# The potential of learner corpora for pedagogical lexicography

Yukio Tono
Tokyo University of Foreign Studies
y.tono@tufs.ac.jp

## 1. 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 chapter, I will summarise recent developments in pedagogical lexicography with special reference to the three areas mentioned above (language corpora, dictionary users and language learners) and argue that significant improvements in user-friendliness could be achieved by using second language (L2 henceforth) learner corpora to inform the making of learners' dictionaries.

## 2. 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 dramatically 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 t for finding out how a word behaves ever since. After Church and Ha (1989) introduced the notion of Mutual Information (a measure of salience of the association between any two words), lexicographers beca more interested in identifying statistically salient collocates. Every publis started to compile its own corpus (e.g. the Cambridge International Corp the Longman Corpus Network, among others) and even its own corpus qu system (a set of tools to help lexicographers use corpus data effectively). 1995, the revised Big 4 (COBUILD, OALD, LDOCE, and CIDE) clain that they were all so-called 'corpus-based.' Since then, using corpora dictionary-making has become standard practice, at least for the ma dictionary publishers in the UK.

Recently, a more sophisticated approach to using corpora has been propos mainly to deal with larger sets of data or more detailed grammatical relati in the text (e.g. the Sketch Engine (Kilgarriff, et al. 2004), the Shogakul Language Toolbox (Nakamura and Tono 2003) among others). Pap presented at Euralex 2004 or 2006 showed that dictionary publishers shifting their attention from using general mega-corpora to more speciali corpora and from the ordinary use of corpora to more purpose-specific us

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

Whilst lexicographers are aware of the importance of user studies, it ta time to apply research findings to the actual production of dictionaries. So notable successful applications of research findings to dictionary-mak include the provision of a 'menu' at the beginning of dictionary entries. W the menu, users can first browse through the various meanings of any gi word, which is a feature widely introduced after a study confirmed the f that the users only look at the beginning of dictionary entries (Tono 19 Dictionary makers are also taking note of research in dictionary use wh the effectiveness of newly introduced organizational devices are put to test. For example, I conducted an experiment to investigate the effects 'Signposts' in LDOCE and 'Guidewords' in CIDE and found that the te used for signposts in LDOCE were more effective in directing users to right meanings while the terms used for Guidewords were often too abst to "signpost" meanings in the dictionary (Tono 1992).

More recently, there has been a renewed interest in the role of dictiona in language learning since electronic dictionaries began growing in popula in Japan several years ago. The market is constantly increasing in size, and major manufacturers such as Casio, Seiko, Sharp, and Canon are competing 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 organised by the JACET Lexicography SIG has been very well attended, where around 50-60 paper presentations are given, in which the use of pocket e-dictionaries has become a favourite theme. 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. The 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 in the 1980s, my primary interest was in the role of dictionary use in language learning. When COBUILD was published in 1987, I realised the potential of corpus-based approach in language studies 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. At that time, I had an opportunity to collect English essays written by Japanese learners of English as part of a large-scale research project on 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. This corpus, the Japanese EFL Learner (JEFLL hereafter) Corpus, has been one of the primary sources of learner corpus research in Japan at present.

## 3. 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. In other cases, learners will underuse or overuse particular linguistic items or constructions, which is again the sign of on-going process of learning the target language.

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 investigate learner language by employing techniques of large-scale textual analysis by computer. What sort 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.

### 3.1. Source of vocabulary selection for L2 learners

One way of analysing learner corpora is to analyse the vocabulary used in the learner production data and compare it with the vocabulary used by native speakers. I worked for NHK (Nihon Hoso Kyokai, Japan Broadcasting Centre) in developing a television English conversation programme (titled *Hyakugo de sutato eikaiwa* which means 'Let's start with 100 basic words in English'). This programme was unique in the sense that it was the first 'corpus-based' English conversation programme on television. It consists of a hundred lessons based on the one hundred 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[1]. 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 television programme, I focused on the most frequent 100 keywords and designed the programme in a lexical syllabus. As I worked on this programme, 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 teaching materials, including textbooks, grammar books and dictionaries.

### 3.2. 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:

(i) *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*)

(ii) *Verb forms*

e.g. You can '**have** problems doing something'. Do not use 'to do'

(*LDOCE*)

(iii) *Verb patterns*

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

(NOT He proposed me a possible solution.)

