English Lexical Phrases Used by Japanese University Students

Tomoko Kaneko

Abstract

This paper explores frequency and characteristics of English vocabulary and lexical phrases used by Japanese university students in written essays using four different native speaker corpora and International Corpus of Learner English Japanese sub-corpus. The results show that learner corpus contains only a few function words and lexical phrases and that the students seem to learn lexical phrases in two ways; by memorizing the string of words as a whole and by combining a few function words to a core content word and also by combining more than two lexical phrases.

In what aspect does English written by Japanese university students differ from that written by native speakers? This paper explores some language gaps between Japanese students' and native speakers' English using corpora. The focus of this study will be on the frequency and characteristics of vocabulary and lexical phrases used by the students. Through this study, it is hoped that we can get an insight into how to improve the students' English language use.

1. The importance of lexical phrases in language learning

How much vocabulary do English native speakers have? Nation (2001) suggests that educated native speakers of English know around 20,000 word families. Word families are defined as classes of words that depend on the knowledge that a person must have in order to use a language acceptably (Miller 1996). Nation (2001) further states that these estimates are rather low because the counting unit is only limited to word families and proper nouns are not included in the count. Nevertheless, he concludes that a very rough rule of thumb would be that for each year of their early life, native speakers add, on average, 1,000 word families a year to their vocabulary. Nation (2001) explains that a word family consists of a head word, its inflected forms and its closely related derived forms. Thus the word family of a verb "to *think*" for example, includes *think* (base form), *thinks, thinking* (regular inflections), *thought* (irregular inflection), *thinker, unthinking* (transparent derivations), *unthinkable* (less transparent derivation) and *t'ink* (other minor variation). He estimates that there are, on average, between 1.5 and 4 derived forms plus 2 or 3 inflections, which leads to a native speaker vocabulary size of between 70,000 and 140,000 words.

The magic of vocabulary learning in mother **tongue** seems to be the acquisition of lexical phrases. Some researchers believe that lexical **phrases** are essential in language learning

because the stored sequences of words are the bases of learning, knowledge and use. Ellis (1994) argues that a lot of language learning can be accounted for by associations between sequentially observed language items, that is, lexical phrases make language learning easier without the need to refer to underlying rules. Erman and Warren (2000) analyze the construction of native English speaker vocabulary and calculate that lexical phrases of various types constitute 58.6% of the spoken English discourse and 52.3% of the written discourse. The fact that native speakers use lexical phrases for more than half of their utterances, leads to the question if non-native speakers, particularly Japanese learners of English, also use as many lexical phrases as native speakers.

Most lexical phrases consist of several function words plus a few content words. Thus, we can suppose that the more lexical phrases are used, the more function words are needed. Consequently, by comparing the ratio of the use of function words and content words, we hope to gain an insight into the nature of lexical phrases used by Japanese students. Another way to answer this question is to look at the n-grams. "N-gram" is a name for a recently developed method for analyzing the frequency of n-words combinations in corpus using a computer program.

In addition to the above mentioned quantitative data, we need qualitative data if we are interested in the way we can improve the learners' English proficiency. This will also be explored by looking at the learners' lexical phrase errors in detail within the present paper.

Thus the following three research questions are raised:

1. How does the ratio of function words and content words in Japanese learner corpus compare to that in native speaker corpora?
2. How does the use of lexical phrases by Japanese university students compare to that by native speakers?
3. What are the typical features of errors seen in the use of lexical phrases by Japanese students learning English?

