origin by native speakers are high frequency words in the BNC and WB. Other relevant topics might be touched on through the discussion based on data collected from the corpora and previous work by Ito (2006).

3. PROCEDURE OF THE RESEARCH

First, the 313 words of Celtic origin analyzed in previous studies were compiled again into a list for the present study. Second, the researcher determined if each word of this 313-word list is listed in the BNC and WB. Third, the frequencies of shared words from the 313-word list and the BNC or
WB corpora were recorded. Finally, a comparison of these frequencies from the corpora and the data from a questionnaire was carried out.

3.1 313-WORD LIST COLLECTED FROM DICTIONARIES
The 313 words of Celtic origin compiled from English dictionaries are as follows:

ach, airt, ambassador, anchor, andiron, ap-, arrah, arpent, ass, bal, ball, bundle, bannock, banshee, bard, basin, bawn, beak, beat, Beltane, ben, betony, billet, bijou, bin, biretta, birlinn, bodkin, bodrag, bog, bonaght, bonnyclabber, booly, bothy, bouge, bowssen, bracken, bragget, brail, brasserie, brat, brattach, bray, brehon, brier, brissance, brock, brogan, brogue, brut, buckeen, bulge, bullace, bushel, butcher, caber, cade, cader, California, cain, caird, cam, cambrel, cammock, cannach, cant, cantankerous, capercaillie, car, carpenter, carrow, carry, caschrom, cashel, cateran, caubeen, cellidh, change, char, char, clabber, clairsach, clachan, clan, claye, claymore, cleave, coarbe, coge, coccagee, cockabondy, colleen, collop, commot, coracle, corgi, corn, corm, coronach, corrie, cosher, coshery, costean, cot, coyne, crag, crannog, cranreuch, creagh, creaght, crine, cro, cromlech, cross, crostorie, crottle, crowd, cudden, cuddy, culdee, curragh, cwm, Dal Eireann, dalt, deasil, dartre, dillue, dolmen, dour, down, drape, druid, drum, duan, dukeen, dulse, dun, dunuivlassal, eisteddfod, elvan, embassy, encumber, eric, esker, fail, fewterer, filibeg, fiorin, flannel, flummery, font, frown, gard, gaddiard, galloglass, gallon, galore, garron, garrote, garter, gelt, gillie, glean, glen, glib, gob, gob, goblet, gombeen, gossan, gowin, gravel, grollach, grave, gravel, grig, grouse, growan, guillem, gull, hubbub, ieroe, inch, ingle, iron, javelin, keen, keen, kern, kex, kibe, killas, kish, kistvaen, knock, kyle, lance, lay, leach, lee, leprechaun, linn, loch, lochan, lough, lough, loy, lozenge, lymphad, marl, menhir, messan, metheglin, minaudiere, mine, minion, mooch, morgay, morglay, mormaor, muley, mun, musha, mutton, och, ohone, ollamh, ough, ovate, oy, palfrey, partan, peat, pen, pendragon, pennill, peulvan, pibroch, piece, piket, pillion, plaid, pollan, port, poteen, ptarmigan, puffin, punt, quaich, rap, rapparee, rath, rich, ruche, sagum, say, scraw, sennachie, shamrock, shanty, shebeen, sivvens, skene, slat. Slob, slogs, sock, socket, sonsy, sorran, spalpeen, splechan, sporrin, spreach, stannum, strathe, tais(ch), tan, tanist, termon, tinche, tocher, toman, tope, tor, Tory, towan, tres, trouse, truant, tuath, tun, union pipes, usquebaugh, vassal, vendace, vergobret, vug, warren, weem, whisky, wirra, wrasse (Ito, 2008).

4. WORDS FOUND IN THE BNC AND WB
The researcher determined if each of the above 313 words are also found in the BNC and WB. The results show that 144 words of the 313 are listed in the BNC and 128 words in the WB.