(*LDOCE*)

(iv) *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*)

(v) *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 the *Longman Active Study Dictionary (LASD) and the 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 usage. The error information should be tuned to the level of learners who would venture to use monolingual dictionaries. The analysis of the JEFLL Corpus shows that there are significant relationships between particular error patterns and stages of acquisition. For example, verb morphology errors are relatively more common for beginning- to intermediate-stage learners than for more advanced learners, while lexical-choice errors are observed more commonly in advanced learners. This is due to the fact that beginning-level learners had difficulties in constructing a longer sentence that contains complex noun phrases, thus there are a smaller number of errors in nouns in comparison to verb errors. For advanced learners, on the other hand, they produce more complex sentences, containing more noun phrases in verb object or prepositional object positions, which produce more noun-related lexical errors, rather than verb errors. In contrast to these, the article (***the***, ***a***, ***an***) errors are persistent throughout the acquisition levels, which shows that there are some errors that are very difficult to overcome for Japanese learners of English.

Secondly, it is difficult to deal with L1-related errors in general-purpose monolingual dictionaries which are not aimed at particular L1-speaking learners. For example, Japanese-speaking learners of English often made the following ungrammatical sentences (examples are taken from JEFLL):
(1) My house is Shinjuku.
(2) We are very hard life everyday.
(3) Keitai is kakaru a lot of money. ('keitai' = 'mobile phone'; 'kakaru' = 'cost')
All these sentences have the 'subject + ***be*** + predicate' pattern, whose

corrected versions are listed as follows:
(4) I live in Shinjuku.
(5) We have a very hard life every day.
(6) Mobile phones cost a lot of money.
It is interesting that each pattern involves slightly different patterns of errors. The sentence (1) needs a human subject as in 'I live in ...', whereas the original sentence starts with 'my house'. This is very common in Japanese, for in Japanese sentences do not always require a subject. The sentence in (7), for example, means 'I live in Shinjuku':

(7) Watashi-no ie-wa Shinjuku-desu.
I-GEN house-TOP Shinjuku-DECL

In (7), 'Watashi-no ie-ha' works as topicalisation, not the subject of the sentence. Many students get confused about the role of a copula ***be*** in this kind of construction. The problem is that a direct translation sometimes works, as in (8), but sometimes not, as in (9):

(8) Watashi-wa kanemochi-desu. (= I am rich.)
I- SUB rich-DECL
(9) Watashi-wa kohi. (=*I am coffee.)
I-TOP coffee (DECL-omitted)

Japanese learners of English often confuse subject-predicate constructions with topic-comment constructions, which is one of the very common error patterns in learner writings.

The second error above in (2) shows a slightly different pattern. In this case, the Japanese equivalent is also a topic-comment structure as in (10):

(10) Mainichi taihen kibishii seikatu desu.
everyday very hard life DECL

The sentence (10) could be translated using a pronoun ***it***, as in (11), or a personal pronoun ***we***, as in (12):

(11) It is a very hard life.
(12) We have a very hard life.

In the sentence (2), the student successfully chose the personal pronoun we, but could not select a right verb to express the message (12). Here there is a strong tendency that Japanese learners of English stick to the 'subject + be + predicate' construction whenever they can express the message by the topic-comment construction in L1.

The sentence (3) also shows the evidence that Japanese learners of English heavily depend on the topic-comment structure whenever they cannot find right words. Here, the student could not come up with words like "mobile phone" or "cost", so she or he used the Japanese equivalents in a sentence. To realise the proposition into English, he deliberately used a copula *be* because again the Japanese translation for the sentence (3) could be expressed using topical particle "*wa*":

(13) Keitai-wa okane ga kakaru.
mobile phone-TOP money-SUB cost

Here it is a little tricky, since, in Japanese, the noun *okane* (i.e. *money*) can become a subject for the verb *kakaru* (i.e. ***cost***) whereas in English it should be placed in the object position. Thus, this student gets confused and simply connect "mobile phone" and "cost" with a copula ***be***.

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, these kinds of L1-related errors should be.treated more extensively in bilingual learners' dictionaries. Unless a careful examination of learner production data, it would not be possible to incorporate such information in learner's dictionaries. Learner corpora should play an important role here as well.

