Dialogue, 2006, Vol.5, pp. 1-18

ISSN 1349-5135

What is the reality of collocation use by native speakers of English?
－Collocation research based on the BNC corpus and the TIME corpus－

TAEKO KOYA
Hosei University

Abstract

This paper attempts to analyze high-frequency verb-noun collocations by native speakers of English in the BNC and the TIME corpus to examine basic collocations for Japanese learners of English. The author believes that identifying the reality of usage of collocations by native speakers of English is the first step to select basic collocations for Japanese learners of English because collocational competence is still regarded as one which only native speakers can establish or confirm (Crystal, 1992; McCarthy, 2004, August).

The findings show that high-frequency collocations in the BNC and the TIME corpus are fairly common, consist of basic verbs and nouns in reference to *JACET 8000* and tend to be used regardless of the topics. However, they are not necessarily equal to what is expected as the basic collocations for Japanese learners of English and they have to be examined in terms of the importance of collocations for them. Therefore, it is argued in the conclusion that this investigation necessitates another research from a pedagogical point of view.

Keywords: high-frequency collocations, native-speaker English, corpus

1. Introduction

Japanese secondary school students learn English to develop general English ability, not English for specific purposes, focusing on practical communication ability and to foster a positive attitude toward communication through English. In order to develop general communicative ability, the acquisition of basic collocations is prerequisite and it is supported in terms of memorization, fluent and appropriate language use, aspects of knowing words, word models and teaching effectiveness (Alexander, 1984; Ellis,

2001; Hill, 2000; Korosadowicz-Struzynska, 1980; Lewis, 1993, 2000; McCarthy, 1984; Nattinger and DeCarrico, 1992; Pawley and Syder, 1983; Yorio, 1980).

In spite of the agreement on the view of collocation mentioned above, there has been little pedagogical research on basic collocations in Japan, which leads to disagreement on basic collocations for Japanese learners of English. In fact, findings of collocation research in English textbooks conducted by Koya (2004) show that the textbook writers are inconsistent in basic collocations and disregard the collocation learning of Japanese learners of English.

In order to examine basic collocations for Japanese learners of English, a reference to the reality of collocation use by native speakers of English is important. This is because collocation is strongly related to culture and collocations have cultural connotations according to McCarthy (2004, August). He mentions that, because of the connotative features, collocations are regarded as right or wrong by native speakers of English and the clarification of the mechanism has been tackled by few researchers. Crystal (1992) also defines collocational competence as one which only native speakers can establish or confirm. I believe that identifying the reality of usage of collocations by native speakers of English is the first step in collocation study, which leads to the selection of collocations to be learned by Japanese learners of English in terms of their purpose of learning English.

2. Purpose and research questions

The purpose of this research is to examine which collocations are frequently used by native speakers of English in order to answer the main question, "What are the basic collocations for Japanese learners of English?" For this study, four research questions were set up:

1. What are high-frequency collocations in large corpora collected from native speakers of English?
2. What are features of those high-frequency collocations by native speakers of English?
   2a. Which levels of words are included in the high-frequency verb-noun collocations, in the word list of basic words for Japanese learners of English?
   2b. Are high-frequency collocations of native-speaker English related to topics?

3. Methodologies

3.1. Material: Corpus

Two corpora were used to examine high-frequency collocations of native-speaker English.

3.1.1. British National Corpus (BNC)

The British National Corpus (BNC) was selected in order to extract target collocations. It is one of the largest monolingual British English corpora in the world, containing some 100 million sample words of both written and spoken English. The reason this corpus was selected was that the whole text was easily obtained, via computer access and it had three main outstanding features as follow:

1. The BNC comprises 100,106,008 words of present-day English.
2. The BNC contains a wide range of both spoken and written British English.
3. The tagging of the BNC is carried out with a version of the CLAWs, a stochastic part-of-speech tagger developed at the university of Lancaster.

3.1.2. Making the TIME corpus

TIME (American edition) was also selected as a written English database of native speakers of English in order to extract target collocations. This weekly magazine has one of the largest circulations in the world and an audience of more than 300 million around the world. It also covers many kinds of topical news such as world, science & technology, art & entertainment, and it attracts many readers. As the English language used in *TIME* is regarded as the standard North American English, the TIME corpus offers the standard North American English collocations, while the BNC provides the standard British English collocations.

Table 1. Tokens and types in *TIME*

TIME	
(from December 1 in 2003 to March 29 in 2004, total 17-volumes)	
Total tokens	453117 words
Total types	36099 words

English of 17 volumes, December 1 in 2003 to March 29 in 2004, of *TIME* were collected and computerized into one large corpus. The size of the TIME corpus results is shown in Table 1.

