The potential of learner corpora for pedagogical lexicography

Yukio Tono

Meikai University
y.tono@meikai.ac.jp

Abstract

There is a growing awareness that a more data-oriented approach can greatly improve the design and content of pedagogical dictionaries. In this paper, I argue that three types of data, general language corpora, research data on dictionary use, and a corpus of learner language, can significantly improve learners' dictionaries. I will review my past research in these areas briefly and focus particularly on how learner corpora can be used to better improve the kinds of information provided to learners in pedagogical dictionaries.

Research paradigm shift in pedagogical lexicography

Lexicography covers a broad range of interdisciplinary areas in linguistics, reference sciences, language teaching and learning, and more recently natural language processing. A common approach in lexicography is to try to apply knowledge and facts found in these disparate areas to the production of dictionaries. These facts are mainly concerned with linguistic observations or theoretical analyses of the system of a language as well as its use. Linguists study languages and inform lexicographers on better or more innovative ways of describing a word, providing usage information, giving illustrative examples and so on.

In the last two decades, however, this traditional approach to dictionary-making has been taken over by a more data-oriented approach to lexicography, which is based upon empirical research on language corpora, dictionary users and language learners. This is particularly true in the field of pedagogical lexicography. In this paper, I would like to briefly summarize recent developments in pedagogical lexicography with reference to the three areas mentioned above (language corpora, dictionary users and language learners) and argue that significant improvements in user-friendliness will be achieved by using second language (L2 henceforth) learner corpora to inform the making of learners' dictionaries.

A data-oriented approach to pedagogical lexicography

In a sense, lexicography has always been 'data oriented.' Lexicographers investigate the use of words and phrases by what lexicographers at Merriam-Webster's call 'reading and marking.' The way lexicographers access language use data, however, has drastically changed since computerised corpora became available in the early 1960s. The first fully corpus-based monolingual dictionary was the *COBUILD English Dictionary* (1987), which was radically different from existing monolingual learners' dictionaries and enthusiastically welcomed among linguists and language educators in Japan as well as in the rest of the world. All the other major monolingual learner's dictionaries have more or less followed this trend and lexicographers have been using KWIC (keyword in context) concordances as their primary tool for finding out how a word behaves ever since. After Church and Hanks (1989) introduced the notion of

Mutual Information (a measure of the salience of the association between any two words), lexicographers became more interested in identifying statistically salient collocates. Every publisher started having its own corpus (e.g. the Cambridge International Corpus, the Longman Corpus Network) and even its own corpus query system (a set of tools to help lexicographers use corpus data effectively). By 1995, the revised Big 4 (*COBUILD*, *OALD*, *LDOCE*, and *CIDE*) were all claiming that they were 'corpus-based.' Since then, using corpora for dictionary-making has become standard practice, at least for the major dictionary publishers in the UK.

Recently, a more sophisticated approach to using corpora has been proposed, mainly to deal with larger sets of data or more detailed grammatical relations in the text (e.g. the Sketch Engine (Kilgarriff, et al. 2004), the Shogakukan Language Toolbox (Nakamura and Tono 2003) among others). Recent papers presented at Euralex 2004 show that dictionary publishers are shifting their attention from using general mega-corpora to more specialised corpora and from the ordinary use of corpora to more purpose-specific uses.

Another important area which underwent a marked shift in the past two decades is research into dictionary use. Up until the 1980s, very little attention was paid to the needs and skills of dictionary users. Reinhart Hartmann was one of the first to enlighten us on the importance of research into dictionary use (Hartmann 1979; 1983). I myself was one of the few researchers who started to conduct experimental studies on dictionary reference skills in the 1980s (Tono 1984, 1986, 1988). There is now a growing body of literature in this field and major works are reviewed in my book in *Lexicographica* series (Tono 2001).

Whilst lexicographers are aware of the importance of user studies, it takes time to apply research findings to the actual production of dictionaries. Some notable successful applications of research findings to dictionary-making include the provision of a 'menu' at the beginning of dictionary entries. With the menu, users can first browse through the various meanings of any given word, which is a feature widely introduced after a study confirmed the fact that the users only look at the beginning of dictionary entries (Tono 1984). Dictionary makers are also taking note of research in dictionary use where the effectiveness of newly introduced organizational devices are put to the test. For example, I conducted an experiment to investigate the effects of 'Signposts' in *LDOCE* and 'Guidewords' in *CIDE* and found that the terms used for signposts in *LDOCE* were more effective in directing users to the right meanings while the terms used for Guidewords were often too abstract to 'signpost' meanings in the dictionary (Tono 1997).

