Developing the Optimal Learning List of Irregular Verbs Based on the Native and Learner Corpora

Yukio TONO
Megumi AOKI

Introduction

There is a growing awareness that corpus data can contribute to the advancement of knowledge in the study of language. There are many fields of language study in which corpus data is recognised as a useful tool for empirical linguistic research (See, for example, McEnery & Wilson 1996: 87ff). The field of language teaching is no exception, where the most notable application of corpus data can be seen in the field of pedagogical lexicography. Since the use of large L1[1] corpora for the compilation of the first edition of the *COBUILD English Dictionary* (1987), such corpora have constantly provided pedagogical lexicographers with naturally occurring examples. It has also provided quantitative guidance on lexical choice.

Besides the area of lexicography, a number of scholars have used corpus data to critically examine existing language teaching materials and syllabuses (for the application of general corpora in ESL textbook analysis, see Kennedy 1987, Holmes 1988, Ljung 1990, 1991, Mindt 1995, Dodd 1997; for examples in improving syllabus design, see Mindt 1996a; for application in CALL, see Johns 1996 : for potentials in language assessment, see Alderson 1996).

While the data provided from general L1 corpora are useful for establishing the target language norms for learners, we must approach L1 corpora with caution. The content of language teaching/ learning is to be determined not only by the information on the native speaker's use of the language, but also by such elements as the L1 background, attitudes and goals of particular learners, the environment which language learning takes place and what we believe needs to be learned in order to be an effective language user. Some grammatical items might be easier to acquire for certain groups of learners because of their L1 background. There may be a case in which the most frequent word in English, such as the determiner *the*, does not necessarily mean it is most easily learned (Tono 1998). Some grammatical items may be governed by the developmental stages of acquisition and others may be free (for instance, see Pienemann 1992).

The recent development of learner corpora is a good indicator of this concern. The comparison of native corpora with learner corpora throws more light on learner behaviour such as overuse/underuse or misuse/errors of some language features, the influence of learner's L1 knowledge, or teaching order of particular grammatical items, and the like. This paper deals with the potential of comparing native speaker corpora with learner corpora in order to provide optimal learning list of particular grammatical items.

A Learning List of Irregular Verbs in English

One of the good examples of corpus-based studies for making a guide for the selection and gradation of particular grammatical items is a series of works carried out by Dieter Mindt and his colleagues (Mindt 1987, 1995, 1996a, 1996b; Grabowski and Mindt 1995). He claims that "a comparative study of authentic language data and textbooks for teaching English as a foreign language has revealed that the use of grammatical structures in textbooks differs considerably from the use of these structures in authentic English" (Mindt 1996: 232). A practical example is a learning list of irregular verbs in English by Grabowski and Mindt (1995). They analysed the use of 160 irregular verb forms in the Brown and LOB corpora and produced a list ranked in order of frequency for the benefit of EFL learners. After learning the forms of the first ten verbs of the list (*say, make, go, take, come, see, know, get, give, find*) the learner has mastered more than 45 % of all forms of irregular verbs in English (Mindt 1996b: 49).

As we looked at this list, we wondered whether this list could be used as it is for EFL learners in Japan. This motivated us to conduct a comparative study of this list and the other two lists: one created by EFL textbook corpora and the other based on our learner corpora.

Purpose

The purpose of this study is twofold: firstly, we try to investigate the frequency patterns of irregular verbs in native and learner corpora to check the validity of the learning list developed by Grabowski and Mindt (1995). Secondly, we would like to see if the data of learner corpora, together with English textbook corpora, can shed more light on the making of an optimal learning list of irregular verbs. As has already been mentioned, it does not seem to be sufficient to provide the learning list of grammatical items based solely upon native corpora data. We hope to clarify the difference between the two types of corpora and make a suggestion for the possible improvement of learning lists.

The Corpora and Tools Used for the Study

For the frequency analysis of irregular verbs, two types of corpora were used to make comparison with the results by Grabowski and Mindt (1995), which were obtained from the Brown and LOB corpora. One is called JEEFL (Japanese EFL Learner) Corpus, which consists of spoken and written corpora produced by Japanese EFL learners aged thirteen to nineteen (the first year students in junior high school to university freshmen). In this study, the subcorpora composed of written essay data were used (each subcorpus size: Junior High = 76,945 words ; Senior High = 74,627 words) . Comparison was also made with frequency data obtained from the English Textbook Corpus, which consists of seven English textbooks officially authorized and used in 3-year course at public junior high schools (c. 61,000 words) and five textbooks used in 3-year course at senior high schools (c. 69,000 words; senior high school is not compulsory in Japan). There are many different courses in senior high school level, such as English I, II (general course), Reading, Writing, Oral Communication, etc., but we chose the textbooks used for English I only.