## 3.3. Identifying the weak areas of learners: underuse of collocations

It is not sufficient to use learner corpora to provide error information only. Another significant application of corpus-based techniques would be to show L2 learners the gap in performance between native speakers and learners and thus encourage them to perform in a more target-like manner. One typical example would be the pattern of use in grammatical and lexical collocations. Table 1 shows the comparison in object-noun collocates of the verb ***make*** between the British National Corpus and the JEFLL Corpus.

**Table 1** Object-noun collocates of the verb make in BNC and JEFLL

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

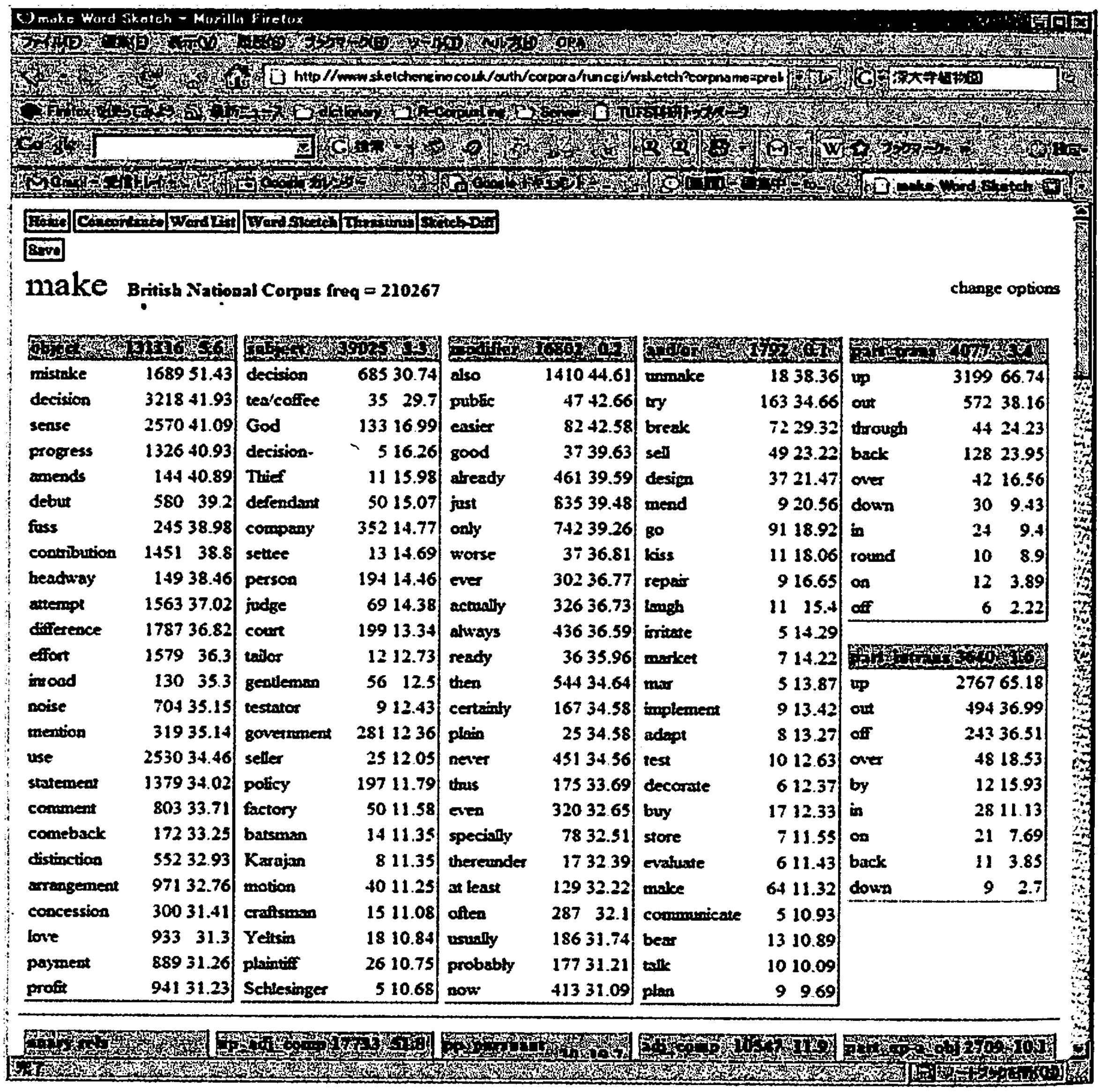
Japanese EFL learners tend to use, as collocates of *make*, relatively concrete objects such as *money*, *food*, *friends*, among others. These collocates are considered to be free combinations with the verb *make* in the sense of 'to produce'. On the other hand, native speakers use the verb *make* with more abstract nouns such as ***sense***, *way*, *use*, *decision* and the like. 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 this way, we can take into account the gap between native speakers and L2 learners (Tono 2001:203ff).