Before we go on to the explanation of data and analysis, we need to define lexical phrases. In fact, lexical phrases seem to exist in so many forms that it is presently difficult to develop a comprehensive definition of the phenomenon. Wray (2002) found over fifty terms to describe the phenomenon of lexical phrases, which includes; canned speech, chunks, collocations, conventionalized forms, formulaic expressions, formulaic language, formulaic sequences, formulaic speech, formulas, holophrases, memorized sequences, multiword units, prefabricated sequences, prefabricated routines, and ready-made utterances. Nattinger and DeCarrico (2001) present an insightful book on this phenomenon emphasizing a lexico-grammatical feature, for which the term "lexical phrases" is used. For the purpose of this paper, we will refer to this phenomenon as "lexical phrases". Matsuno & Sugiura in Sugiura (2004) review most of the definitions of collocations and explain that in addition to the central notion that "collocations are

actual words in habitual company" proposed by Firth (1957), there are other criteria proposed by several researchers. Those are:

1) Collocations should be frequently used combinations of words.

2) The words used in collocations should be strongly combined, and must be learnt as a whole.

3) The way the target words are combined in collocations should be limited by structure, semantics and arbitrariness.

4) The meaning of collocations should be conceptually predictable (if not, they are idioms). In this study, we'd like to consider all the criteria mentioned above in order to decide which combinations of words we would define as "lexical phrases". Thus the researcher's position is that no idioms and a limited number of n-grams will be considered as lexical phrases.

2. The study

In this section, the data for the present study will be described first. Then the procedure of the study will be explained. Finally after the statement of the results, a brief discussion about the results will follow.

2-1. The data

The data used as Japanese learners' English samples come from ICLE-J (International Corpus of Learner English-Japanese sub-corpus). In total, it includes 144 essays written by Japanese advanced level (junior & senior) university students between 1998 and 2000. Four kinds of native speaker corpus are compared. The FLOB (The Freiburg-LOB (The Lancaster-Oslo/Bergen) Corpus) is used to provide samples of British English and the FROWN corpus (The Freiburg-Brown Corpus – The Standard Sample of Present-Day American English) is used to provide samples of American English. Both of these corpora were collated from 15 different sources, including newspapers and novels written during 1991. LOCNESS (Louvain Corpus of Native Essays), a parallel native version of ICLE, will also serve as samples of native English. This corpus has been informally collated by Univeristé catholique de Louvain in Belgium in order to serve as comparative native speaker corpus to ICLE. In the present study, only the file collated from American English speakers is used. WS (The Wellington Corpus of Written New Zealand Corpus) is used as another sample of native English. It was developed in the Department of Linguistics at Victoria University of Wellington between 1986 and 1992.

ICLE-J also has error tagged sub-corpus. The sub-corpus will be utilized to find out what kind of errors the students make in using lexical phrases.

2-2. Analysis and Results

Chart 1 is the list of frequency of words in the five corpora.

Chart 1. Descriptive summary of the five corpora

corpora / frequency	NNS	NS			
	Japanese	British	American		Australian
	ICLE-J	FLOB	FROWN	LOCNESS	WC
tokens	76,875	1,237,425	1,383,467	143,463	1,244,453
types	5,718	45,089	43,252	10,667	40,363
type/token ratio	7.44	3.64	3.13	7.44	3.24

Notes) NNS = non-native speakers NS = native speakers

There were two steps employed to find out the ratio of function words and content words in each corpus. Firstly, the top fifty words from each frequency list were grouped by the researcher according to parts of speech (Appendix 1). As Nation (2001) states, the list shows surprisingly high agreement on frequently used words found in the three large size corpora (FLOB, FROWN & WC). LOCNESS still shows about 80% agreement to the other three corpora although its size is less than 12% of them. In contrast, ICLE-J only shows a 58% correlation. This in itself highlights an substantial difference in the way Japanese learners use English. Chart 2 summarizes the frequency of function words and content words according to the parts of speech. The grouping of content words and function words is not consistent among researchers, thus we'd like to follow a traditional categorization here. Some words are used in several different categories. For the purpose of this analysis, *that*, *so*, and *as* are categorized as conjunctions, *which* as an interrogative, and *more* as an adverb. *Be* verbs are categorized as verbs.