The frequency analysis was done by the query tool called TXTANA (by Shiro Akasegawa, URL: http://www.biwa.or.jp/~aka-san/index.html in Japanese). This concordancer has a function called Synonym Dictionary. This is basically a list of lemmatised words with their possible variants in the text. The Synonym Dictionary allows all the possible occurrences of a lemma to be searched for without actually lemmatising the text.

Results and Discussion

Table 1 shows the results of Spearman rank-order correlation among the six different lists obtained from the corpora. NATIVE denotes the original list made by Grabowski and Mindt (1995). JHTEXT is the list produced from the junior high school textbook corpus. SHTEXT is based on the senior high school textbook corpus. The other three lists (JALL, SALL and ALL) are from written essay data of JEFLL corpus[2]. As you can see in Table 1, there is a significant positive correlation between the lists, even though the lists were derived from the corpora of English learners in Japan and the English textbooks. It is quite natural to have a high correlation if the entire lists of irregular verbs are compared.

We found that in both learner and textbook corpora, the verbs below the 100th rank were rarely used[3]. Many of the verbs between rank 50 and 100 are not very frequent, either. There are some exceptions, such as a group of verbs in the lower rank which was used quite frequently by learners and appeared in the textbooks, too[4]. We will discuss the possibility of revising the list later. In the meantime, let us limit our scope to the top 30 and look at the order closely. Table 2 shows the summary of correlations among the top 30 lists. We could observe that there is still a very positive correlation between the lists in every case.

Table 1: Spearman's rank order correlation of the irregular verb lists (n=160)

			NATIVE	JHTXT	SHTXT	JALL	SALL	ALL
Spearman's rho	Correlation Coefficient	NATIVE	1.000	.847*	.912*	.831*	.835*	.846*
		JHTXT	.847*	1.000	.879*	.877*	.864*	.873*
		SHTXT	.912*	.879*	1.000	.853*	.854*	.864*
		JALL	.831*	.877*	.853*	1.000	.927*	.961*
		SALL	.835*	.864*	.854*	.927*	1.000	.980*
		ALL	.846*	.873*	.864*	.961*	.980*	1.000

**. Correlation is significant at the .01 level (2-tailed).

Table 2: Spearman's rank-order correlation among the frequency lists (top 30)

			NATIVE	JHTXT	SHTXT	JALL	SALL	ALL
Spearman's rho	Correlation Coefficient	NATIVE	1.000	.814*	.898*	.774*	.780*	.783*
		JHTXT	.814*	1.000	.873*	.699*	.649*	.664*
		SHTXT	.898*	.873*	1.000	.784*	.745*	.766*
		JALL	.774*	.699*	.784*	1.000	.970*	.991*
		SALL	.780*	.649*	.745*	.970*	1.000	.987*
		ALL	.783*	.664*	.766*	.991*	.987*	1.000

**. Correlation is significant at the .01 level (2-tailed).

As we narrow down the number of verbs in the list, however, there is a distinct tendency for the correlation to become less significant. The rank order is more sensitive to the types of corpora. Table 3 and 4 shows the correlations in top 20 and 10 verbs in each list.

Table 3: Spearman's rank order correlation of the irregular verb lists (n=20)

			NATIVE	JHTXT	SHTXT	JALL	SALL	ALL
Spearman's rho	Correlation Coefficient	NATIVE	1.000	.889*	.832*	.556*	.573*	.554*
		JHTXT	.889*	1.000	.884*	.555*	.513*	.512*
		SHTXT	.832*	.884*	1.000	.607*	.572*	.561*
		JALL	.556*	.555*	.607*	1.000	.968*	.983*
		SALL	.573*	.513*	.572*	.968*	1.000	.991*
		ALL	.554*	.512*	.561*	.983*	.991*	1.000

**. Correlation is significant at the .01 level (2-tailed).

*. Correlation is significant at the .05 level (2-tailed).

