Using learner corpora to teach authentic English

Junko Shirato
Graduate School of International Media and Communication, Hokkaido University

Reference data:

Shirato, J. (2006). Using learner corpora to teach authentic English.

In K. Bradford-Watts, C. Ikeguchi, & M. Swanson (Eds.) *JALT2005 Conference Proceedings*. Tokyo: JALT.

Insights from corpus linguistics have come to be seen as having a significant impact in second language pedagogy. Recently, learner corpora, or collections of texts spoken or written by non-native speakers (NNS) of a language, are now being used for the purposes of enhancing language teaching. Specifically, by comparing vocabulary in learner corpora with that of native speakers (NS) it is possible to investigate how far, and in what ways the vocabulary of NNS deviate from NS' norms.

The present research describes how the author compared a spoken learner corpus of Japanese NNS with an established NS corpus. Results revealed that the spoken vocabulary of Japanese NNS differed markedly from that of NS in many areas. It was suggested that learner corpora have important pedagogical implications which include giving higher priority to certain classes of words and lexical phrases that appear to be underused among NNS.

コーパス言語学の発展は第二言語教育に大きな影響を与えてきた。特に近年、非母語話者の話し言葉や書き言葉を集めた学習者コーパスが語学教育の分野で応用され始めている。学習者コーパスと母語話者コーパスにおける使用語彙を比較分析する事により、学習者の使用語彙と母語話者が日常的に用いている語彙の規範との間にはどのような相違があるのかを明らかにすることが可能である。本研究では、日本人英語話者の話し言葉学習者コーパスと英語母語話者話し言葉コーパスの比較分析を行い、その結果として、両者の使用語彙には際立った相違があることを実証した。特に母語話者に比べて日本人学習者の使用頻度が著しく低い単語や連語の中には、今後の語彙指導に取り入れるべき要素が存在していることが明らかになった。この結果を踏まえて、学習者コーパスを用いた研究成果を今後、どのように語彙指導へ応用していくべきかに関して考察を行う。

Acquisition of native-like proficiency in oral English skill has always been the ultimate goal for EFL learners. To attain this competency, it is imperative to have as a precondition a complete mastery of necessary vocabulary together with lexical phrases peculiar to day-to-day interaction. Remarkable progress made by corpus linguistics and related technological applications over the past few decades has

clarified the specific features of conventional vocabulary. NS corpora provide us with important information as to the frequencies of specific linguistic features and their distributions across registers. Information of this kind alone, however, is insufficient for effective second language teaching. In order to compensate for this, spoken learner corpora have recently been created and are extensively utilized as important resources to access information about the degree of difficulty inherent in a NNS vocabulary. In comparisons of learner and NS corpora, most studies look for significant differences between NS and NNS language use in single word and multiword cluster counts. In essence, learner corpora offer potential insights for language pedagogy.

The present study aims at ascertaining what vocabulary NNS under- or overuse compared to NS norms in informal face-to-face interaction by comparing a spoken learner corpus of Japanese NNS and an established NS spoken corpus. The pedagogical significance of the findings will also be discussed.

Spoken features

The literature discussing corpora shows that various features of spoken language can be isolated based on whether they appear as single words or multiword clusters.

Single words

Single words form a substantial part of the lexicon of English and have been perceived in pedagogy as the central units to be acquired. Some attempts have been made to categorize spoken single words in recent corpus-based analysis. For example, in an analysis of 3-million samples of the CANCODE, McCarthy (1999) generated nine broad categories within a 2000-word basic vocabulary for spoken communication:

1. *modal item*--referring to degree of certainty or necessity including modal verbs and other high frequency items that carry related meanings
2. *delexical verbs*--such as *do, make, take,* and *get* in their collocations with nouns, prepositional phrases and particles
3. *interactive words*--representing speaker attitudes and stances towards the content communicated
4. *discourse markers*--whose function is to organize talk and monitor its progress
5. *basic nouns*--which have very general, non-concrete and concrete meanings
6. *general deictic items*--including demonstratives that relate the speaker to the world in relative terms of time and space
7. *basic adjectives*--for communicating everyday positive and negative evaluations of people, situations, events and things
8. *basic adverbs*--referring to time, frequency and habituality, and manner and degree
9. *basic verbs*--for actions and events denoting everyday activity
10. *Multi-word Clusters*

Phraseology, which is basically defined as the study of word combinations, has attracted considerable attention over

the past two decades. The notion of multi-word clusters has been studied under a number of rubrics by researchers, for example, *lexical phrases* (Nattinger & DeCarrico, 1992), *lexical bundles* (Biber, et al.1999), *formulaic sequences* (Wray 2000), as well as the more conventionally understood labels such as collocations, or idioms. In this study, we define *multi-word clusters* as a group of two or more words that repeatedly appear in the corpora as fixed items.