More recently, there has been a renewed interest in the role of dictionaries in language learning since electronic dictionaries began growing in popularity in Japan several years ago. The market is constantly increasing in size, and major manufacturers such as Casio, Seiko, Sharp, Canon and Sony all spend time and money on the development of new types of hand-held e-dictionaries. As the number of university and high school students who own pocket e-dictionaries grows rapidly, more research has been conducted on the effects of using pocket e-dictionaries in reading and writing. The dictionary workshop organized by the JACET Lexicography SIG this March was very well attended, where approximately sixty paper presentations were given. The time is ripe for further research on the effects of using pocket e-dictionaries in terms of the medium or interface (paper vs. electronic), L2 vocabulary learning, and dictionary skills training.

The above two factors, the advent of language corpora and research into dictionary use, have contributed greatly to the improvement of learners' dictionaries. A third area which I will now focus on is the study of language learners themselves. Dictionaries serve many different purposes. Pedagogical lexicography is mainly concerned with dictionaries designed to help foreign learners of the language. Language learners as dictionary users need to be investigated more seriously. Pedagogical lexicography should take into account L2 learners' learning habits, learning styles, learning strategies and learning processes. There is a large body of research in the field of foreign language learning and second language acquisition (SLA henceforth), but unfortunately very little effort has been made to apply SLA research findings to the study of dictionary-making and on how language learners actually use dictionaries.

For the past twenty years, I have been conducting research in all the above three areas. At the beginning of my research career, my primary interest was in the role of dictionary use in language learning. When the *COBUILD English Dictionary* was first published, I realised the potential of corpus-based research, and this led me to pursue my doctoral research at Lancaster in the 1990s. There I learned about the various branches of corpus-based research and saw examples of corpus applications in different fields. I became convinced that the use of corpora would make a major difference in the field of English language teaching in Japan. Before going to Lancaster, I had collected English essays written by Japanese learners of English as part of a large research project investigating the effects of teacher feedback in L2 writing. I started turning this valuable data into a corpus so that I could more systematically investigate the characteristic features of the writing of Japanese learners of English.

Learner corpora and L2 lexicography

A learner corpus is a collection of speech or writing by foreign language learners. By looking at the learner performance data, we can find many interesting patterns of use which are quite different from those of native speakers. In many cases, these differences are due to the fact that learners are still in the process of acquiring a language, and they naturally make errors or mistakes. Studying learner errors is not new. The research area called 'Error Analysis' has been around for more than 30 years. What is new is that we can now employ the techniques of corpus linguistics to investigate learner language in a more empirical, data-based fashion. What sorts of information can we extract from learner corpora? How can we apply such findings to pedagogical dictionary-making? Let me describe some of these areas in detail.

(a) Source of vocabulary selection for L2 learners

One way of analysing learner corpora is to create a wordlist from learner writing and compare it with a comparable wordlist derived from the writing of native speakers. I have been working for NHK (Nihon Hoso Kyokai, Japan Broadcasting Center) in developing a television English conversation program (titled '*Hyakugo de sutato eikaiwa*' which means *'Let's start with 100 basic words in English'*). This program is unique in the sense that it is probably the first 'corpus-based' English conversation program on TV. It consists of a hundred lessons based on 100 key vocabulary items which were chosen based on corpus analysis. It is a well-known fact that the high-frequency lexical items in English (or any language) will cover a very high proportion of the words in any text; the most frequent 100 words (lemmas) in English, for example, will cover approximately 70% of words in a spoken corpus (e.g., the

spoken component of the British National Corpus). Many of these are core lexical items (verbs, prepositions, personal and *wh*-pronouns, determiners, adverbs and conjunctions) that play a crucial role in constructing basic English structures (See Lee 2001 for more discussion on core vocabulary). There are relatively few nouns (only six!) and adjectives in the top 100 words. In this TV program, I focus on the most frequent 100 keywords and design the program around a lexical syllabus. As I worked on this program, I became convinced that beginning-level students should study a set of basic core vocabulary again and again in a series of different language tasks. These core vocabulary items are at the heart of English grammar and are rich in meanings and functions, and it takes time to acquire a satisfactory productive and receptive grasp of them.