Chart 2. Frequency of function words and content words seen in the top 50 words

corpus / word types		NNS	NS			
		Japanese	British	American		Australian
		ICLE-J	FLOB	Brown	LOCNESS	WS
C W	noun	7	0	0	1	0
	verb	12	9	9	7	10
	adjective	2	1	1	1	1
	adverb	4	3	3	4	4
	total	25 (50%)	13 (26%)	13 (26%)	13 (26%)	15 (30%)
F W	preposition	7	9	9	9	9
	pronoun	6	14	16	12	13
	article	2	3	3	3	3
	conjunction	8	7	6	7	7
	auxiliary	2	2	2	4	1
	interrogative	0	2	1	2	2
	total	25 (50%)	37 (74%)	37 (74%)	37 (74%)	35 (70%)

Notes) CW = content words FW = function words

As a next step, 2, 3, 4, 5, and 6-grams of the five corpora were listed using a computer analysis tool called *Collocate* by Barlow (2004). Based on our definition of lexical phrases stated in section 1, only n-grams which carry lexico-grammatical function were picked up. The lists of lexical phrases picked up in this way from the original n-gram lists are shown in Appendix 2. This list only includes the sequences of words designated by native speakers of English as lexical phrases and selected by the researcher as not specially triggered by the content of the essay among the top 100 n-grams. Again, FLOB, FROWN & WC show high agreement, and LOCNESS shows a certain level of agreement. On the other hand, ICLE-J shows only a low level of agreement with the other corpora. Chart 3 shows a comparison made between NS and NNS corpora in 3 categories: 1) phrases common in NS and NNS corpora, 2) phrases appearing only in more than three NS corpora, and 3) phrases appearing only in the NNS corpus.

Chart 3. Comparison of lexical phrases in the five corpora

category grams	Common in NS & NNS corpora	Appearing only in more than 3 NS corpora	Appearing only in NNS corpus
2-grams	have to	such as out of more than	for example each other not only so on I think
3-grams	a lot of in order to	as well as a fact that as $\sim$ as a number of in terms of in front of a couple of as a result a series of	and so on more and more
4-grams	at the same time on the other hand in the $\sim$ of at the $\sim$ of	in the case of as a result of for the first time for a long time in the first place when it comes to	it is said that
5-grams	(none)	(none)	what I want to say
6-grams	(none)	(none)	it will be possible to say

In order to find out what kind of errors the students make in using lexical phrases, 266 usage errors listed as lexical phrase errors in error tagged ICLE-J data were further analyzed. As for the total number of lexical phrases used in ICLE-J, Murao & Sakagami in Sugiura (2004) recognize a total of 459 examples. Although we can not gain a statistically reliable accuracy

rate using the numbers shown by two different studies, it is shown that 44.2% is an approximation of lexical phrase accuracy in ICLE-J by comparing Sugiura's total number of collocations with the number of lexical phrase errors in the ICLE-J error tagged sub-corpus.

The types of Japanese students' lexical phrase errors are varied. Sometimes errors occur because the form of the phrase is not correct, and sometimes because the linguistic contexts where they are used are not suitable. However, in most of the cases, they occur because ill-forms are used in non-appropriate linguistic contexts as shown in Chart 4. In the Chart, the mark [LP] shows the beginning of the lexical phrase error and the word(s) enclosed by $ marks show(s) the suggested correct form.