Regarding application to L2 pedagogy, McCarthy and Carter (2002) via an analysis of CANCODE argued that specific interactive functions in face-to-face communication are inherent in multi-word units. The functions they specified are as follows:

1. *discourse markers*--these clusters are used to signal a transition in a conversation, and an interactive relationship between a speaker and a listener
2. *face and politeness*--clusters encoding this function save face for a receiver and show the speaker's politeness
3. *hedges*--these convey imprecision and make statements less assertive and less open to challenge or refutation
4. *vagueness and approximation*--these refer to semantic categories in an open-ended way and help the conversation go smoothly.

These four types of word combinations appear to play a major role in spontaneous spoken interaction where speakers have to plan, encode, and actually produce their utterances in real-time. In addition they function to "allow speakers to talk without too much hesitation or without too many pauses" (Aijmer, 1996, p. 9). However, they have long been neglected in language teaching.

The comparative analysis

Although previous studies have shown how corpus-based findings on spoken vocabulary have been applied in vocabulary teaching, the present study seeks to contrast specific features of conversational vocabulary of Japanese EFL learners with those of NS. In particular, the following research questions are proposed.

Research questions

1. What are the tendencies of NNS vocabulary under- or overuse in informal face-to-face interaction as compared to NS?
2. What pedagogical implications can be drawn from these findings?

Data and method

The Japanese Learner Corpus (the JLC)

One hundred and four sample conversations, totaling 43,600 orthographic words were collected from 2003 to 2004. The informants of the JLC were 117 Japanese native speakers, 18 through 74 years of age. They can be further divided into two groups. Face-to-face in class conversations among 87 informants and a NS or NNS English teacher were recorded

by using an IC recorder, and semi-formal interviews with 30 college students containing a suggestion for creating a story and enacting a role play were recorded by a NS interviewer. As would be expected, the speaking abilities of informants differed greatly. Informants with lower speaking abilities were able to understand and hold short conversations, including asking and answering simple questions, while those with higher level abilities were proficient enough to converse about matters of daily life, such as giving simple explanations and conducting basic transactions.

A criterial corpus of NS (the BNC)

The demographically sampled portion of the British National Corpus (the BNC) was used as criterial data for the contrastive analysis. In the construction of this corpus, 153 informants were asked to record the everyday face-to-face conversations in which they participated over a two- to seven-day period. The BNC was chosen because it is currently the only spoken corpus publicly available consisting of NS face-to-face spontaneous conversations.

Procedure

All of the audio data in the JLC was transcribed into an English-based phonetic orthography. In order to extract recurring words and phrases, we generated the rank-order frequency lists of single word and two- through four-word sequences by using *Wordsmith Tools* (Scott 1999) analytical software. The next step for isolating distinctive characteristics of vocabulary in the JLC was to compare each frequency list to that in the BNC. The final step was to compile concordance lines of those words and phrases whose occurrences showed large differences from those in NS utterances.

Results

Single words

Table 1 shows the top 20 single words in both the JLC and the BNC. Frequencies are pro-rated at a rate per 100,000 words.

With regard to the nine single word categories introduced by McCarthy (1999), considerable differences appear between NNS and NS in four of these: modal items, delexical verbs, interactive words, and basic adjectives. We confine ourselves here to focusing on the specific elements in the four categories that showed marked contrasts.

a) Modal items

The total occurrences of four modal verbs, *will, can, could,* and *would* in the JLC account for 93% of the entire modal verb tokens, among which *can* ranks the highest, followed by *will* and *could*. This result roughly parallels the BNC. Furthermore, the concordance lines of these four auxiliary verbs show that NNS frequently used them to form direct polite forms, such as *could you...?* and *would you...?* However, differences between the two corpora arise in certain usages. NS use *could* to suggest doing something and to show something might be possible or happen. The concordance lines show that NS use *could* in this way very frequently, however, this usage rarely appeared in the JLC. *Should* was used by both groups to express obligation or need. There were also, however, clear divergences in the usage of *should*. NS use *should* to indicate what can