4.1 144 Words in the BNC
The 144 words found in the BNC are as follows:

ach, ambassador, anchor, andiron, arpent, ass, bal, bannock, banshee, bard, basin, beak, beat, Beltane, betony, billet, bijou, bin, birlinn, bodkin, bog, brasserie, brogan, brogue, brut, bulge, bushel, butcher, cade, cader, cailleach, cain, cantankerous, car, carpenter, carry, cateran, cellidh, change, clachan, clan, claymore, colt, colleen, coracle, corgi, corn, corrie, cromlech, cross, cromw, Dal Eireann, dolmen, dour, drape, druid, dulse, eisteddfod, elvan, embassy, encumber, eric, esker, fail, fewterer, filibeg, fiorin, flannel, flummery, font, frown, gard, gaddiard, galloglass, gallon, galore, garron, garrote, garter, gelt, gillie, glean, glen, glib, gob, gob, goblet, gombeen, gossan, gowin, gravel, grollach, grave, gravel, grig, grouse, growan, guillem, gull, hubbub, ieroe, inch, ingle, iron, javelin, keen, keen, kern, kex, kibe, killas, kish, kistvaen, knock, kyle, lance, lay, leach, lee, leprechaun, linn, loch, lochan, lough, lough, loy, lozenge, lymphad, marl, menhir, messan, metheglin, minaudiere, mine, minion, mooch, morgay, morglay, mormaor, muley, mun, musha, mutton, och, ohone, ollamh, ough, ovate, oy, palfrey, partan, peat, pen, pendragon, pennill, peulvan, pibroch, piece, piket, pillion, plaid, pollan, port, poteen, ptarmigan, puffin, punt, quaich, rap, rapparee, rath, rich, ruche, sagum, say, scraw, sennachie, shamrock, shanty, shebeen, sivvens, skene, slat. Slob, slogs, sock, socket, sonsy, sorran, spalpeen, splechan, sporrin, spreach, stannum, strathe, tais(ch), tan, tanist, termon, tinche, tocher, toman, tope, tor, Tory, towan, tres, trouse, truant, tuath, tun, union pipes, usquebaugh, vassal, vendace, vergobret, vug, warren, weem, whisky, wirra, wrasse (Ito, 2008).
The 128 words found in the WB are as follows:

ach, ambassador, anchor, arrah, ass, bal, bannock, banshee, bard, basin, bawn, beak, beet, Beltane, betony, billet, bijou, bin, biretta, bodkin, bog, bothy, brasserie, brogue, brut, bulge, bullace, bushel, butcher, caber, cade, cader, cairn, cam, cantankerous, car, carpenter, carry, cashel, change, clan, claymore, cleave, cobe, colleen, coracle, corgi, corm, corrie, crannog, cromlech, cross, currach, Dail Eireann, dolmen, dour, drape, druid, eisteddfod, embassy, encumber, fail, flannel, flummery, frown, gallon, galore, garter, gillie, glen, goblet, gouge, gravel, grouse, guillem, gull, hog, hubbub, ingle, iron, javelin, keen, kibe, lance, lawn, leprechaun, linn, loch, lochan, lozenge, marl, menhir, meth EGLIN, mine, minion, mooch, mullein, mutton, och, peat, pendragon, piece, pillion, plaid, pollan, poteen, ptarmigan, puffin, rath, rich, ruche, say, shamrock, shanty, slogan, socket, tan, toman, tor, Tory, trews, truant, vassal, vendace, vug, warren, whisky, wrasse.

5. RESULTS

5.1 163 Words Identified By Native Speakers Of English

In a previous study, the 313 word-list was divided into four groups according to provenance: Celtic words from Irish, Celtic words from Scottish Gaelic, Celtic words from French, and Celtic words of Brythonic origin. The subjects completing the questionnaire were adults, mainly with a background of higher education, and who were, at the time, English teachers in universities. The nationalities of the subjects were mainly American, English and British. The number of subjects for each questionnaire included 25 males and 25 females (Ito, 2006).

A matching type of questionnaire and a word completion type of questionnaire were made for each of the four different groups of words listed above. The matching portion required the subjects to identify definitions for the different words listed in a group. The completion part required subjects to fill in a blank space with one word for each of the definitions provided. As a clue, the initial two letters for each word was given.

Out of 313 words, 163 words were correctly identified by the native speakers of English on the matching part of the questionnaire; 123 words were identified correctly on the completion portion. The 163 words from matching test were used in the present study because this list also includes those 123 words from the completion part.