Chart 4. Types of lexical phrase errors by Japanese learners

errors / types		examples	appro-priate context	correct form	frequency (total 267)	
1		~ their country has power [LP] at the same time $while$ it also ~	NO	YES	2 (0.8%)	
2	A	[LP] At the result $As a result$, all of the ~	YES	NO	27	62 (23.2%)
	B	~ read and write English [LP] a certain extent $to a certain extent$ ~	YES	NO	10	
	C	USA [LP] is used to $used to$ be a British colony ~	YES	NO	4	
	D	~ a listener can understand [LP] intention of us $our intention$.	YES	NO	21	
3		~ was abolished [LP] by the reason that $because$ it ~	NO	NO	203 (76.0%)	

In type 1, the form of the lexical phrases is correct, however, the linguistic context is not appropriate. In type 2, although the context is appropriate, an ill-form appears. In 2A, a part of the target form is changed to another word. In 2B, a part of the target form is missing. In 2C, an unnecessary word is added to the target form, and in 2D, the construction of the target form is incorrect. This illustrates that type 2 may represent lexical phrases which students are in the process of learning. By far, type 3 errors, where an incorrect form is used in a non-appropriate context, outnumbers the other two types.

3. Discussion

In this section, we'd like to discuss the results in order to answer the research questions. The type/token ratio shows that NNS have much less variation in vocabulary. The striking difference from the NS corpora is that NS corpora show high agreement concerning what particular words should be included in the top 50 list, while the NNS corpus doesn't show the same phenomenon. In all the 4 NS corpora, *the* is the most frequently used and *of*, the second.

The low ranks of these words in the NNS corpus suggest the underuse of these function words by Japanese students. Furthermore all the NS corpora hold *the, of, and, to* and *a* in the top five, while in the NNS corpus only the words *to* and *the* are shared with the NS corpora. The high frequency of the word *English* may be due to context. It is notable that the word *I* is ranked third, considering the fact that the word is, at highest, around the top 15 in the NS corpora. Although the high frequency of *I* is an interesting phenomenon, considering the fact that Japanese speakers often omit the subject *I* in their L1, the possible reasons for this phenomenon are a topic for another study. In the following three sections, we would like to focus on the ratio of function words and content words to find out some typical features in the use of lexical phrases by Japanese learners of English.

3-1. Function words and content words

There are many function words in NS corpora, while there is only a few of them in the NNS corpus as shown in Chart 2. By utilizing more function words appropriately, the students will be able to make the most of the few learned content words. Use of function words makes it possible for students to send clearer and more detailed messages. For example, if a student only says *book* instead of *the book*, it is impossible for the student to communicate that she is talking about a book she has mentioned before.

3-2. Lexical phrases

The frequency and variation of lexical phrases in the NNS data is much less than that in the NS data. We can find some typical features of the lexical phrase use by NNS. First of all, NNS used a totally different set of lexical phrases from NS (ex. *I think, and so on, it is said that* and *what I want to say*). Most of them seem to be transfers from L1 due to the fact that those lexical phrases are often taught in classrooms based on the frequent requests from the learners in L1-L2 translation activities. Secondly, NNS do not use lexical phrases which can be applied to a variety of expressions. Ellis (in press) calls those types of lexical phrases "patterns". Examples are *as ～ as, in (at) the ～ of*, and so forth. Thirdly, NNS overuse 2-grams but they underuse more than 3-grams. De Cook (2000) finds that when compared to native norms, some formulaic sequences were overused, some underused, and others misused by NNS. In the data presented, the lexical phrases in the column "appeared only in NS corpus" are the ones underused by students, while, the ones in the column "appeared only in NNS corpus" are overused. Fourthly, the phenomenon that the phrase *I want to* appears in the NNS column while *I'd like to* appears in the NS column is insightful in that it shows, like the previous results, that the amount of genuine input is not enough for the NNS to be able to use those lexical phrases correctly.

As for the poor use of lexical phrases by L2 learners, there are several arguments. One argument concerns the reason for the less frequent use of lexical phrases by L2 learners. Irujo

(1993) studied the formulaic language of L2 learners and concluded that their language tended to lag behind other linguistic aspects. This may again be partly due to a lack of rich input. Irujo (1986) suggests that idioms are often left out of speech addressed to L2 learners and that there is a tendency for learners to avoid the use of idiomatic language. What Irujo states is very plausible, at least through the researcher's own experience as an EFL learner. The second argument concerns the small variety of lexical phrases used by L2 learners. Schmitt (2004) introduces Granger (1998) and points out her statement that there is a tendency to stick with familiar and 'safe' sequences which the learners feel confident in using. This may partly explain the overuse of lexical phrases like *and so on*. Nation (2001) argues that collocational knowledge is in the center of language knowledge. This implies that all fluent and appropriate language use requires collocational knowledge. Chart 3 shows how little lexical phrases, especially 3/4-grams, NNS use in writing.