Table 1. Most frequent 20 single word lists

(Frequencies are pro-rated at a rate per 100,000 words)

	JLC	Frequency	BNC	Frequency
1	I	5,254	I	3,349
2	AND	2,391	YOU	2,958
3	YOU	2,054	THE	2,924
4	TO	1,988	AND	2,303
5	YES	1,965	IT	2,279
6	THE	1,928	A	1,960
7	IS	1,614	TO	1,865
8	MY	1,443	THAT	1,534
9	A	1,193	YEAH	1,487
10	IN	1,138	IN	1,104
11	WAS	994	OF	1074
12	DID	928	NO	1,051
13	GO	914	OH	963
14	IT	898	WELL	905
15	DID	861	HE	894
16	OF	788	IT'S	870
17	HOW	781	ON	859
18	WENT	696	WHAT	849
19	DO	685	WAS	813
20	WHAT	666	KNOW	798

reasonably be expected to happen, such as "*They should be in the car"*. This usage of *should* was quite frequent in the NS corpus, however, NNS almost never used it.

Among other modal expressions, *maybe* was the only item that frequently occurred in the JLC, while other items, such as *seem, sound, certain, definitely,* and *probably* rarely appeared. *Maybe* occurred in the JLC almost five times as frequently as in the BNC. On the contrary, occurrences of the other modal items in the BNC were 26 times as high as those in the JLC.

b) Delexical verbs

Very high-frequency verbs such as *do, make, take,* and *get*, whose meaning is determined and tailored by the company they keep (Allerton, 1984), have received considerable attention. The frequency of delexical verbs in the JLC was much lower than in the BNC. For example, the get-passives, such as *get locked in* and *get done,* which are often used to reflect the speaker's opinion on an event, seldom appeared in the JLC. Phrasal verbs, e.g., *take after* and *get over*, also rarely occurred. The result shows that NNS tended to use high frequency verbs only with their lexical meaning.

c) Interactive words

According to McCarthy (1999), interactive words such as *pretty, actually, basically,* and *really* may variously soften or make indirect potentially face-threatening utterances or intensify and emphasize affective stances towards the content of utterances. Again, limitations in the JLC were apparent. For example, while the intensifiers *pretty* appeared frequently, their usage was mostly confined to adjacency pairs such as *how are you* and *pretty good*. NS more frequently used *pretty* in collocations with other adjectives, such as *bad*, *well*, and *sure,* as well as with *good*.

Unlike the items discussed above, other interactive items, such as *actually* and *basically* rarely occurred in the JLC, while they frequently appeared in the BNC. For example, *actually* occurred in the BNC almost 40 times

more frequently than in the JLC. NS often use *actually* as a discourse particle "to change perspective and call attention to something that they have just come to think of" (Aijmer, 2002, p. 252). The underuse of *actually* in utterances may cause an abrupt change of perspective, which causes a lack of coherence in NNS' utterances.

d) Basic adjectives representing evaluations

The JLC shows that common adjectives such as *good, bad, nice* and *fine* frequently occurred, while *lovely, horrible* and *terrible*, representing more specific evaluations, rarely occurred. *Good* was used in place of a wide range of words, e.g., *wonderful, excellent, strong, delicious,* and *splendid* by NNS indicating its overuse. On the contrary, all of the adjectives mentioned above have a very high frequency in the BNC. It is important, however, to ascertain how each adjective commonly collocates with other items. *Horrible* and *terrible*, for example, are close in meaning, but the BNC shows that *terrible* usually collocates with *situation* and *state*. In contrast, no such collocation with *horrible* occurred.

Multi word clusters

Table 2 to 4 shows the most frequent 20 two- to four-words cluster lists of the JLC and the BNC. Differences between the two corpora will be discussed in terms of the following four categories: discourse markers, vagueness and approximation, face and politeness, and hedges.