One hundred words might seem too few in number and some people claim that to be functional in English one should know at least the top 2000 words, which would typically cover about 90% of the words in a spoken corpus. Leftover words (i.e. those below the 2000 word level) are said to be mostly those which are affected by particular topics or situations and which can therefore be learned independently from the first 2000 basic items. However, how exactly can we determine the next set of words to learn (after the first 2000)? In an EFL environment like Japan, most L2 input will come from the classroom especially for beginning-level learners. The language spoken and written in the classroom is different from that of everyday conversations encountered by native speakers. It is natural, therefore, that the vocabulary covered in classroom settings will be different from those used in everyday life in Britain, and EFL learners' dictionaries should meet the specific communicative needs of L2 learners in terms of vocabulary selection. For such purposes, learner corpora collected from particular L2 learner groups would be most useful. By comparing well-balanced learner corpora with native-speaker corpora, both in spoken and written modes, we can possibly identify a list of words which are significantly more frequently used by L2 learners. These are the candidate words that learners want to express in English. In this way, we could exploit learner corpora to improve the selection of vocabulary for more user-friendly bilingual dictionaries.

(b) Identifying L2 learners' common errors

Recently, monolingual dictionaries such as *LDOCE*, *CALD*, and *Longman Essential Activator* all feature common learner errors as part of the usage information. The primary aim of this information is to give learners information on correct usage based on common errors as shown in the learner corpus data collected by the dictionary publishers. The types of errors highlighted in learners' dictionaries may be classified as follows:

(a) Lexical choice

e.g. Do not say 'injure someone's health'. Say 'damage someone's health.'
(LDOCE)

e.g. The words *'not ... either'* are used to add another piece of negative information.

Helen didn't enjoy it either.

~~Helen didn't enjoy it too.~~ (CALD)

(b) Verb forms

e.g. You can 'have problems doing something'. Do not use 'to do' (LDOCE)

(c) Verb patterns

e.g. You propose something to someone: *He proposed a possible solution to me.*

(NOT He proposed me a possible solution.) (LDOCE)

(d) Word position

e.g. Especially never comes at the start of a sentence: *He loves fruit. He especially likes kiwis.*

(NOT *Especially he likes ...*) (LDOCE)

(e) Grammatical/lexical collocation

e.g. Be careful to use the correct verb.

I have to make a speech.

~~I have to do a speech.~~ (CALD)

Whilst such error information is valuable in itself, the way the information is provided in pedagogical dictionaries still needs to be refined. Firstly, the selection of errors is not always appropriate. Some information is too basic for those who would use monolingual dictionaries. There is a trend to provide simple error information in beginners' monolingual dictionaries such as *Longman Active Study Dictionary* (*LASD*) and *Cambridge Learner's Dictionary* (*CLD*), but most of those who would dare to use a monolingual dictionary are likely to be already familiar with such information. The error information should be tuned to the level of learners who would venture to use monolingual dictionaries.

Secondly, it is difficult to deal with L1-related errors in general-purpose monolingual dictionaries which are not aimed at particular L1 speakers. For example, in the case of Japanese-speaking learners of English, incorrect sentences such as '*My house is Kyoto*' are quite common because of its parallel Japanese sentence '*Watashi no uchi ha Kyoto desu*'. This type of error cannot be adequately described in monolingual dictionaries because it is often caused by learners' L1 knowledge and error patterns are different from L1 to L1. Thus, this type of L1-related errors should be treated more extensively in bilingual learners' dictionaries.

