Recurrent Word Clusters Used by Asian Learners

—A Statistical Study of Differences—

Yuka ISHIKAWA
Nagoya Institute of Technology

Abstract

Recurrent word clusters have received more and more attention recently in applied linguistics. The aim of this study is to explore which word clusters are over- or underused by particular groups of Asian learners of English and to clarify this aspect of their language-learning processes. We focus on word clusters consisting of three or four words, and analyze essays written by several groups of Asian learners and by native speakers on the same topic under the same constraints in order to answer research questions 1) Are there any differences among certain groups of Asian learners or between them and native speakers in their use of word clusters? and 2) Do any word clusters tend to be used more frequently or less frequently than others by a particular group of Asian learners? The results suggest that the answers to both these questions are positive and that the essays can be divided into four groups by the features of their word cluster use: those written by Japanese learners, by other EFL learners, by ESL learners, and by native speakers. Specific findings include that Japanese learners tend to overuse "I think that" and "I agree that," while ESL learners tend to overuse certain linking adverbials, such as "on the other hand" and "at the same time."

Keywords

Word clusters, Asian learners, ICNALE, Essay writing

I Introduction

Since the 1980s, scholarly analysis of recurrent word clusters or "lexical bundles" in English texts has become more and more focused on applied linguistics, or the field of language learning and teaching (Granger, 1998; Schmitt and Carter, 2004; Hyland, 2008). Learners tend to use a far more limited number of words and strings of words than native speakers, and to use them repeatedly; in contrast, native speakers use a greater variety of *hapax legomena*, words used only once in the text. In this study, we will analyze English essays written by college students from various Asian countries

and autonomous entities (see section 3.2 below) in order to clarify which word combinations tend to be used recurrently by which group of learners.

The essay data to be analyzed are taken from the ICNALE, or International Corpus Network of Asian Learners of English, which is one of the largest learner corpora compiled in Asia. One thousand three hundred Asian learners and one hundred native speakers of English are involved in the ICNALE project. Each of them writes two essays: one on the topic "it is important for college students to have a part time job" and the other on the topic "smoking should be completely banned at all the restaurants in the country." Writing conditions are strictly controlled: learners are required to write a 250–300 word essay within 20–30 minutes, without using a dictionary. It therefore becomes easy to conduct a comparative study between essays written by this group of learners or a subgroup and those written by native speakers.

On the basis of these data, and focusing on the word clusters used by Asian learners in various countries and autonomous entities, this study will explore the linguistic features of their essays and compare them with those of essays written by native speakers, which may help illuminate the English learning processes that Asian learners undergo. There are two research questions here to answer in this study:

1. Are there any differences among certain groups of Asian learners or between them and native speakers in their use of word clusters?
2. Do any word clusters tend to be used more frequently or less frequently than others by a particular group of Asian learners?

II Literature Review

Language teachers and researchers in the field of language teaching and learning have recently become disenchanted with the Chomskian notion that a speaker can create a limitless number of utterances using a limited number of rules, and have instead focused on increasing evidence that most of texts are prefabricated word clusters combined to create meanings. Altenberg (2005), investigating the London–Lund Corpus of Spoken English, suggests that even native speakers uses recurrent word clusters and that especially in speaking, "over 80 per cent of the words in the corpus form part of a recurrent word-combination" (p. 102).

The general consensus among researchers and teachers is that successfully acquiring frequent word clusters or lexical bundles is crucial to language learning (Pawley & Syder, 1983; Wrey, 2002; Schmitt & Cater, 2004); however, there has been some confusion as a result of the terminology used. As Chen and Baker (2010) pointed out, the same term is sometimes used with different meanings, or different terms used to refer to the same thing. Biber et al. (1999) use the term "lexical bundle," defining it as follows:

> Lexical bundles can be regarded as extended collocations: bundles of words that show a statistical tendency to co-occur […] Lexical bundles are recurrent expressions, regardless of their idiomaticity, and regardless of their structural status. That is, lexical bundles are simply sequences of word forms that commonly go together in natural discourse. (pp. 989–990)

Altenberg (1998) uses the term "recurrent word combination," focusing on word clusters consisting of at least three words and occurring at least ten times in his data. Gläser (1998) uses "phraseological unit," defining it as "a lexicalized, reproducible bilexemic or polyleximic word group in common use, which has relative syntactic and semantic stability, may be idiomatized, may carry connotations, and may have an emphatic or intensifying function in a text" (p. 125). Hyland (2008) uses the term "clusters," defined as "recurrent strings of uninterrupted word-forms" (p. 43).