Table 2. Most frequent 20 two-word cluster lists

(Frequencies are pro-rated at a rate per 100,000 words)

	JLC	Frequency	BNC	Frequency
1	DID YOU	753	YOU KNOW	688
2	I WENT	554	I DON'T	270
3	WENT TO	458	IN THE	250
4	YOU GO	339	I MEAN	225
5	DO YOU	314	I THINK	204
6	I SEE	289	DO YOU	202
7	IT WAS	266	IT WAS	194
8	I LIKE	250	ON THE	170
9	HOW WAS	238	AND I	163
10	GO TO	208	I KNOW	157
11	THANK YOU	181	I SAID	148
12	I HAVE	172	DON'T KNOW	140
13	WHAT DID	163	OF THE	135
14	WITH MY	153	AND THEN	134
15	IN THE	144	HAVE TO	131
16	I DON'T	131	I WAS	129
17	GO THERE	126	YOU CAN	128
18	NAME IS	117	IF YOU	126
19	MY HUSBAND	117	IS IT	116
20	YOU DO	115	GOT A	113

Table 3. Most frequent 20 three-word cluster lists

(Frequencies are pro-rated at a rate per 100,000 words)

	JLC	Frequency	BNC	Frequency
1	I WENT TO	398	I DON'T KNOW	117
2	DID YOU GO	254	I DON'T THINK	49
3	WHAT DID YOU	172	DO YOU WANT	48
4	HOW WAS YOUR	133	A LOT OF	44
5	MY NAME IS	112	WHAT DO YOU	32
6	YOU GO THERE	105	A BIT OF	26
7	DO YOU LIKE	103	HAVE YOU GOT	26
8	WHERE DID YOU	103	DO YOU KNOW	26
9	WHEN DID YOU	96	YOU HAVE TO	24
10	DID YOU DO	94	YOU WANT TO	24
11	WAS YOUR WEEKEND	87	YOU KNOW WHAT	23
12	I DON'T KNOW	85	I MEAN I	22
13	WHAT KIND OF	82	AND I SAID	22
14	HOW WAS THE	82	HAVE A LOOK	21
15	I WANT TO	82	I DON'T WANT	21
16	WHO DID YOU	79	YOU'RE GOT TO	21
17	DID YOU EAT	76	DON'T KNOW WHAT	21
18	HOW LONG DID	76	MM MM MM	21
19	I GO TO	69	BE ABLE TO	20
20	I HAD A	69	DO YOU THINK	20

Table 4. Most frequent 20 four-word cluster lists

(Frequencies are pro-rated at a rate per 100,000 words)

	JLC	Frequency	BNC	Frequency
1	DID YOU GO THERE	87	MM MM MM MM	16
2	HOW WAS YOUR WEEKEND	76	I DON'T KNOW WHAT	15
3	WHEN DID YOU GO	69	KNOW WHAT I MEAN	9
4	WHERE DID YOU GO	66	YOU KNOW WHAT I	9
5	WHAT DID YOU DO	62	I THOUGHT IT WAS	9
6	HOW WAS THE WEATHER	44	DO YOU WANT TO	8
7	DID YOU GO TO	39	I DON'T KNOW WHETHER	8
8	I WANT TO GO	34	WHAT DO YOU WANT	8
9	I WENT TO A	32	I DON'T WANT TO	7
10	I WENT TO THE	32	THE END OF THE	7
11	WHO DID YOU GO	32	A BIT OF A	7
12	HOW LONG DID YOU	32	HAVE A LOOK AT	7
13	WHAT DID YOU SEE	30	DO YOU KNOW WHAT	7
14	HOW LONG DID IT	27	ARE YOU GOING TO	6
15	LONG DID IT TAKE	27	DO YOU WANT A	6
16	WHO DID YOU HAVE	27	I THINK IT WAS	6
17	I READ A BOOK	27	AT THE END OF	6
18	LONG DID YOU STAY	25	A CUP OF TEA	6
19	NICE TO SEE YOU	25	WELL I DON'T KNOW	6
20	WHAT DID YOU BUY	25	I TELL YOU WHAT	6

a) Discourse markers

You know and *I mean* are the most commonly used phrases in spoken interaction among NS; the former marks a projection of the speaker's understanding of the conversational topic or shared knowledge of the situation, and the latter marks a speaker's orientation toward *own talk* (Schiffrin, 1987). Both of them rarely appeared in the JLC.