## 4. Profiling learner language for pedagogical lexicography

So far I have discussed the potential of learner corpus research for pedagogical lexicography in three major areas. The contribution of learner corpora sorely depends on the quality of the data. I have recently completed the projects of compiling two different corpora of Japanese EFL learners. One is called the NICT JLE Corpus (Izumi et al. 2004)[2], the world-largest spoken learner corpus to date. The total size is approximately 2 million words. It is comprised of 1,281 subjects' transcripts of 15-minute oral proficiency interviews taken as part of the Standard Speaking Test, the localised version of the ACTFL Oral Proficiency Interview, developed by ALC Press. Each subject's transcript is tagged for nine spoken English proficiency levels, which makes it possible to investigate characteristics of learner language based on proficiency. The second corpus is called the JEFLL Corpus (c. 700,000 words; see Tono 2007 for more details), which is a corpus of free compositions (in-class, timed essays written without recourse to dictionaries) by approximately 10,000 students.

By comparing spoken and written productions by Japanese-speaking learners of English in terms of overuse, underuse, and misuse in lexis and structures, we could provide systematic usage notes in learner's dictionaries. They are not just common errors, but errors characterising particular proficiency groups of learners in spoken and written modes, which would be very different from the common learner error notes provided in LDOCE, CALD and so forth.

Another breakthrough has come when lexical profiling tools such as the Sketch Engine (See Figure 1) and the Shogakukan Corpus Network (See Figure 2) became available these several years. Both interfaces can deal with not only learner corpora[3] but also native-speaker corpora such as BNC, making it possible to compare native vs. non-native performance in various linguistic features. The Sketch Engine is especially useful to analyse the grammatical relations of the target word, such as verb and prepositional complements, adjectival and adverbial modifiers, etc. It is now possible to gain an overall picture of learners' use of core and specialised vocabularies at various proficiency levels. This will provide very useful input for vocabulary learning theories and for syllabus design.


make Word Sketch - Mozilla Firefox
http://www.sketchengine.co.uk/auth/corpora/run.cgi/wsketch?corpname=preł
Home
Concordance
Word List
Word Sketch
Thesaurus
Sketch-Diff
Save
change options
make Word Sketch