As it has been stated before, there is a consensus that children learn and use L1 lexical phrases before they have mastered the sequences' internal makeup. Schmitt (2004) states an interesting view on how much lexical phrases children use in learning their mother tongue. He explains that the acquisition of lexical phrases might depend to some extent on whether they are 'system learners' or 'item-learners'. He introduces Nelson (1973), who found that children who had referential preferences (naming things or activities and dealing with individual word items) usually learned more single words, particularly nouns, while children who had more expressive tendencies (having interactional goals and focusing on the social domain) were more likely to learn whole expressions which were not segmented. Schmitt (2004) explains that one of the reasons for these preferences may reflect wheather "the child 'supposes the language to be useful for' predominantly naming things in the world or engaging in social interaction (p.11)." He concludes that we need to wait for further studies on why learners use lexical phrases, yet "regardless of the underlying reason, there seems to be a link between the need and desire to interact and the use of the formulaic sequences (p.11)." In SLA, we need to wait even longer to find out why this phenomena takes place in second language learning, nevertheless, it is a profoundly interesting question to ask.

Although Japanese learners of English at this stage of interlanguage do not use as many lexical phrases as NS, they seem to be in the process of learning more of them. At the same time, we need to offer a lot of chances for students in order for them to be able to catch more lexical phrases and to build up their own system in order to produce more target-like lexical phrases. It is because language is learned both as a system and as a whole chunk, although some learners prefer chunk learning and others prefer system learning.

3-3. Lexical phrase errors by NNS

Over 75% of collocation errors occurred in non-appropriate contexts using ill-forms, however, nearly a quarter of all collocation errors occurred in appropriate context using ill-forms.

It can be argued that the second group of examples shows that those lexical phrases are now in the process of being learned.

We definitely need some developmental data to say anything about the process of learning, however, the result of this study suggests that the students do not learn lexical phrases only as chunks. For learners to use a lexical phrase *in the case of*, for example, there are two different ways to master it; starting from chunks or from separate words. Type 1 in Chart 4, is a typical example of chunk learning, top-down processing, while type 2C and 2D show the learning starting from separate words, bottom-up processing. It is true that the students learn lexical phrases as chunks only through genuine input, just like in L1 learning. Thus it seems useful for students to memorize some essential chunks, especially at the beginning stage of language learning. The other way for students to acquire lexical phrases is to learn how to combine learned words together. For example, *in the case of* may be learned firstly as separate words; *case* (N), *the* (ARTICLE), *in* (PRERP), and *of* (PREF), secondly as a 2 word-set; *the case* (ARTICLE + N), *case of* (N + PREP), thirdly as a 3 word-set; *in the case* (PREP + ARTICLE N), *the case of* (ARTICLE + N + PREP), and finally *in the case of*, which represents Type 2. This suggests that students acquire lexical phrases by starting with short basic lexical phrases, which they then break in smaller pieces, so that the pieces will be combined with other words and phrases as they are exposed to lots of genuine input. Schmitt (2004) states that the automatic use of lexical-phrase sequences allows chunking, while freeing up memory and processing resources. He concludes that "the capacity surplus gained in this way can then be utilized to deal with conceptualizing and meaning, which must surely aid language learning (p.12)."