The extended clusters, *(do) you know what I mean* and *if you see what I mean,* encoding a similar function of checking shared knowledge never occurred in the JLC, while they were common in the BNC.

b) Vagueness and approximation

Among all of the strings used as interactive units, those encoding vagueness and approximation displayed the largest differences between NNS and NS. McCarthy and Carter (2002) listed seven items that are inherently used by NS; *a couple of, or something like that, that sort of thing, this that and the other,* and *all this/that sort of thing* in this category. Aijmer (2002) listed a total of 56 variants of the above strings by examining the London-Lund Corpus and demonstrated that such vagueness is ubiquitous in the native speaker's spoken data. Despite the frequency and importance of these vague expressions appearing in the BNC (38 variants), none of these variants occurred in the JLC.

c) Face and politeness

According to McCarthy (1999), "Speakers use indirect forms to perform speech acts such as directives and requests in order to protect the face of their receivers" (p. 22). He lists the following five items in this category: *do you think, do you want (me) (to), I don't know if/whether, what do you think,* and *I was going to say.* Overall, these items rarely appeared in the JLC, while all of these are very frequent in the BNC. *can you* and *can I* occurred 25 times and 14 times more frequently in the JLC, respectively. The corpus data show that NNS tended to overuse direct polite forms with modal verbs, such as the speaker-based form, *Modal + I + VP* and the hearer-based form, *Modal + you + VP*, common for less formal situations with close friends in any register.

d) Hedges

Hedging is also used as an important aspect of face-protection and politeness in spoken interaction. Among NS, *sort of* is the most frequently used discourse particle whose meaning can be expressed by adverbs, such as *approximately, roughly,* or *rather*. The data show that NS used *sort of* almost 150 times as frequently as did NNS. NNS used *sort of* preceding only noun phrases, such as *What sort of sushi do you like?* However, NS used it in a wide range of syntactic environments: it preceded not only nouns, but verbs, adverbs, adjectives, and preposition phrases in the BNC. Its flexibility is reflected in a range of different positions.

Discussion

The results of this study indicate that the NNS participants clearly deviated from NS in the frequency with which they use both single word and multiword items. Below, we will discuss the significance of the findings and suggest pedagogical implications for the Japanese context.

Single words

The relative infrequency of discourse markers in the JLC was one of the most notable differences isolated by this study. The less intuitive usage of modals, such as when *should* is used to indicate expectation, needs to be brought into the curriculum, judging from the minimal usage it appears to presently have among NNS. In the same vein, it appears that *maybe* is used as a catchall term to cover a wide range of probability for Japanese learners. The simple addition of a few terms such as *probably* and *definitely* to the lexical syllabus would greatly enhance their limited repertoire. A similar argument can be made for the adjectives and interactive words discussed in the results. A very limited expansion or earlier introduction in the syllabus of as few as a dozen words (*actually, basically, terrible*) would seem to enhance naturalness considerably at the small expense of memorizing a few new words.

Naturally, there are some caveats when making changes to established curriculums. McCarthy (1999) suggests that some justification is necessary for incorporating modals into a lexical syllabus at each level of pedagogy. However, the learnability and familiarity of auxiliary verbs are generally considered higher than those of other modal items for Japanese EFL learners because they are introduced earlier in junior and senior high school. Therefore, it is recommended that other modal items be phased in according to learner level even though their frequencies are high among NS.

The relatively infrequent use of delexical verbs in the JLC brings up an important distinction that we suggest be recognized when devising the L2 curriculum. Because the participants showed reluctance to use these terms, i.e., preferring *depart* over *take off*, learners may benefit from the realization that delexical verbs are more appropriate in spoken language, while low frequency single item verbs are more often used in writing. This knowledge coupled with the learning of the most commonly used forms as chunks could result in more natural language. To impart this realization, one strategy may be to integrate an actual explanation into the L2 syllabus to describe the differences between spoken and written discourse.

Multi-word clusters

It is worth noting how high frequency clusters compare to the distribution of single words in corpora. McCarthy and Carter (2002) argue that "an over-emphasis in language teaching on single words out of context may leave second language learners ill-prepared both in terms of the processing of heavily-chunked input such as casual conversation, as well as in terms of productive fluency"(p. 38).

One strategy for teaching clusters is suggested by Nattinger and DeCarrico (1992) who illustrated the advantages of teaching lexical phrases by following their acquisition process. At the early stage of language learning, learners are taught easier concordances then gradually exposed to more complex ones. The phrases can then be stored and rearranged in sentences and retrieved as required as whole chunks.