Despite this variety of definitions and emphases, there is general agreement on the basic definitions of the linguistic features they refer to: as Wood (2002) points out, word clusters or lexical bundles can basically be regarded as "multiword units of language that are stored in long-term memory as if they were single lexical units" (p. 2).

In this paper, we use "word cluster" to refer to any sequence of word forms consisting of two or more words used recurrently (at least twice) in a text.

III Research Design

3.1 Corpus Analysis Tool

This study uses AntConc 3.2.4w (Anthony, 2011), to make a list of word clusters, called "Clusters" or "N-Grams" by the software. Punctuation marks are not counted as words in this study; they are replaced with a space stroke in the counting process. Thus, for example, a three-word cluster, "part-time job isn't," with a hyphen and apostrophe, is identified as a five word cluster "part time job isn t," as two space strokes are added to the string as punctuation is processed. This means that "t," for instance, functions as a presumptive word, a negative contraction. In the analysis applied in this study, the case of letters is not relevant.

3.2 Data

The ICNALE, a learner corpus with special attention to controlling writing condition, involves 200 essays written by native speakers and 2,600 by learners of English from ten Asian countries and autonomous entities. The Asian learners involved can be divided into two groups: learners with an ESL background and learners with an EFL background. The number of learners belonging to the former group—from Hong Kong, Indonesia, Pakistan, the Philippines, and Singapore—is 700, while the number in the latter group, from China, Japan, Korea, Thailand, and Taiwan, is 1,900.

As we have defined "word cluster" as such, in this study, we initially counted word clusters consisting of at least two words and occurring at least twice. The results show that, for example, in the English native speaker (ENS) files, which are essays by native speakers, there are 29,055 different types of word cluster used recurrently in the corpus. The total number of word clusters used in the corpus, tokens, is 120,561. As Fig. 1 shows below, more than half of them are two-word clusters. However, many of these are fragments of larger structures, as Altenberg (1998) points out—for instance, "a part," "part time" and "time job," which are a part of "a part time job," or "in the" and "is a," which are of only secondary importance to semantic meaning. Therefore, as a matter of convenience, we will limit our examination here to word clusters composed of three and four words, discarding shorter and longer clusters.

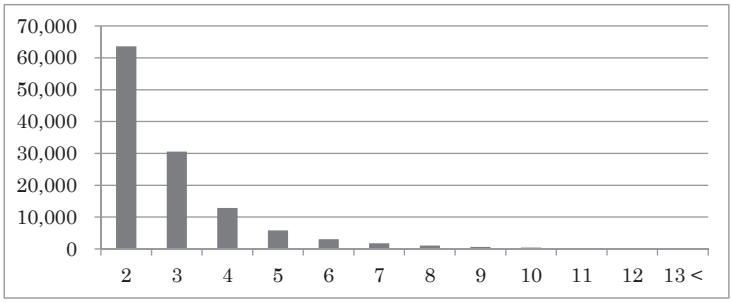

Fig. 1 Frequency and length of recurrent word clusters in ENS data

The data to be examined in this study are those from ICNALE and contained in the "Base" folder, released in Jan. 2013 (Ishikawa, 2013). Setting aside the ENS files, the files for Asian learners are classified into ten groups according to the country or entity from which the learner comes. In addition, each group is divided into two sub-groups by topic: PTJ (about the part time job) and SMK (about smoking). Thus, a total of 22 sub-corpora are used, and the outline is shown in Table 1 below.