As mentioned above, the number of subjects completing the questionnaire was 50; therefore, the maximum frequency of identification by the subjects is 50; the minimum being one. The researcher made a list of the 163 words with identification frequency by the subjects, and then added the frequencies from the BNC and WB corpora. The 163-word list is as follows:

airt, ambassador, andiron, arpent, arrah, bannock, banshee, bard, basin, beak, Beltane, betony, bijou, billet, bin, bog, bonaght, bonnyclabber, brair, brasserie, brogan, brogue, brut, buckeen, bulge, bullace, bushel, butcher, caber, cade, cailleach, cairn, cantankerous, car, carpenter, carrow, carry, change, clabber, clan, claymore, cockabondy, colleen, coracle, corgi, corn, coronach, corrie,
crannog, crine, cromlech, crostarie, currach, cwm, dalt, deasil, dillue, dolmen, dour, drape, druid, duan, dudeen, dulse, duniwassal, eisteddfod, elvan, embassy, encumber, esker, fewterer, filibeg, flannel, flummery, frown, galliard, gallon, galore, garron, garty, gleam, glem, glosk, gossan, gore, grave, gravel, grig, grise, growan, gull, ingle, javelin, keen, kern, kibe, killas, kistvaen, kyle, lance, lawn, leprechaun, loch, lochan, loy, lozenge, marl, mavourneen, menhir, methhegin, minaudiere, mine, minion, mook, morglay, mormaor, mullein, mutton, och, ohone, ouch, palfrey, pendragon, pennill, pibroch, piece, pikelet, pillion, pled, poteen, ptarmigan, puffin, quaich, ruche, sennachie, shamrock, shanty, shebeen, slogan, socket, sowens, spalpeen, spleuchan, strath, taish, tanist, tocher, toman, tor, Tory, trews, truant, union pipes, usquebaugh, vassal, vendace, vug, warren, whisky, wirra, wrasse.

5.2 Frequencies by BNC and WB

When considering these frequencies, it is important to remember that some of the words from the list above are only found in one of the two corpora, BNC or WB, while others are not found in either of them. Out of the 163 words, 112 words are found in the BNC, and 85 words are found in the WB. Thus, the 163-word list was reduced to 112 words in order to make a comparison with the frequencies in the BNC.

The 112-word list with frequencies by the subjects and those in the BNC and WB is as follows:

<table>
<thead>
<tr>
<th>item</th>
<th>S</th>
<th>BNC</th>
<th>WB</th>
<th>item</th>
<th>S</th>
<th>BNC</th>
<th>WB</th>
<th>item</th>
<th>S</th>
<th>BNC</th>
<th>WB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ambassador</td>
<td>48</td>
<td>1346</td>
<td>1273</td>
<td>dolmen</td>
<td>33</td>
<td>26</td>
<td>6</td>
<td>mine</td>
<td>35</td>
<td>1150</td>
<td>644</td>
</tr>
<tr>
<td>andiron</td>
<td>29</td>
<td>2</td>
<td>0</td>
<td>dour</td>
<td>28</td>
<td>152</td>
<td>125</td>
<td>minion</td>
<td>11</td>
<td>79</td>
<td>30</td>
</tr>
<tr>
<td>arpent</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>druid</td>
<td>17</td>
<td>182</td>
<td>25</td>
<td>moom</td>
<td>4</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td>bannock</td>
<td>17</td>
<td>9</td>
<td>6</td>
<td>dulse</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>moraor</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>banshee</td>
<td>36</td>
<td>47</td>
<td>18</td>
<td>eisteddfod</td>
<td>21</td>
<td>48</td>
<td>70</td>
<td>mullein</td>
<td>3</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>bard</td>
<td>18</td>
<td>143</td>
<td>241</td>
<td>elvan</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>mutton</td>
<td>49</td>
<td>162</td>
<td>85</td>
</tr>
<tr>
<td>basin</td>
<td>33</td>
<td>1511</td>
<td>389</td>
<td>eisteddfod</td>
<td>21</td>
<td>48</td>
<td>70</td>
<td>mullie</td>
<td>3</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>beak</td>
<td>50</td>
<td>388</td>
<td>136</td>
<td>embassy</td>
<td>43</td>
<td>1491</td>
<td>1557</td>
<td>och</td>
<td>13</td>
<td>208</td>
<td>23</td>
</tr>
<tr>
<td>Beltane</td>
<td>12</td>
<td>8</td>
<td>2</td>
<td>encumber</td>
<td>37</td>
<td>4</td>
<td>21</td>
<td>palfrey</td>
<td>18</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>betony</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>filibeg</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>pibroch</td>
<td>6</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>bijou</td>
<td>28</td>
<td>26</td>
<td>12</td>
<td>flannel</td>
<td>41</td>
<td>243</td>
<td>174</td>
<td>piece</td>
<td>45</td>
<td>14145</td>
<td>7467</td>
</tr>
<tr>
<td>billet</td>
<td>6</td>
<td>98</td>
<td>18</td>
<td>flummery</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>pillar</td>
<td>12</td>
<td>47</td>
<td>27</td>
</tr>
<tr>
<td>bin</td>
<td>28</td>
<td>1251</td>
<td>591</td>
<td>gallon</td>
<td>44</td>
<td>852</td>
<td>529</td>
<td>plaid</td>
<td>21</td>
<td>116</td>
<td>127</td>
</tr>
<tr>
<td>bog</td>
<td>45</td>
<td>522</td>
<td>151</td>
<td>garron</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>poteen</td>
<td>9</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>brasserie</td>
<td>31</td>
<td>73</td>
<td>80</td>
<td>gare</td>
<td>23</td>
<td>125</td>
<td>122</td>
<td>ptarmigan</td>
<td>27</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>brogan</td>
<td>12</td>
<td>2</td>
<td>0</td>
<td>gillie</td>
<td>10</td>
<td>13</td>
<td>2</td>
<td>sennachie</td>
<td>9</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>brogue</td>
<td>2</td>
<td>94</td>
<td>38</td>
<td>glee</td>
<td>30</td>
<td>190</td>
<td>25</td>
<td>shamrock</td>
<td>39</td>
<td>91</td>
<td>143</td>
</tr>
<tr>
<td>brut</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>glosk</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>sconnachie</td>
<td>9</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>bulge</td>
<td>1</td>
<td>226</td>
<td>75</td>
<td>glee</td>
<td>24</td>
<td>367</td>
<td>46</td>
<td>shanty</td>
<td>21</td>
<td>118</td>
<td>110</td>
</tr>
<tr>
<td>bushel</td>
<td>27</td>
<td>45</td>
<td>23</td>
<td>glee</td>
<td>24</td>
<td>367</td>
<td>46</td>
<td>shanty</td>
<td>21</td>
<td>118</td>
<td>110</td>
</tr>
<tr>
<td>butcher</td>
<td>48</td>
<td>476</td>
<td>275</td>
<td>gossan</td>
<td>47</td>
<td>181</td>
<td>55</td>
<td>shebeen</td>
<td>7</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>caber</td>
<td>20</td>
<td>8</td>
<td>8</td>
<td>gossan</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>slogan</td>
<td>7</td>
<td>801</td>
<td>539</td>
</tr>
<tr>
<td>cailleach</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>gare</td>
<td>19</td>
<td>45</td>
<td>5</td>
<td>socket</td>
<td>21</td>
<td>929</td>
<td>270</td>
</tr>
<tr>
<td>caim</td>
<td>32</td>
<td>198</td>
<td>31</td>
<td>gare</td>
<td>47</td>
<td>1245</td>
<td>303</td>
<td>sowens</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>cantankerous</td>
<td>17</td>
<td>40</td>
<td>24</td>
<td>gare</td>
<td>38</td>
<td>203</td>
<td>82</td>
<td>spalpeen</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>car</td>
<td>48</td>
<td>33936</td>
<td>23643</td>
<td>gare</td>
<td>22</td>
<td>513</td>
<td>61</td>
<td>strath</td>
<td>3</td>
<td>24</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: Frequencies in BNC and WB
As mentioned above, numbers under the label S come from earlier research that was conducted on native speakers of English. The number of subjects was 50; therefore, the maximum frequency is 50; the minimum is one (Ito, 2006). Numbers under the labels of BNC and WB are frequencies from the one-million-word text in each corpus (Shogakukan BNC, 2008; Shogakukan Wordbanks, 2008).