Table 4: Spearman's rank order correlation of the irregular verb lists (n=10)

			NATIVE	JHTXT	SHTXT	JALL	SALL	ALL
Spearman's rho	Correlation Coefficient	NATIVE	1.000	.602	.818**	.491	.576	.515
		JHTXT	.602	1.000	.851**	.377	.267	.249
		SHTXT	.818**	.851**	1.000	.418	.309	.285
		JALL	.491	.377	.418	1.000	.903**	.939**
		SALL	.576	.267	.309	.903**	1.000	.988**
		ALL	.515	.249	.285	.939**	.988**	1.000

**. Correlation is significant at the .01 level (2-tailed).

Table 4 shows that the rank order of irregular verbs in learner corpora have no positive correlation with those of native corpora and English textbooks. It is noteworthy that the correlation becomes no more significant between the native corpora list and the junior high textbook corpora whereas the correlation remains significantly positive between the lists in native corpora and senior high textbook corpora. This seems to indicate that there are particular types of irregular verbs which are used relatively more frequently in learner's writing as well as English textbooks for novice learners, in contrast to the frequency list made by the L1 corpora. If we can find out why these particular verbs are used more often than in the standard list, based on the use of native speakers, we could possibly account for the significant mismatch between the naturally occurring frequencies of the verbs in L1 corpora and those occurring in L2 or L2 oriented data. There is a possibility, therefore, that the learning list could be revised and improved based upon the information from learner corpora so that the list may be better tuned to the needs of English learners in a particular learning context.

Let us now turn to the detailed analysis of a group of verbs which contributed to changing the rank order. Table 5 describes the frequency lists obtained from each corpus. There seem to be two main reasons for the low correlation in the top 10 verb lists. Firstly, the first two most frequent verbs in the native corpora, *say* and *make*, were used less frequently in learner data. Since the essays in learner corpora do not contain any narrative essay, the learners did not have a chance to use the verb *say* more frequently. Secondly, there are a group of verbs which do not appear in the native corpora or English textbook corpora, but appeared in learner corpora. Many of these verbs are characterised by the daily routines (*wake, sleep, eat, buy*) while others are particularly influenced by the essay topic (such as *run* in the topic "What do you take out with you when a big earthquake happens?") We can argue that these verbs appeared simply because one of the essay tasks was a description of daily routines such as "Which do you prefer for breakfast, rice or bread?" However, we should not ignore the fact that some of these verbs are strongly influenced by the language activities in the classroom and also the textbooks as primary input source.

Table 5: The frequency lists of top 20 in each corpus

native	all	jall	sall	jhtxt	shtxt
say	eat*	eat*	eat*	go	say
make	take	go	take	come	go
go	go	think	go	say	see
take	think	take	think	know	come
come	buy*	buy*	get	see	think
see	get	get	buy*	make	make
know	bring	bring	bring	get	know
get	become	become	make	think	take
give	make	see	run*	write*	get

find	come	make	become	take	feel
think	see	come	come	let*	tell
tell	run*	say	say	speak*	give
become	say	run*	see	become	become
show	give	know	feel	give	begin
leave	feel	feel	give	eat*	leave
feel	know	give	wake*	tell	write*
put	wake*	wake*	know	fly	find
bring	sleep*	sleep*	find	meet*	put
begin	find	find	sleep*	show	run*

Table 6 indicates the list of irregular verbs whose ranks were radically different from the native speaker's list. Those verbs which do not appear in the rank difference lists of the textbook corpus (for example, *wake, sleep, drink, forget, win, break, spend*) were presumably influenced by the essay topics whereas other verbs in common with the learner and textbook lists (i.e. *eat, swim, buy, fly*) were the potential words which were encouraged to use through classroom instruction using these textbooks.