Table 1 Types and tokens of three and four word clusters for each group

Topic	#	NS	ESL						EFL			
		ENS	HKG	PAK	PHL	SIN	CHN	IDN	JPN	KOR	THA	TWN
PTJ	Type	5616	3423	7958	7617	6371	16380	8880	14712	10302	15518	6344
	Token	20407	12857	27470	28703	26187	74113	31291	75117	43860	63207	26314
SMK	Type	5557	2941	7714	7773	6048	14321	8285	13701	9652	15072	5757
	Token	18683	9787	23800	25996	22775	62994	26378	61614	37187	53917	22296

3.3 Method

Several word clusters found in the data seem to be peculiar to the two topics dealt

with in the ICNALE. Thus, in order to eliminate these word clusters (such as “a part time job” or “smoking in restaurants”) and obtain only word clusters used by a particular group regardless of topic, we took the following steps. First, all word clusters consisting of three or four words in each sub-corpus were listed to make 22 lists. Then, the 22 lists were paired by country or area, to make 11 pairs of lists. For example, the word cluster list made from the ENS-PTJ corpus was paired with the one made from the ENS-SMK corpus. Next, word clusters common to the two lists in each pair were extracted to make a new list, and word clusters peculiar to one list deleted. Some clusters that were retained, such as “part time job” or “the restaurants in,” appeared in both lists but disproportionately in one due to their special relation to the topic; while these are likely common clusters in general language, their frequency is likely boosted in the present study by this special relation. Therefore, all possible word clusters with three and four words in the two topics given—“it is important for college students to have a part time job" and “smoking should be completely banned at all the restaurants in the country”—were deleted from the lists. Their variations such as “it’s important” or “have part time jobs” were retained. Fig. 2 below shows examples of deleted word clusters.

```
it is important
it is important for
   is important for
   is important for college
      important for college
      important for college students
                for college students
                for college students to
                    college students to
                    college students to have …
```

Fig. 2 Example of deleted word clusters

Thus, we obtain 11 new lists of word clusters that can be supposed to be used frequently by learners in a particular place or by native speakers, respectively. It is highly probable that the word clusters on the list would be used by the same writers in similar argumentative essays under the same conditions, even with a different topic.

The next step is to combine the 11 lists and sort the word clusters by frequency. The 100 most frequent word clusters are used to conduct a correspondence analysis (CA) in order to answer the research questions presented in the first chapter.

IV Results and Discussion

The CA results are shown in Table 2 below. The first items are the 11 groups of essay

writers and the second items are the 100 most frequent word clusters. The cumulative contribution rate up to the second axis exceeds 50%, and the eigenvalue of the first and second axes is over 0.1. We will therefore examine a map where the items on the first (F1) and second (F2) axes are plotted separately, as seen in Figs. 3 and 4.

Table 2 Summary of correspondence analysis results

	F1	F2	F3	F4	F5	F6	F7	F8
Eigenvalue	0.168	0.101	0.074	0.047	0.029	0.027	0.021	0.018
Contribution (%)	31.64	19.12	14.04	8.91	5.48	5.10	3.94	3.47
Cum. cont. (%)	31.64	50.76	64.80	73.71	79.19	84.30	88.24	91.71
Countries and entities								
ENS	-0.03	-1.70	-0.28	-0.17	1.20	1.88	-0.67	2.04
HKG	1.68	-0.40	0.22	0.60	-0.37	0.62	3.40	-1.93
IDN	0.06	0.79	1.35	-0.53	0.14	-1.90	-1.80	-0.05
PAK	1.40	-0.05	2.36	3.38	0.51	-0.59	0.69	1.30
PHL	1.58	-0.79	-0.20	0.05	-2.83	1.13	-1.33	-0.72
SIN	1.68	-2.48	-0.62	-1.60	0.82	-2.61	0.42	-0.03
CHN	0.70	1.38	-1.52	0.41	0.28	-0.22	-0.35	0.45
JPN	-1.37	-0.49	-0.41	0.59	-0.39	-0.38	0.17	-0.27
KOR	-0.06	0.14	0.92	-0.18	1.31	0.84	-0.98	-1.74
TWN	0.15	0.66	-0.45	-0.85	0.92	0.77	1.39	-0.53
THA	-0.35	0.75	1.09	-1.30	-0.80	0.15	0.69	0.98
ENS	-0.03	-1.70	-0.28	-0.17	1.20	1.88	-0.67	2.04

[NB: “cum.cont” stands for cumulative contribution]

4.1 RQ 1: Are there any differences among certain groups of Asian learners or between them and native speakers in their use of word clusters?