(c) Identifying the weak areas of learners: underuse of collocations

It is not sufficient to use learner corpora to provide error information only. A more significant application of corpus-based techniques would be to show the gap in performance between native speakers and learners and so encourage learners to perform in a more target-like manner. One typical example would be the pattern of use of grammatical and lexical collocations. Table 1 shows the object noun collocates of the verb 'make' found in the British National Corpus and the Japanese EFL Learner (JEFLL) Corpus. As one can clearly see, Japanese EFL learners tend to use relatively concrete objects such as *money*, *food*, *friends*, and so on. These collocates can be regarded as free combinations with the verb *make* in the sense 'to produce.' On the other hand, native speakers use the verb *make* with more abstract nouns such as *sense*, *way*, *use*, *decision* and so on. Since phrases such as *make sense*, *make a decision*, etc.

are all highly frequent collocations used by native speakers of English, but constantly underused by Japanese EFL learners, it would be desirable to highlight these differences in dictionaries and to advise learners to use the keyword in a more target-like manner. One way to do this is to allocate more space to the item which needs more attention. In this particular case, one could describe the basic use of the verb *make* (i.e. the core meaning of 'produce') more extensively in a beginner's dictionary and give more space and treatment to the extended and often metaphorical meanings in advanced learner's dictionaries. In so doing, we would be taking into account the gap between native speakers and L2 learners (Tono 2001:203ff).

Rank by Freq.	BNC	JEFLL
1	sense	money
2	way	food
3	use	breakfast
4	decision	friends
5	mistake	story

Figure 1. Object noun collocates of the verb *make* in BNC and JEFLL

Profiling learner language for pedagogical lexicography

Finally, I would like to describe my on-going project on profiling learner language and examining its implications for pedagogical lexicography. I am currently working on two large corpora of Japanese EFL learners: one is called the SST Corpus (currently called the NICT JLE Corpus; approximately 2 million words; see Tono, et al. 2002), which is a corpus of 1,200 oral proficiency interviews taken as part of the Standard Speaking Test. Each transcript is based on a 15-minute speaking test. The second corpus is called the JEFLL Corpus (700,000 words; see Tono 2004 for more details), a corpus of free compositions by approximately 8,000 students (in-class, timed essays written without recourse to dictionaries). A project team is now working on the data of these two corpora and will publish the first report sometime toward the end of this year. This will be the first large-scale research project report based on Japanese EFL-learner corpora.

There is tremendous potential in the exploitation of these resources for improving EFL syllabuses and materials design. Using analysis tools such as the Sketch Engine or the Shogakukan Language Toolbox, it is now possible to gain an overall picture of learners' use of core and specialized vocabularies at various proficiency levels. This will provide very useful input for vocabulary learning theories and for syllabus design. We can identify not only the major error patterns of the learners but also the overuse and underuse of particular words, to which special attention could be drawn by way of usage notes or extra practice in relevant places in textbooks and reference materials.

If developmental errors are identified for each proficiency level, dictionaries can then be customized to specifically address the relevant weak points for different levels of users. Electronic dictionaries, in particular, could change their interfaces and even their content according to individual user settings. It would be ideal to have multiple-levels of information in a dictionary, leaving it to end-users to choose the level, amount and type of content they see according to their needs. At the moment, we have very little of this sort of proficiency level-based information, but as relevant corpora grow in size and coverage, the type of customizable dictionary described above should become a reality in the not-too-distant future. Pedagogical dictionaries should deal with

all the issues I have discussed in this paper and provide the kind of support detailed above. One last point to be made, however, is that proper dictionary training also needs to be given so that learners can learn to access and exploit such information for their own ends and thus become more successful language users.

References

Church, K. W. and P. Hanks (1989), Word association norms, mutual information and lexicography. *Proceedings of the 27th Annual Meeting of ACL*, Vancouver, pp.76-83.

Hartmann, R.R.K. (ed.) (1979), Dictionaries and Their Users, Papers from the 1978 BAAL Seminar on Lexicography. *Exeter Linguistic Studies*, Vol. 1. No.4. University of Exeter.

Hartmann, R.R.K. (ed.) (1983), *Lexicography: Principles and Practice*. London: Academic Press.

Lee, D. (2001), Defining core vocabulary and tracking its distribution across spoken and written genres: evidence of a gradience of variation from the British National Corpus. *Journal of English Linguistics*, Vol. 29(3), pp. 250-278.

Nakamura, T. and Y. Tono (2003), Lexical profiling using the Shogakukan Language Toolbox. In Murata, Yamada & Tono (eds.) *Dictionaries and Language Learning: How can Dictionaries Help Human & Machine Learning?* (ASIALEX 2003 Proceedings), pp. 170-176.