Another strategy suggested by Nesselhauf (2003), is to point out contrasts in collocations with the learners' native language. For example, in English it is common to say, *a serious illness*, whereas the Japanese equivalent substitutes *heavy* for the word *serious*. In order to avoid such non-native

usage in EFL contexts, teachers may be well-advised to draw L2 learners' attention to these contrasts.

One specific area of cluster deficiency among the NNS was in their use of vagueness markers. Vagueness is central to informal conversation and can be a ploy used when interlocutors cannot find the right words (Channell, 1994). Again, as per Savignon (1997), we recommend the introduction of terms that enhance strategic competence at an early stage of language learning, perhaps "even before the acquisition of any grammatical competences" (Savignon, 1997, p. 49). The same can be said about face and politeness clusters that also play an important role in the mutual preservation of face and smooth interaction between interlocutors, as well as the polite progression of conversation (McCarthy, 1999). Finally, as indicated by the results of this study, and in line with the present recommendations, we also suggest adding hedging devices (*sort of*, *kind of*, etc.) to the lexical syllabus. Regardless of teaching strategy, the necessity of bringing clusters into the curriculum is underlined by a comparison of the learner and native corpora. A recent corpus-based study on spoken vocabulary in textbook conversations, however, identified that these clusters are rarely incorporated to EFL materials for Japanese learners (Shirato, 2005).

Conclusion

The results of this study lead us to the conclusion that the vocabulary currently acquired by Japanese NNS differs markedly from the NS norm. Our qualitative analysis has illustrated significant aspects of the characteristics of NS conversational vocabulary, in which softness, indirectness, hedges, and vagueness are abundant. These characteristics may be considered a basic defining feature of spoken lexis, although they have been given scant attention in Japanese formal education, in which a mastery of written language has been a major concern. Furthermore, an overemphasis has been put on teaching individual words in formal education, causing a general neglect of multi-word clusters in language teaching. Consequently, a lack of these interactive words and phrases is likely to make NNS utterances sound blunt and pedantic, explaining learner *foreign-soundingness* (Crystal and Davy, 1975).

In conclusion, it appears essential to provide learners with the lexical tools that will bring them closer to the native norm in each register. The most pressing task for us now is to prepare a more use-centered vocabulary instruction based on the findings of corpus-based research. We believe that the comparison of learner corpora with NS corpora can add insight into this goal.

Junko Shirato is a doctoral student in the Graduate School of International Media and Communication at Hokkaido University. She is currently employed as a part-time English instructor at Sapporo City University. Her research interests include computer-mediated contrastive learner corpus analysis, second language vocabulary acquisition, and phraseology. <shirato@ilcs.hokudai.ac.jp>

References

Aijmer, K. (1996). *Conversational routines in English. Convention and creativity*. London: Longman.

Aijmer, K. (2002). *English discourse particles*. Amsterdam: John Benjamins Publishing Company.

Allerton, D. J. (1984). Three (or four) levels of word co-occurrence restriction. *Lingua* 63, 17–40.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Channell, J. (1994). *Vague language*. Oxford: Oxford University Press.

Crystal, D., & Davy, D. (1975). *Advanced conversational English*. London: Longman.

McCarthy, M. (1999). What constitutes a basic vocabulary for spoken communication? *SELL 1*, 233-249.

McCarthy, M., and Carter, R. (2002). This that and the other: Multi-word clusters in spoken English as visible patterns of interaction. *Teanga: The Yearbook of the Irish Association for Applied Linguistics*, *21*, 30-52.

Nattinger, J. R., and DeCarrico, J. S. (1992). Lexical phrases and language teaching. Oxford: Oxford University Press.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24, 223-242.

Savignon, S. J. (1997). *Communicative competence--Theory and classroom practice* (2nd ed.). New York: The McGraw-Hill Companies, Inc.

Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press.

Scott, M. (1999). *Wordsmith tools*. Software. Oxford: Oxford University Press.

Shirato, J. (2005). A Corpus-based analysis of basic spoken vocabulary in EFL textbook conversations. *Hokusei Gakuen University Graduate School Literature Review* 2, 15-31.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics* 21, 463-489.