To answer RQ 1, a correspondence analysis was conducted. The first axis classifies Asian ESL learners (Singaporean, Philippine, Hong Kong, and Pakistani) at the right end as one group, and Japanese learners at the left end as another, with the others in between, as Fig. 3 shows below. The second axis distinguishes, at the respective ends, native speakers and Singaporean learners from some Asian EFL learners (Taiwan, Indonesian, Thai, and Chinese), with the others in between.

Japanese leaners are located far away from others, as Fig. 3 shows. It can therefore be said that Japanese learners use word clusters rather differently from other Asian EFL learners and from Asian ESL learners. We will examine those clusters more closely later in order to determine the details of this difference.

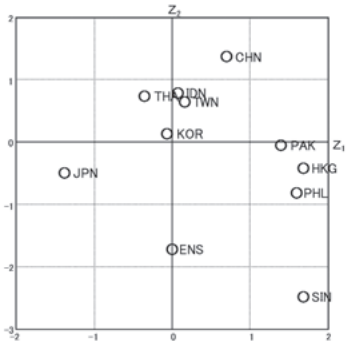

Fig. 3 Countries and entities plotted by CA

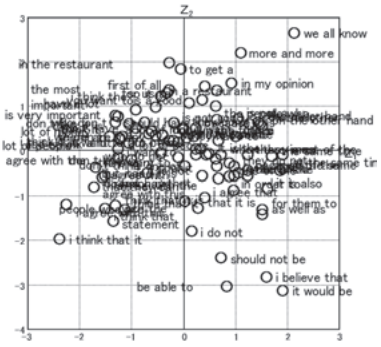

Fig. 4 Word clusters plotted by CA

The second axis distinguishes, at the respective ends, native speakers and Singaporean learners from Chinese learners of English, with the others in between. Word clusters that tend to be used frequently by Chinese learners include “we all know,” “more and more,” and “in my opinion,” and those that tend to be used frequently by Singaporean learners are “it would be” and “I believe that.”

The answer to RQ 1 is, therefore, positive. Asian learners can be roughly classified into three groups: a Japanese learner group, an EFL learner group, and ESL learner group. As for the last group, Singaporean learners can be further distinguished from other ESL learners on the second axis (Z2), as Fig. 3 shows. In this paper, we will limit our discussion to the first group, and examine the word clusters that are frequently used by Japanese learners in the next section.

4.2 RQ 2: Do any word clusters tend to be used more frequently or less frequently than others by a particular group of Asian learners?

We will take the Japanese learners to answer this question for simplicity, as they constitute a single national group as well as a single learner group as identified by the correspondence analysis. The items plotted at the right end of the map in Fig. 4 reveal that Japanese learners use particular word clusters, such as “at the same time,” “it would be,” “I believe that,” “on the other hand,” or “as well as,” less frequently than other ESL learners ($p<.01$). On the other hand, word clusters including “I think that” and “I agree with,” plotted at the left end, are used frequently by Japanese learners, in a rather peculiar way. For example, Japanese learners use “I agree with” approximately twenty times as frequently as ESL learners in Asia, and seven times as frequently as native speakers. However, as Fig. 5 below shows, they also use “I believe that” less frequently than ESL learners or native speakers. Native speakers use this word cluster almost seven times more frequently, and ESL learners 3.5 times more, than Japanese learners.