Tono, Y. (1984), *On the Dictionary User's Reference Skills*. Unpublished B.Ed. Dissertation. Tokyo: Tokyo Gakugei University.

Tono, Y. (1986), A scientific approach toward lexicography: the user perspective. *LEO* 15, pp.37-54.

Tono, Y. (1988), Can a dictionary help one read better? On the relationship between E.F.L. learners' dictionary reference skills and reading comprehension. In G. James (ed.) *Lexicographers and Their Works*. Exeter: University of Exeter Press, pp. 192-200.

Tono, Y., T. Kaneko, H. Isahara, T. Saiga and E. Izumi (2002), The Standard Speaking Test Corpus. *Studies in Lexicography* 11 (2), pp. 7-18.

Tono, Y. (2001), *Research on Dictionary Use in the Context of Foreign Language Learning*. (LEXICOGRAPHICA Series Maior 106). Tübingen: Max Niemeyer Verlag.

Tono, Y. (2004), Multiple comparisons of IL, L1 and TL corpora: the case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. In G. Aston, S. Bernardini, & D. Stewart (eds.) *Corpora and Language Learners*. Amsterdam: John Benjamins, pp. 45-66.

Pocket e-dictionaries in Japan: new developments

Yukio Tono **Miwa Nakamura** **Shinichi Yagi**
Meikai University
y.tono@meikai.ac.jp

Abstract

This study investigates the comparisons of misinterpretations between handheld electronic dictionary (ED) and printed dictionary (PD), which were analysed from the marketing data of the 2004 Pop Song Translating Competition in Japan. This competition was held by Seiko Instruments Inc (SII)., the pioneer maker of handheld dictionaries in Japan. For the past three years, the use of handheld dictionaries in high school has taken a steep increase, which indicates growing importance and need for handheld dictionaries in schools. As such it is imperative that the dictionary manufacturers seek to understand the needs of this growing market and continually develop user-friendly dictionaries. This comparative study found two plausible results.

First, the limited size of ED screens did not pose a significant factor in causing misinterpretations by the user. Second, for PD users, there was a high tendency to skip words more frequently than ED users when they met unknown words while translating.

The gap between the importance level indication in English-Japanese (E-J) dictionaries and the vocabulary in English textbooks.

Tsuyako Touno
Kwansei Gakuin University
touno-tsuyako@mug.biglobe.ne.jp

Abstract

In this study, I will choose some words from English textbooks in Japan and look into the importance level indication of those words in English-Japanese (E-J) dictionaries and discuss (1) what kind of vocabulary should be taught in Japanese junior and senior high schools and (2) how E-J dictionaries have to be improved to meet the needs of Japanese learners of English.

1. Introduction

Almost a hundred E-J dictionaries are currently on the market, and a dozen influential E-J dictionaries adopt a notable common practice of giving several thousand important or highly frequent words 'labels of importance' or at which stage of the learning process those words have to be learned. Although each dictionary has its own criteria for labeling those words, quite a few are given different importance labels from dictionary to dictionary.

There are also other facts in which vocabulary used in English textbooks are not listed in those influential dictionaries or simple words used in our daily lives are labeled as 'college level'.

After revealing these facts, I will compare the importance level indication of words in English textbooks and discuss the two research points shown in the abstract.

2. The importance level indication in learner's E-J dictionaries

2.1 The criteria for the selection of headwords

As for the selection of headwords of learner's E-J dictionaries, compilers of those dictionaries have their own criteria for the number of headwords and the level indication for each word. From the criteria of some E-J dictionaries, the following three facts can be understood.

(1) Each dictionary specifies its target users and has vocabulary which is necessary for them. Then it defines several thousand headwords as basic

important words, label them as such and gives them detailed description of their meanings and usages.

(2) Each dictionary states what kind of vocabulary is contained as headwords in the dictionary.