4. Summary

The use of only a small amount of function words by NNS leads to the finding that the students underuse most lexical phrases. The NNS also use peculiar phrases which are not used by NS. Over half of the phrases are errors. Since lexical phrases facilitate further language learning (Schmitt 2004), the most effective way for NNS to be able to use target-like language is to use more target-like lexical phrases. It will be attained mainly by receiving lots of genuine inputs, which are full of lexical phrases.

By raising the learners' consciousness of function words, and the forms and meanings of lexical phrases, I believe that it is possible to promote accuracy, fluency and complexity of lexical phrase use by the learners. Combining content words and function words to make lexical phrases, and deconstructing lexical phrases to produce more varieties of language are the two effective routes to improve their English proficiency. Once the students learn that both the top-down and bottom-up routes for learning are effective, it is plausible that learning lexical phrases facilitates a more rapid gain of more target-like proficiency in the students' English.

References Consulted

Barlow, M. 2004. *Collocation*. Software introduced at the workshop at Showa Women's University, October 24, 2004.

Dagneaux, D., S. Denness, S. Granger, and F. Meunier. 1996. *Error Tagging Manual Version 1.1*. Centre for English Corpus Linguistics, Univeristé catholique de Louvain.

De Cock, S. 2000. Repetitive phrasal chunkiness and advanced EFL speech and writing. In *Corpus Linguistics and Linguistic Theory*, C. Mair and M. Hundt (eds), 51-68. Rodopi.

Ellis, R. 1994. *The Study of Second Language Acquisition*. Oxford University Press.

Ellis, R. In press. *Analyzing Learner Language*. Oxford University Press.

Erman, B. and B. Warren. 2000. The idiom principle and the open-choice principle. *Text* 20: 29-62.

Firth, J.R. 1957. *Paper in Linguistics 1934-1951*. Oxford University Press.

Granger, S. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In *Phraseology: Theory, Analysis and Applications*, S.P. Cowie (ed.), 145-160. Oxford University Press.

Irujo, S. 1986. A piece of cake: Learning and teaching idioms. *ELT Journal* 40: 236-242.

Irujo, S. 1993. Steering clear: Avoidance in the production of idioms. *International Review of Applied Linguistics in Language Teaching* 31: 205-219.

Miller, G. 1996. *The Science of Words*. Scientific American Library.

Nation, P. 2001. *Learning Vocabulary in Another Language*. Cambridge University Press.

Nattinger, J. R. and J. S. DeCarrico. 2001. *Lexical Phrases and Language Teaching*. Oxford University Press.

Nelson, K. 1973. *Structure and Strategy in Learning to Talk*. Monographs of the Society for Research in Child Development 149: 1-2.

Schmitt, N. (eds.) 2002. *An Introduction to Applied Linguistics*. Arnold.

Schmitt, N. (eds.) 2004. *Formulaic Sequences*. John Benjamins.

Sugiura, M. (eds.) 2004. *How can the native English speakers tell that a speaker is not a native speaker but a learner once they hear the learner speak?* Report of the Grant-in-Aid for scientific Research (C)(2)(2001-2003) Supported by Japan Society for the Promotion of Science. Graduate School of International Development, Nagoya University.

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press.

Appendix 1. A list of the top 50 words and parts of speech in the 5 corpora

FO	British		American			Australian		Japanese	
	FLOB		FROWN		LOCNESS	WS		ICLE-J	
1	the	A	the	A	the	A	the	A	to	P
2	of	P	of	P	to	P	of	P	English	N
3	and	C	and	C	of	P	and	C	I	PN
4	to	P	to	P	and	C	to	P	the	A
5	a	A	a	A	a	A	a	A	is	V
6	in	P	in	P	is	V	in	P	and	C
7	that	PN	that	PN	in	P	is	V	a	A
8	is	V	is	V	that	PN	was	V	we	PN
9	was	V	was	V	for	P	that	C	in	P
10	for	P	he	PN	it	PN	for	P	of	P
11	it	PN	for	P	be	V	it	PN	it	PN