Table 6: Rank differences in the textbook corpora as compared with the native corpora

all [illegible]	rank [illegible]	[illegible]
wake	86 17	69
eat	62 1	61
sleep	75 18	57
drink	70 20	50
swim	84 39	45
buy	48 5	43
fly	73 40	33
forget	56 27	29
win	55 29	26
break	43 22	21
spend	44 23	21

Table 7: Rank differences in the textbook corpora as compared with the native corpora

jhtext [illegible]	rank [illegible]	[illegible]
fly	73 17	56
swim	84 33	51
eat	62 15	37
teach	61 24	37
speak	32 12	20
send	39 27	12

shtext [illegible]	rank [illegible]	[illegible]
fly	73 22	51
eat	62 31	31
buy	48 30	18
speak	32 21	11
read	35 24	11

It can be argued that statistics such as those on verb-form frequency reflect their particular textual sources, and that corpora of learner English or textbooks would certainly lead to different relative frequencies. However, the broad picture from the present study is clear. Most English verb forms do not seem to be frequent enough to warrant pedagogical attention until quite advanced stages of the second language acquisition process (Kennedy 1998:284). In addition, for pedagogy, corpus-based descriptions like the one by Grabowski and Mindt (1995) should be supplemented by the communicative needs of the real learners and the particular EFL context in that country. The learner and textbook corpora together could enhance the value of the native corpus-based learning list. Further studies will be needed to determine the optimal learning list of irregular verbs due to the lack of representativeness of learner data, but we hope that this study will open the door to the possibility of developing learner resources based on a healthy fusion of L1 and L2 data.

REFERENCES

Alderson, C. (1996) Do corpora have a role in language assessment? In Thomas, J. and Short, M. (eds.) *Using Corpora for Language Research*. London: Longman: 248-259.

Dodd, B. (1997) Exploiting a corpus of written German for advanced language learning. In Wichmann, A., Fligelstone, S., McEnery, T., and Knowles, G. (eds.) *Teaching and Language Corpora*. London: Longman: 131-145.

Grabowski, E. and Mindt, D. (1995) A corpus-based learning list of irregular verbs in English. *ICAME Journal* 19: 5-22.

Holmes, J. (1988) Doubt and certainty in ESL textbooks. *Applied Linguistics*, 9:21-44.

Johns, T. (1996) Contexts: the background, development and trialling of a concordance-based CALL program. In Wichmann, A., Fligelstone, S., McEnery, T., and Knowles, G. (eds.) *Teaching and Language Corpora*. London: Longman: 100-115.

Kennedy (1987) Quantification and the use of English: a case study of one aspect of the learner's task. *Applied Linguistics*, 8:264-286.

Ljung, M. (1990) *A Study of TEFL Vocabulary*. Stockholm Studies in English 78. Stockholm: Almqvist & Wiksell.

Ljung, M. (1991) Swedish TEFL meets reality. In Johansson, S. & Stenström, A.-B. (eds.) *English Computer Corpora*. Berlin: Mouton de Gruyter: 245-256.

McEnery, T. and Wilson, A. (1996) *Corpus Linguistics*. Edinburgh University Press.

Mindt, D. (1987) *Sprache – Grammatik – Unterrichtsgrammatik: Futurischer Zeitbezug im Englishcen I + Darstellungen*. Diesterweg, Frankfurt am Main.

Mindt, D. (1995) *An Empirical Grammar of the English Verb: Modal Verbs*. Berlin: Cornelsen.

Mindt, D. (1996a) English corpus linguistics and the foreign language teaching syllabus. In Thomas, J. and Short, M. (eds.) *Using Corpora for Language Research*. London: Longman: 232-247.

Mindt, D. (1996b) Corpora and the teaching of English in Germany. In Wichmann, A., Fligelstone, S., McEnery, T., and Knowles, G. (eds.) *Teaching and Language Corpora*. London: Longman: 40 – 50.

Pienemann, M. (1992) COALA – a computational system for interlanguage analysis. *Second Language Research* 8, (1): 59-92.

Tono, Y. (1998) A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. In *TALC (Teaching and Language Corpora) 98 – Conference Proceedings*, Keble College Oxford, 24-27 July 1998: 183-187.

[1] L1 = first language

[2] JALL = junior high; SALL = senior high; ALL = JALL + SALL

[3] The following verbs did not appear at all in either learner corpora or textbook corpora: *sweep, undertake, withdraw, slide, swear, thrust, split, wind, undergo, spin, creep, fling, grind, weave, dwell, shrink, bleed, sew, sow, tread, wet, saw, string, rid, strew, slit, shear, wring, forsake, shoe, beseech, smite, rend, hew, slay, slink*

[4] For example, the verbs which fall into this category are as follows: *win, forget, eat, drink, fly, sleep, wake, swim, buy, spend, break, fall, lose*