make British National Corpus freq = 210267

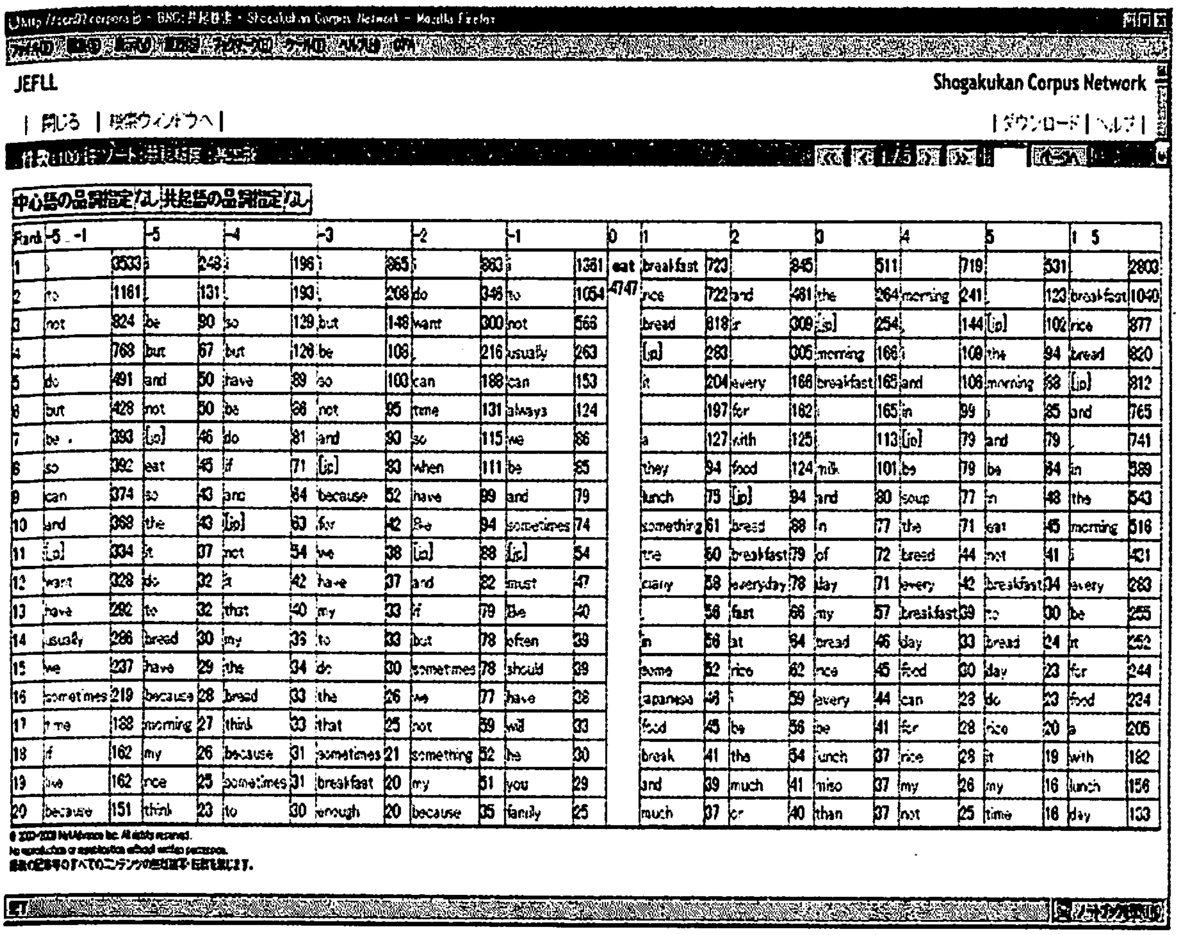
| object | 131316 | 5.6 | subject | 39025 | 1.3 | modifier | 16802 | 0.2 | and/or | 1792 | 0.1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| mistake | 1689 | 51.43 | decision | 685 | 30.74 | also | 1410 | 44.61 | unmake | 18 | 38.36 |
| decision | 3218 | 41.93 | tea/coffee | 35 | 29.7 | public | 47 | 42.66 | try | 163 | 34.66 |
| sense | 2570 | 41.09 | God | 133 | 16.99 | easier | 82 | 42.58 | break | 72 | 29.32 |
| progress | 1326 | 40.93 | decision- | 5 | 16.26 | good | 37 | 39.63 | sell | 49 | 23.22 |
| amends | 144 | 40.89 | Thief | 11 | 15.98 | already | 461 | 39.59 | design | 37 | 21.47 |
| debut | 580 | 39.2 | defendant | 50 | 15.07 | just | 835 | 39.48 | mend | 9 | 20.56 |
| fuss | 245 | 38.98 | company | 352 | 14.77 | only | 742 | 39.26 | go | 91 | 18.92 |
| contribution | 1451 | 38.8 | settee | 13 | 14.69 | worse | 37 | 36.81 | kiss | 11 | 18.06 |
| headway | 149 | 38.46 | person | 194 | 14.46 | ever | 302 | 36.77 | repair | 9 | 16.65 |
| attempt | 1563 | 37.02 | judge | 69 | 14.38 | actually | 326 | 36.73 | laugh | 11 | 15.4 |
| difference | 1787 | 36.82 | court | 199 | 13.34 | always | 436 | 36.59 | irritate | 5 | 14.29 |
| effort | 1579 | 36.3 | tailor | 12 | 12.73 | ready | 36 | 35.96 | market | 7 | 14.22 |
| inroad | 130 | 35.3 | gentleman | 56 | 12.5 | then | 544 | 34.64 | mar | 5 | 13.87 |
| noise | 704 | 35.15 | testator | 9 | 12.43 | certainly | 167 | 34.58 | implement | 9 | 13.42 |
| mention | 319 | 35.14 | government | 281 | 12.36 | plain | 25 | 34.58 | adapt | 8 | 13.27 |
| use | 2530 | 34.46 | seller | 25 | 12.05 | never | 451 | 34.56 | test | 10 | 12.63 |
| statement | 1379 | 34.02 | policy | 197 | 11.79 | thus | 175 | 33.69 | decorate | 6 | 12.37 |
| comment | 803 | 33.71 | factory | 50 | 11.58 | even | 320 | 32.65 | buy | 17 | 12.33 |
| comeback | 172 | 33.25 | batsman | 14 | 11.35 | specially | 78 | 32.51 | store | 7 | 11.55 |
| distinction | 552 | 32.93 | Karajan | 8 | 11.35 | thereunder | 17 | 32.39 | evaluate | 6 | 11.43 |
| arrangement | 971 | 32.76 | motion | 40 | 11.25 | at least | 129 | 32.22 | make | 64 | 11.32 |
| concession | 300 | 31.41 | craftsman | 15 | 11.08 | often | 287 | 32.1 | communicate | 5 | 10.93 |
| love | 933 | 31.3 | Yeltsin | 18 | 10.84 | usually | 186 | 31.74 | bear | 13 | 10.89 |
| payment | 889 | 31.26 | plaintiff | 26 | 10.75 | probably | 177 | 31.21 | talk | 10 | 10.09 |
| profit | 941 | 31.23 | Schlesinger | 5 | 10.68 | now | 413 | 31.09 | plan | 9 | 9.69 |