12	he	PN	it	PN	are	V	on	P	that	C
13	as	C	with	P	not	AD	with	P	language	N
14	with	P	as	C	this	PN	as	C	Japanese	N
15	on	P	his	PN	they	PN	I	PN	for	P
16	be	V	on	P	as	C	be	V	think	V
17	I	PN	be	V	have	V	he	PN	but	C
18	his	PN	at	P	with	P	at	P	have	V
19	by	P	by	P	on	P	by	P	so	C
20	at	P	I	PN	their	PN	are	V	people	N
21	had	V	this	PN	people	N	from	P	are	V
22	but	C	had	V	or	C	had	V	speak	V
23	not	AD	not	AD	was	V	this	PN	not	AD
24	this	PN	are	V	by	P	his	PN	as	C
25	are	V	but	C	would	AV	not	AD	can	AV
26	have	V	from	P	I	PN	but	C	they	PN
27	from	P	or	C	has	V	have	V	if	C
28	her	PN	have	V	an	A	they	PN	be	V
29	she	PN	an	A	more	AD	were	V	with	P
30	which	I	they	PN	from	P	she	PN	master	V
31	an	A	which	I	one	PN	or	C	many	AJ
32	or	C	one	PN	but	C	her	PN	study	V
33	you	PN	you	PN	because	C	an	A	students	N
34	they	PN	were	V	he	PN	you	PN	very	AD
35	were	V	all	PN	can	AV	their	PN	need	V
36	one	PN	her	PN	if	C	new	AJ	will	AV
37	all	PN	she	PN	these	PN	one	PN	Japan	N
38	their	PN	there	AD	many	AJ	which	I	my	PN
39	been	V	would	AV	there	AD	all	PN	because	C
40	there	AD	their	PN	at	P	there	AD	world	AV
41	has	V	we	PN	when	C	we	PN	when	C
42	we	PN	him	PN	we	PN	has	V	second	AJ
43	more	AD	been	V	all	PN	been	V	there	AD
44	if	C	has	V	his	PN	would	AV	more	AD
45	would	AV	when	C	do	V	up	AD	was	V
46	will	AV	who	PN	what	I	when	C	this	PN
47	so	C	will	AV	should	AV	so	C	learn	V
48	when	C	more	AD	also	AD	who	I	want	V
49	who	I	no	AJ	will	AV	out	AD	by	P
50	no	AJ	if	C	who	I	said	V	on	P

Note) A＝article AD＝adverb AJ＝adjective AV＝auxiliary verb C＝conjunction I＝interrogative N＝noun
P＝preposition PN＝pronoun V＝verb FO＝frequency order

Appendix 2. Comparison of lexical phrases in the five corpora

	British	American		Australian	Japanese
	FLOB	BROWN	LOCNESS	WS	ICLE-J
2-grams	out of 45	out of 32	have to 20	have to 63	I think 1
	more than 58	more than 50	such as 34	such as 67	want to 6
	have to 64	such as 59	because of 48	had to 74	need to 9
	of course 89	have to 81	out of 52	more than 95	have to 31
	a few 90	a few 83	could be 73		for example 34
	such as 93	want to 95	need to 74		of course 71
		at least 98			each other 74
					not only 86
					so on 99
3-grams	as well as 2	as well as 3	the fact that 2	as well as 8	a lot of 7
	a number of 8	a lot of 9	in order to 4	a number of 12	and so on 21
	the fact that 10	the fact that 13	as well as 9	a lot of 13	more and more 36
	in order to 22	in front of 14	should not be 11	the fact that 16	in order to 93
	in terms of 23	a number of 16	a lot of 21	in front of 40	
	a lot of 32	in order to 18	a part of 37	a couple of 41	
	in front of 38	in terms of 37	in favor of 56	in terms of 43	
	a couple of 45	as a result 47	as long as 63	would have been 52	
	as a result 63	in addition to 9	this type of 80	in order to 68	
	a series of 79	a couple of 60	whether or not 85	a series of 84	
	in relation to 82	as soon as 60	as a result 100	as far as 90	
	in the past 91	in the past 65			
	as soon as 94	a series of 72			
		a variety of 76			
		a kind of 80			
		by the time 80			
		as long as 85			
		as much as 95			
4-grams	for the first time 2	at the same time 2	on the other hand 2	at the end of 1	all over the world 8
	at the same time 4	on the other hand 4	as a result of 18	for the first time 3	at the same time 54
	in the case of 5	at the end of 7	a great deal of 23	at the same time 5	on the other hand 62
	on the other hand 7	for the first time 8	in the case of 32	in the middle of 6	for a long time 92
	as a result of 8	in the middle of 10	at the same time 32	on the other hand 7	it is said that 92