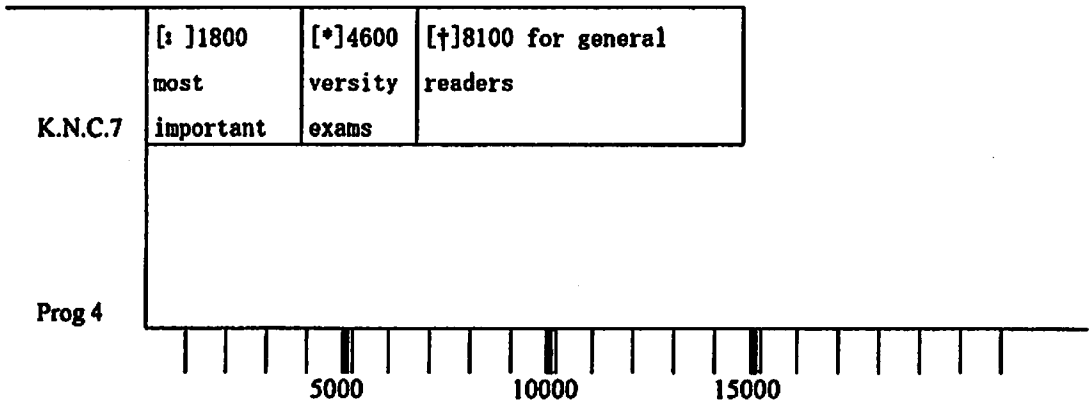
(3) Some dictionaries state what areas of vocabulary they have tried to contain most.

2.2 The importance level indication of some E-J dictionaries

The following table shows the importance level indication of 7 E-J dictionaries.

Table 1 Importance level indication of 7 E-J dictionaries names of the dictionaries

Y	[⁝]1300 for J.H	[:]1800 for S.H	[*]2200 important words	
W	[⁝]900 for J.H	[:]2900 for S.H	[*]3400 for U.S. and G.R.	[†]8000 for G.R.
G3	[:]1100 for J.H.	[*]3400 for S.H.	[†]5100 for U.S. and G.R.	
Lexis	[⁝]1300 for J.H. and S.H.	[:]1900 for S.H.	[* in red] 6000 for center exam	[* in black] 1900 for state universities
	[⁝]1100 must	[:]2000 most important	[*]5000 important	
A.F.	[:]2000 most important	[*]5000 important	[†]8000 next important	



cf.) J.H. indicates junior high school students / S.H. indicates senior high school students
U.S. indicates university students / G.R. indicates general readers

From this table, we understand that there are three differences in the criteria of the importance level indication.

(1) The way headwords are divided [e.g. Are they labeled according to the stage of learning process or are they labeled just according to their importance?]
(2) The number of labels each dictionary has.
(3) The number of headwords each label contains.

3. How the words in high school English textbooks are labeled in E-J dictionaries

In this section, I will choose some words from the viewpoint of their meaning and compare their level indication among seven E-J dictionaries.

3.1 Daily words

I chose 278 daily words from the word lists of the textbooks and looked into their level indication and characteristics.

First, the three words, 'piroshki', 'wonton' and 'pa' were not listed as headwords in some of the dictionaries. It should be noted that the first two words are the names of food. As an example, I will show you a chart of the level indication of daily words in a textbook.

Chart 1 The level indication of daily words in a textbook

eel	o	*	†	* in red	o	*	*
heart attack	□*1	□*1	□*1	□*1	o	*	*
lasagna	o	o	o	o	o	o	o
octopus	o	†	†	o	o	o	o
piroshki	×	×	×	o	o	o	o
scone	o	†	o	o	o	†	†
sushi	o	o	o	o	o	o	o
wonton soup	o*2	×	×	o*2	o*2	o*2	o*2

meaning of symbols) o: listed as a headword /×:not listed as a headword
*1: listed as a compound of 'heart' / *2: listed only as 'wonton'

Then I checked the level indication of the 278 words and found that 169 words among them (about 60 %) had a level for university students or general readers. Many of the 169 words were derivatives or compound words and were related to food and cooking, with the percentage of 15 % and 14 % respectively. From these facts, I learned that basic daily words such as 'toothpaste' and 'kilogram' had a level for university students or general readers in the E-J dictionaries, and I think the compilers of the dictionaries should try to rethink the level indication of those words for their necessity in everyday life.