| part_trans | 4077 | 3.4 |
|---|---|---|
| up | 3199 | 66.74 |
| out | 572 | 38.16 |
| through | 44 | 24.23 |
| back | 128 | 23.95 |
| over | 42 | 16.56 |
| down | 30 | 9.43 |
| in | 24 | 9.4 |
| round | 10 | 8.9 |
| on | 12 | 3.89 |
| off | 6 | 2.22 |

| part_intrans | 3640 | 1.6 |
|---|---|---|
| up | 2767 | 65.18 |
| out | 494 | 36.99 |
| off | 243 | 36.51 |
| over | 48 | 18.53 |
| by | 12 | 15.93 |
| in | 28 | 11.13 |
| on | 21 | 7.69 |
| back | 11 | 3.85 |
| down | 9 | 2.7 |

unary_rels | np_adj_comp 17733 51.8 | pp_pursuant [illegible] 10.7 | adj_comp 10547 11.9 | part_up_n_obj 2709 10.1

**Figure 1** The Sketch Engine (word sketch on BNC)



JEFLL

Shogakukan Corpus Network

| 閉じる | 検索ウィンドウへ |

| ダウンロード | ヘルプ |

1/5

中心語の品詞指定なし 共起語の品詞指定なし

| Rank | -5 _ -1 | | -5 | | -4 | | -3 | | -2 | | -1 | | 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | I | 3533 | I | 248 | I | 196 | I | 365 | I | 383 | I | 1361 | eat<br>4747 |
| 2 | to | 1161 | , | 131 | , | 193 | , | 208 | do | 346 | to | 1054 | |
| 3 | not | 824 | be | 90 | so | 129 | but | 148 | want | 300 | not | 566 | |
| 4 | | 768 | but | 67 | but | 126 | be | 108 | , | 216 | usually | 263 | |
| 5 | do | 491 | and | 50 | have | 89 | so | 100 | can | 188 | can | 153 | |
| 6 | but | 428 | not | 50 | be | 38 | not | 95 | time | 131 | always | 124 | |
| 7 | be . | 393 | [jp] | 46 | do | 81 | and | 90 | so | 115 | we | 86 | |
| 8 | so | 392 | eat | 45 | if | 71 | [jp] | 90 | when | 111 | be | 65 | |
| 9 | can | 374 | so | 43 | and | 64 | because | 52 | have | 99 | and | 79 | |
| 10 | and | 368 | the | 43 | [jp] | 63 | for | 42 | the | 94 | sometimes | 74 | |
| 11 | [jp] | 334 | it | 37 | not | 54 | we | 38 | [jp] | 88 | [jp] | 54 | |
| 12 | want | 328 | do | 32 | it | 42 | have | 37 | and | 82 | must | 47 | |
| 13 | have | 292 | to | 32 | that | 40 | my | 33 | if | 79 | Be | 40 | |
| 14 | usually | 286 | bread | 30 | my | 36 | to | 33 | but | 78 | often | 39 | |
| 15 | we | 237 | have | 29 | the | 34 | do | 30 | sometimes | 78 | should | 39 | |
| 16 | sometimes | 219 | because | 28 | bread | 33 | the | 26 | we | 77 | have | 36 | |
| 17 | time | 188 | morning | 27 | think | 33 | that | 25 | not | 59 | will | 33 | |
| 18 | if | 162 | my | 26 | because | 31 | sometimes | 21 | something | 52 | he | 30 | |
| 19 | we | 162 | rice | 25 | sometimes | 31 | breakfast | 20 | my | 51 | you | 29 | |
| 20 | because | 151 | think | 23 | to | 30 | enough | 20 | because | 35 | family | 25 | |