Table 2. Four main categories

Main topics	Subordinate topics
Social science	Nation, World, Business, Society, Crime, Religion, Education
Science & Technology	Medicine, Space, Time in Depth, Health, Technology, Environment
Art & Entertainment	Book, Theater, Movie, Music, Television, Sports
Others (essays & opinions)	Essays, Interviews, Letters, Notebook, People, Your time, Life style, Viewpoints

Because *TIME* has more than 20 categories, the corpus was recategorized into main topics and subordinate topics as in Table 2.

3.2. Selection of collocations

In this analysis, verb-noun collocations were targeted because they are most frequently used combinations, are regarded as key combinations in producing clauses and sentences, and they are the most often selected in the previous empirical research (Bahns and Eldaw, 1993; Caroli, 1998; Nesselhauf, 2003).

In specifying verb-noun collocations which are used in the BNC and the TIME corpus, we used one word list needed for Japanese learners of English *the JACET List of 8000 Basic Words* (2003)(*JACET 8000*), and four collocation dictionaries*, *COBUILD English Collocations on CD-ROM* (1995), *Oxford Collocations Dictionary for Students of English* (2002), *the BBI Dictionary of English Word Combinations* (1997) and *The Kenkyusha Dictionary of English Collocations* (1995).

JACET 8000 is a latest word list combining the scientific viewpoint and the educational viewpoint. It not only refers to the rank of words calculated from the data of the BNC and a set of various sub corpora, such as a TOEFL corpus and a science magazine corpus, but also modifies the rank of words by examining how they are used in school textbooks. The word list made in this way consists of the basic 8000 words for Japanese learners of English and is ranked from the first 1000 basic words (Level 1) to the 7001-8000 words (Level 8). This scientific and educational word list is important to investigate high-frequency collocations by native speakers of English because this investigation is carried out to identify basic collocations for Japanese learners of English and they have to consist of necessary words for them. Therefore, all the nouns listed in this word list were extracted, amounting to 4986 nouns.

All the verbs collocated with the selected 4986 nouns in *JACET 8000* were checked if they were included in the four collocation dictionaries, resulting in 1572 collocations (Koya, 2005, pp. 276-300).

3.3. Procedure

First, the BNC was installed in my computer, then the TIME corpus was completed, and a selection of targeted collocations was finished. Then, whether the target collocations occurred in these two corpora was examined. As for span size, four words on either side of node are considered appropriate in this investigation, following previous research by Berry-Roghe (1973) and Jones and Sinclair (1974).

Collocations in the BNC were retrieved with SARA concordance software. Examination of the frequency of collocates for given nodes was conducted by using a collocation dialog box in SARA. In addition to the frequency check of the target collocations, the z-score was used, which is one of the most reliable statistical measures in calculating the strength of combinations between nodes and collocates within a certain span. The z-score shows that the higher it is, the more significant the clustering is. According to Barnbrook (1996) and Berry-Roghe (1973), more than a *three* z-score is significant as a collocation.

Target collocations in the TIME corpus were retrieved with TXTANA, a concordance software, to show whether a keyword collocates with certain other words within a certain span of context in a set-up corpus. In addition to the lemmatized forms of target verb-noun collocations (e.g. *make mistake*), infinitive and -ing forms of verbs (e.g. *to make a mistake* and *making a mistake*), and plural forms of nouns (e.g. *make mistakes*) were shown as occurrences of the related types of combinations in the TIME corpus. Checking the context where target collocations occur and deleting inappropriate examples that target node and the collocate occur in different sentences in reference to z-scores, the search for 1572 target nodes and the collocates was carried out.

4. Results

4.1. Analysis of the BNC

SARA, a concordance software, was used to calculate the frequency of 1572 target collocations in the BNC and the z-score per collocation. As is seen in Table 3, verb-noun combinations regarded as collocations were 1502 in the BNC as the z-score of 70 verb-noun combinations is under *three* (＜3.0), indicating that the combinations are clustered weakly and they are not judged as collocations by Bahnbrook (1996) and Berry-Rogghe (1973). According to these two papers, 1502 verb-noun combinations

were considered as collocations in the BNC.

Selected target collocations were checked with two corpus-based collocation dictionaries and two native speakers' intuition-based collocation dictionaries.