3.2 Words for high-school life

I chose 51 words from an Oral Communication textbook and compared their level indication. The words are related to everyday life of high school students. Again, such words as 'textbook', 'eraser', 'picnic', 'pajamas' had a level indication for university students or above. And like the case with daily words, compound words had higher level indication.

Then I looked into the situations in which a lot of words with high level indication were used. The result showed that high school textbooks need to include vocabulary for various kinds of activities other than school life, though such vocabulary had high level indication. From these facts, it can be said that the difference in the level indication for the 51 words shows that there is a difference in the number of words for high school students and in the number of words for their school life each dictionary contains.

4. Comparison of vocabulary between English textbooks in Japan and coursebooks for EFL/ESL students

Next, I will compare the vocabulary between 16 English textbooks in Japan (henceforth 16 textbooks) and a series of coursebooks with four different levels (henceforth the coursebooks). The reasons for the selection of this series are: (1) the original series 'Headway' received an excellent teaching material award and (2) this series and English textbooks have a similarity in that they both are organized to improve students' four skills of English. The following is the original data for the two word lists.

	16 textbooks	the coursebooks
total types	6117	6242
total tokens	81686	118761

First, I deleted proper names such as the names of people, countries from the lists and compared the percentage of the proper names. 16 textbooks had about 11.2 % and the coursebooks had about 18.4 % of proper names in each list. Then I looked into the differences in both lists. As for morphological aspects, vocabulary in the coursebooks had many derivatives and compound words. It also had many prefixes and suffixes for users of the books to increase their vocabulary. As for semantic aspects, the coursebooks contained many words for daily life (ex. air-conditioned, ATM), and more grammatical words than the 16 textbooks do.

From these results, I suggest that English textbooks in Japan should have more derivatives, compound words and affixes for Japanese students of English to increase their vocabulary and they should contain more words for food.

5. Suggestions for the improvements of E-J dictionaries

Frequency is an important criterion for the selection of headwords and the order of their definitions. Carter (1998:46) raises three problems associated with frequency counts. First, as for the lemmatization problem, E-J dictionaries don't list all the meanings of a homograph, and frequency or historical order is used to arrange homographs as headwords. I suggest that E-J dictionaries should choose and list homographs according to their frequency and the needs of learners. I will use BNC corpus, English textbooks in Japan, TOEIC test and JACET 8000 word list to get the relative frequency of English words. Second, as for inflections and derivatives of words, I think few E-J dictionaries indicate derivatives of words. So as many derivatives of important high-frequency words should be listed as possible to increase learners' vocabulary. Third, Stubbs (2001:30-31) defines 'lexical item' as follows; 'The term 'lexical item' is therefore used to cover a range of individual words and phrases'. We should interpret 'word' to suggest that it implies individual words and phrases so that important phrases are listed as run-ons or headwords in E-J dictionaries.

I think we should take into account not only the frequency of words but also the needs of learners and the critical judgment of compilers of E-J dictionaries in choosing headwords of an E-J dictionary.

6. Conclusion

The importance level indication of headwords in E-J dictionaries differs from dictionary to dictionary. I think we should look into words from various viewpoints (e.g. their place in corpora, familiarity, range of use) and select them as headwords as objectively as possible.

References

Carter, R. 1998. *Vocabulary* 2nd edn London : Routledge
Stubbs, M. 2001. *Words and Phrases.* Oxford : Blackwell Publishers Inc.

English-Japanese dictionary
Asano Hiroshi, et al. 2002. *Advanced Favorite English-Japanese Dictionary.* Tokyo:Tokyo Shoseki
Hanamoto Kingo, et al. 2003. *Lexis English-Japanese Dictionary.* Tokyo:Obunsha
Inoue Nagayuki, et al. 2003. *Wisdom English-Japanese Dictionary.* Tokyo:Sanseido
Konishi Tomoshichi, et al. 2001. *Genius English-Japanese Dictionary* 3rd edn. Tokyo:Taishukan Shoten
Kunihiro Tetsuya, et al. 2003. *Progressive English-Japanese Dictionary* 4th edn. Tokyo:Shogakukan
Takebayashi Shigeru, et al. 2003. *New College English-Japanese Dictionary* 7th edn. Tokyo: Kenkyusha
Yagi katsumasa, et al. 2004. *Youth Progressive English-Japanese Dictionary.* Tokyo:Shogakukan