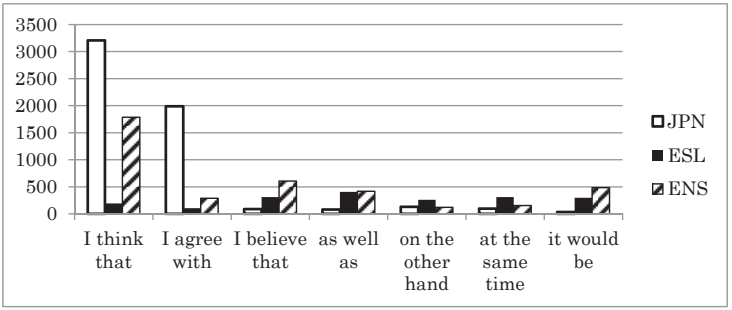

Fig. 5 Word clusters tending to be used frequently by Japanese learners of English (Y-axis points represent adjusted frequencies of each word cluster per one million.)

These results lead us to the conclusion that Japanese learners of English do not use in their writing, at an appropriate time in an appropriate way, so-called downtoner or hedge expressions, such as "I think that" "I believe that," and "it would be." They also use linking adverbials, such as "as well as" and "at the same time," which function to connect additional information to existing information, far less frequently. Japanese learners of English do, however, use linking adverbials that function to add concessive information to existing information such as "on the other hand." These tendencies might be attributed to L1 transfer, as expressions equivalent to "I think that" and concessive conjunctions are frequently used in Japanese texts.

Thus, for Japanese learners at least, the answer to RQ 2 is also positive. Japanese learners use "I think that" and "I agree with" more frequently, and "I believe that," "it would be," "as well as," and "at the same time" less frequently, than other Asian learners.

V Conclusion

Examination of recurrent word clusters used by Asian learners of English can illuminate various aspects of their language-leaning process. Some word clusters may be used with different frequency and in different ways depending on language proficiency and others depending on the linguistic or cultural background of the learner. In this paper, we have used ICNALE data and focused on word clusters consisting of three and four words to answer two research questions: 1) Are there any differences among certain groups of Asian learners or between them and native speakers in their use of word clusters? and 2) Do any word clusters tend to be used more frequently or less frequently than others by a particular group of Asian learners?

Analyzing the data, we have concluded that the answers to both research questions

are positive. Correspondence analysis has led to a rough grouping of essays written by Asian learners and native speakers respectively into those written by Japanese learners, other EFL learners, ESL learners and native speakers. Japanese learners use the word clusters "I think that" and "I agree that" highly frequently, but use "I believe that" and "it would be" far less frequently than other groups. Similarly, the linking adverbials "at the same time" and "on the other hand" tend be used more frequently by Asian ESL learners than by native speakers. Chinese learners and Singaporean learners also have a slight tendency to peculiar or particular use of some word clusters.

There are extensive implications here for the ESL/EFL teacher. For instance, teachers could guide Japanese learners to pay attention to and limit their usage of "I think that" and "I agree with" and to try to use more "supplementing" linking adverbials in argumentative essay writing, or they could help other Asian ESL learners not to use certain linking adverbials too frequently. By considering these tendencies, English learners can make their writing more native-like.

In future research, we need to examine the use of word clusters more closely, in order to clarify the language learning processes Asian learners follow. One research topic will be what types of Japanese learners tend to use the word clusters identified here more frequently or less frequently than others and why.

References

Altenberg, B. (2005). On the phraseology of spoken English: The evidence of recurrent word combinations. In A. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 101-122). Oxford, UK: Oxford University Press.

Anthony, L. (2011). AntConc 3.2.4w [computer software]. Available from http://www.antlab.sci.waseda.ac.jp/antconc_index.html

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

Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, *14*(2), 30-49.

Gläser, R. (1998). The stylistic potential of phraseological units in the light of genre analysis. In A. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 125-143). Oxford, UK: Oxford University Press.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. Cowie (Ed.) *Phraseology: Theory, analysis and applications* (pp.145-160). Oxford, UK: Oxford University Press.

Hyland, K. (2008). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41–62.

Ishikawa, S. (2013). *ICNALE: The international corpus network of Asian learners of*

English [data files]. Available from http://language.sakura.ne.jp/icnale/index.html

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike frequency. In J. C. Richards, & Schmidt, R. W. (Eds.), *Language and communication* (pp. 191-226). London: Longman.

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 1–22). Amsterdam: Benjamins.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.