| Rank | 1 | | 2 | | 3 | | 4 | | 5 | | 1 5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | breakfast | 723 | . | 845 | . | 511 | . | 719 | . | 531 | . | 2803 |
| 2 | rice | 722 | and | 481 | the | 264 | morning | 241 | , | 123 | breakfast | 1040 |
| 3 | bread | 818 | , | 309 | [jp] | 254 | , | 144 | [jp] | 102 | rice | 877 |
| 4 | [jp] | 283 | . | 305 | morning | 168 | . | 108 | the | 94 | bread | 820 |
| 5 | it | 204 | every | 188 | breakfast | 165 | and | 106 | morning | 68 | [jp] | 812 |
| 6 | | 197 | for | 162 | . | 165 | in | 99 | . | 85 | and | 765 |
| 7 | a | 127 | with | 125 | . | 113 | [jp] | 79 | and | 79 | . | 741 |
| 8 | they | 94 | food | 124 | milk | 101 | be | 79 | be | 84 | in | 589 |
| 9 | lunch | 75 | [jp] | 94 | and | 80 | soup | 77 | in | 48 | the | 543 |
| 10 | something | 61 | bread | 88 | in | 77 | the | 71 | eat | 45 | morning | 516 |
| 11 | the | 60 | breakfast | 79 | of | 72 | bread | 44 | not | 41 | . | 431 |
| 12 | daily | 58 | everyday | 78 | day | 71 | every | 42 | breakfast | 34 | every | 283 |
| 13 | . | 56 | fast | 66 | my | 57 | breakfast | 39 | to | 30 | be | 255 |
| 14 | in | 56 | at | 54 | bread | 46 | day | 33 | bread | 24 | it | 252 |
| 15 | some | 52 | rice | 62 | rice | 45 | food | 30 | day | 23 | for | 244 |
| 16 | japanese | 48 | . | 59 | every | 44 | can | 23 | do | 23 | food | 234 |
| 17 | food | 45 | be | 56 | be | 41 | for | 28 | rice | 20 | a | 206 |
| 18 | break | 41 | the | 54 | lunch | 37 | rice | 28 | it | 19 | with | 182 |
| 19 | and | 39 | much | 41 | miso | 37 | my | 26 | my | 16 | lunch | 156 |
| 20 | much | 37 | or | 40 | than | 37 | not | 25 | time | 16 | day | 133 |

© 2003-2008 NetAdvance Inc. All rights reserved.
No reproduction or republication without written permission.
掲載の記事等のすべてのコンテンツの無断転載・配信を禁じます。

**Figure 2** The Shogakukan Corpus Network (JEFLL Corpus on the Web)

## 5. Conclusion

In this chapter, I have argued that a computational analysis of a large amount of L2 learner language will shed light on better understanding of language learning processes and incorporation of such information into pedagogical lexicography will improve learner's dictionaries tremendously.

If developmental errors are identified for each proficiency level, dictionaries can then be customised 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 customisable dictionary described above should become a reality in the foreseeable future. Pedagogical dictionaries should deal with all the issues I have discussed in this chapter and provide necessary support. 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.

## Notes

1. The statistics is based on the spoken component of the British National Corpus.
2. The NICT JLE Corpus was completed by the National Institute of Information & Communications Technology, but I was heavily involved in planning the corpus building project initially and asked for help from NICT.
3. The Sketch Engine does not contain learner corpora at default setting, but we can upload our own data for analysis.