Word list
The Japan Association of College English Teachers (JACET). 2003. *JACET List of 8000 Basic Words.* Tokyo : JACET

Coursebooks
John and Liz Soars. 2002. *American Headway (Starter, 1, 2, 3).* New York:Oxford University Press

English textbooks
Hanamoto Kingo et al. 1999. *Step English Revised.* Tokyo:Obunsha
Kaneda Michikazu et al. 2003. *Acorn English Course.* Osaka:Keirinkan
Matsumura Kenichi et al. 2000. *Genius English Course Revised.* Tokyo:Taishukan Shoten

Morizumi Mamoru et al. 2002. *Exceed English Series.* Tokyo:Sanseido / Nakau Minoru et al. 2001. *Quest English Course 2* Revised. Tokyo: Shogaku Tosho
Ogata Takao et al. 1999. *Access to English.* Tokyo:Kaitakusha
Suenaga Kuniaki et al 1998. *Unicorn English Course.* Kyoto: Buneido
Yoneyame Asaji et al. 2002. *Clipper English Course 2* Revised. Tokyo:Taishukan Shoten
Yoshida Kensaku et al. 1997. *Birdland Oral Communication A* 2nd edn Kyoto:Buneido

16 textbooks
Hanamoto Kingo et al. 2000. *Step English* Revised. Tokyo:Obunsha
Hanamoto Kingo et al. 2002. *Step English.* Tokyo:Obunsha
Hanamoto Kingo et al. 1999. *Step English* Revised. Tokyo:Obunsha
Hisamura Ken et al. 2002. *All Aboard!.* Tokyo:Tokyo Shoseki
Hisamura Ken et al. 2003. *All Aboard!.* Tokyo:Tokyo Shoseki
Ikeda Satoshi et al. 2002. *Vista.* Tokyo: Sanseido
Jinbo Hisatake et al. 1997. *Hello there! Oral Communication A* .Tokyo:Tokyo Shoseki
Jinbo Hisatake et al. 2002. *Power On.* Tokyo:Tokyo Shoseki
Matsumura Kenichi et al. 2000. *Genius English Course.* Revised Tokyo:Taishukan Shoten
Matsumura Kenichi et al. 2000. *Genius English Course* Revised. Tokyo:Taishukan Shoten
Morizumi Mamoru et al. 2002. *Exceed.* Tokyo:Sanseido
Nakau Minoru et al. 2002. *Quest English Course 1* Revised. Tokyo:Shogaku Tosho
Sano Masayuki et al. 2002. *Captain.* Tokyo:Taishukan Shoten
Takanashi Kenkichi et al. 1999. *Daily English Reading.* Tokyo:Ikeda Shoten
Yoneyama Asaji et al. 2000. *Clipper.* Tokyo:Taishukan Shoten
Yoshida Kensaku et al. 1999. *Birdland Oral Communication.* A. Kyoto:Buneido

Reverse indexing and customization – future trends in bilingualized dictionaries

Hugo T. Y. Tseng

Soochow University
hugozeng@scu.edu.tw

Abstract

In spite of all the kudos garnered from both students and teachers, monolingual English learner's dictionaries (MELDs) pose an inherent problem of not addressing target users' specific needs. Their bilingualized adaptations, however, enjoy unparalleled advantage over their parent works because they are dictionaries with a focus. Bilingualized dictionaries (BDs), in this sense, combine the strengths of two types of dictionaries – the authenticity and reliability of MELDs and the accessibility of native-language equivalents / translations of traditional bilingual dictionaries. However, the BDs are bound to lag behind the MELDs in terms of the currency of the contents. In order to compensate for this irreversible inherent shortcoming, we should engage in making BDs linguistically and culturally more focused, something which MELDs can never emulate. Since the contents are electronically stored, reverse indexing, with proper tagging when translated, is only a click away. This index functions as a mini, self-contained Native Language-English (e.g. Chinese-English) dictionary which directs EFL learners to the idiomatic English expressions with ample illustrations in the dictionary proper. A special appendix on core culture-specific words in the learners' mother tongue could also be added. These culture-specific words are among the hardest to translate. With