6. DISCUSSION
Since main purpose of the present study is to examine if the results obtained in an earlier study (Ito, 2006) are in accord with the frequencies of the identified words in the BNC and WB, a statistical analysis was applied to the results.

A correlation between the frequencies in the BNC and those in the WB was 0.97, which means that both corpora show the same tendency to exhibit popularity among native speakers of English. However, a correlation of the results between the subjects from the Celtic word study (Ito, 2006) and those taken from the BNC exhibited 0.31, which does not show much of a correlation between the two groups. Next, a rank order was applied to the list. This list is as follows;

Table 2: Word List in Rank Order

| item       | S  | BNC | item       | S  | BNC | item       | S  | BNC | item       | S  | BNC |
|------------|----|-----|------------|----|-----|------------|----|-----|------------|----|-----|------------|----|-----|------------|----|-----|
| ambassador | 4  | 11  | change     | 12 | 2   | glean      | 42 | 34  | palfrey    | 61 | 79  |
| andiron    | 44 | 96  | clan       | 21 | 19  | glen       | 52 | 25  | pibroch    | 84 | 96  |
| arpent     | 89 | 96  | claymore   | 61 | 71  | goblet     | 11 | 36  | piece      | 14 | 4   |
| bannock    | 64 | 78  | colleen    | 33 | 73  | gossan     | 106| 96  | pillion     | 70 | 59  |
| banshee    | 33 | 59  | coracle    | 50 | 69  | gouge      | 60 | 61  | plaid      | 55 | 43  |
| bard       | 61 | 39  | corgi      | 14 | 56  | gravel     | 9  | 13  | poteen     | 77 | 87  |
| basin      | 36 | 8   | corn       | 106| 77  | grouse     | 30 | 32  | ptarmigan  | 48 | 71  |
| beak       | 1  | 24  | corrie     | 81 | 61  | gull       | 54 | 21  | puffin     | 38 | 51  |
| Beltane    | 70 | 79  | cromlech   | 89 | 89  | ingle      | 80 | 105 | quaich     | 98 | 85  |
| betony     | 86 | 94  | cwm        | 75 | 87  | javelin    | 14 | 53  | sennachie  | 77 | 96  |
| bijou      | 45 | 66  | dolmen     | 36 | 66  | keen       | 67 | 91  | shamrock   | 29 | 47  |
| billet     | 84 | 45  | dour       | 45 | 38  | keen       | 73 | 96  | shanty     | 55 | 42  |
| bin        | 45 | 12  | drape      | 64 | 25  | killas     | 106| 79  | shebeen    | 81 | 91  |
Correlation between the subjects' results and the BNC results according to rank order analysis was 0.75, which means that the two types of results are highly correlated to each other.

### 7. CONCLUSION

A conclusion of the results is as follows:

1. The degree of recognizing English words of Celtic origin by native speakers of English on a questionnaire is similar to the degree of frequencies found in the BNC and WB. This was obtained by rank-order correlation, not by comparing frequencies between the results of the questionnaire to those of the BNC and WB.

2. The words collected from the dictionaries mentioned in the study are not all found in the BNC and WB corpora. This implies that the dictionaries cite words that even large corpora do not include.

3. Predictably, even educated adult native speakers of English do not know all of the words of Celtic origin in English dictionaries.

4. Identifying some of the original 313 words was impossible in the corpora because of completely different meanings in one, and the same spelling under the same part of speech. For example, *down* was impossible to obtain a frequency as a word of Celtic origin because one *down* means 'soft, fine, fluffy feathers,' and the other *down* means 'a gently rolling hill.' Thus, corpora are not always suitable for lexical studies.

### NOTE

(1) This study was supported in part by a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (Foundation Studies (C) ; No. 20520438).

### REFERENCES

(4) Shogakukan BNC Corpus (2008).