by the end of 10	in the case of 13	at the end of 32	a wide range of 9
in the middle of 12	as a result of 19	in the long run 39	as a result of 13
on the basis of 12	in the context of 19	when it comes to 39	in the case of 16
in the context of 15	in the form of 19	at the beginning of 39	in the form of 27
at the time of 19	in the first place 26	I would like to 39	a large number of 39
in the first place 19	on the other side 29	in the case of 78	from time to time 39
in the light of 19	at the beginning of 32	all over the world 78	at the same time 39
in the form of 22	in the face of 32		as a means of 49
a wide range of 24	by the end of 37		in the case of 49
the way in which 27	at the time of 39		nothing to do with 60
at the beginning of 31	on the basis of 40		at the back of 60
from time to time 37	on the one hand 40		on the other hand 67
in the course of 40	a great deal of 44		in the face of 67
for a long time 50	in a way that 44		as soon as possible 79
it is clear that 50	the back of the 44		I would like to 79
on the part of 55	nothing to do with 50		in the first place 79
a great deal of 66	for a long time 54		when it comes to 100
in the face of 66	in the course of 54		the same way as 100
in the same way 66	for the most part 60		for a few days 100
at the age of 77	on the part of 72		at the age of 100
at the top of 77	in the absence of 90		in the context of 100
in the history of 87	in the midst of 90		in the light of 100
in the sense that 87	the way in which 90		at the bottom of 100
towards the end of 87	turned out to be 90		for a long time 100
in so far as 99	when it comes to 90		

5-grams	the turn of the century 2	on the other side of 3	due to the fact that 4	the rest of the world 4	what I want to say 48
	by the end of the 9	the rest of the world 10	the rest of the lives 14	at the back of the 10	

in the same way as 21	has nothing to do with 26	only a matter of time 51	on the other side of 10	
it is not surprising that 32	the end of world war 26	in a number of ways 51	the turn of the century 17	
as a matter of fact 46	the turn of the century 48	from the long term effects 51	due to the fact that 19	
in a wide range of 46	had nothing to do with 68	the end of the season 51	on the far side of 19	
in such a way as 46	if it weren't for 68		much the same way as 51	
it has been suggested that 75	play an important role in 68		in the same way as 51	
it should be remembered that 75			it should be noted that 51	
on the face of it 75			there is no doubt that 51	
there is no doubt that 75			it is not surprising that 51	
			a long way to go 51	
			there was no need to 51	
6-grams in such a way as to 10	it has nothing to do with 14	the last day of the season 5	in much the same way as 20	it will be possible to say 80
from the point of view of 17	all the way up to the 66	for the rest of their lives 22	there can be no doubt that 20	
this is not to say that 17	from a moral point of view 66	it is only a matter of 60	in spite of the fact that 20	
before the end of the year 48	if it hadn't been for 66	it is important to remember that 60	does not mean to say that 20	
by the turn of the century 48		it is in these instances that 60		
this is not the place to 48				
we are glad to note that 48				

Note) The numbers show the ranking of the frequencies.

（金子　朝子　　英語コミュニケーション学科）