Table 3. Number of verb-noun collocations in the BNC

z-score under 3.0 (＜3.0)	70
No. of verb-noun collocations in the BNC	1502

Table 4. Level of the nodes and collocates per 100 collocations in the BNC

BNC		**L1**	L2	L3	L4	L5	L6	L7	L8	TOTAL	**L1+L2**
～100	No.	**78**	19	0	2	0	0	0	0	99	**97**
	%	**78.0**	19.2	0.0	2.0	0.0	0.0	0.0	0.0	100.0	**98.0**
～200	No.	**56**	31	5	7	1	0	0	0	100	**87**
	%	**56.0**	31.0	5.0	7.0	1.0	0.0	0.0	0.0	100.0	**87.0**
～300	No.	**42**	39	7	12	0	0	0	0	100	**81**
	%	**42.0**	39.0	7.0	12.0	0.0	0.0	0.0	0.0	100.0	**81.0**
～400	No.	**39**	33	10	12	2	4	0	0	100	**72**
	%	**39.0**	33.0	10.0	12.0	2.0	4.0	0.0	0.0	100.0	**72.0**
～500	No.	**32**	31	14	16	4	3	0	0	100	**63**
	%	**32.0**	31.0	14.0	16.0	4.0	3.0	0.0	0.0	100.0	**63.0**
～600	No.	**27**	36	11	15	4	3	2	0	98	**63**
	%	**27.6**	36.7	11.2	15.3	4.1	3.1	2.0	0.0	100.0	**64.3**
～700	No.	**21**	31	12	22	8	3	3	0	100	**52**
	%	**21.0**	31.0	12.0	22.0	8.0	3.0	3.0	0.0	100.0	**52.0**
～800	No.	**14**	37	20	20	6	2	1	0	100	**51**
	%	**14.0**	37.0	20.0	20.0	6.0	2.0	1.0	0.0	100.0	**51.0**
～900	No.	**4**	28	33	18	8	5	2	1	99	**32**
	%	**4.0**	28.3	33.3	18.2	8.1	5.1	2.0	1.0	100.0	**32.3**
～1000	No.	**14**	19	27	17	15	5	2	0	99	**33**
	%	**14.1**	19.2	27.3	17.2	15.2	5.1	2.0	0.0	100.0	**33.3**
～1100	No.	**4**	17	21	23	17	8	3	3	96	**21**
	%	**4.2**	17.7	21.9	24.0	17.7	8.3	3.1	3.1	100.0	**21.9**
～1200	No.	**1**	18	23	16	19	12	4	1	94	**19**
	%	**1.1**	19.1	24.5	17.0	20.2	12.8	4.3	1.1	100.0	**20.2**
～1300	No.	**4**	10	27	10	24	13	7	1	96	**14**
	%	**4.2**	10.4	28.1	10.4	25.0	13.5	7.3	1.0	100.0	**14.6**
～1400	No.	**1**	9	21	12	18	17	12	3	93	**10**
	%	**1.1**	9.7	22.6	12.9	19.4	18.3	12.9	3.2	100.0	**10.8**
～1500	No.	**0**	5	16	5	12	18	24	8	88	**5**
	%	**0.0**	5.7	18.2	5.7	13.6	20.5	27.3	9.1	100.0	**5.7**
～1572	No.	**5**	8	6	4	12	10	14	5	64	**13**
	%	**7.8**	12.5	9.4	6.3	18.8	15.6	21.9	7.8	100.0	**20.3**

L = Level

Table 4 shows frequency of 1502 collocations in the BNC and the levels of nodes and collocates per 100 collocations. Level 1 (L1) means the node and the collocate of a collocation consist of the first 1000 basic words in accordance with *JACET 8000* (2003). In cases where either the node or the collocate of a collocation is found the higher levels, they belong to that level. For example, when a collocation consists of L1 node and level 2 (L2) collocate, it is regarded as L2 collocation, because an individual who knows more than 2000 words is most likely to understand and produce L1 node and L2 collocations.

As is also seen in Table 4, 78.0% of collocations consisted of L1 nodes and collocates in the first 100 high-frequency collocations, after which the proportion of L1 nodes and collocates was steadily reduced: to 56.0 % in the second 100 high-frequency collocations, to 42.0% in the third, to 39.0 % in the fourth and so on. For L1 and L2 nodes and collocates, which are subject to be acquired by the Japanese twelfth graders, the first 100 high-frequency collocations made up 98.0 % and then the proportion of L1 and L2 nodes and collocates slightly decreased. In other words, the higher the frequency of collocations was, the more basic words might be comprised in the collocations, which leads to the observation that high-frequency collocations in the BNC consist of basic level verbs and nouns. These collocations are to be learned before entering universities.

4.2. Analysis of the TIME corpus

TXTANA, another concordance software, was used to calculate the frequency of 1572 target collocations in the TIME corpus. Table 5 shows that 581 combinations out of 1572 target collocations occurred in it, although 466 nouns and 180 verbs of target collocations individually appeared. About a third of targeted collocations appeared and especially the number of type of collocates was limited.

Table 5. Number of collocations appearing in the TIME corpus

	No.
Noun (type)	466
Verb (type)	180
TOTAL COLLOCATIONS	581(/1572)

A few collocates were used quite often among the collocations appearing in the TIME corpus (see Table 6). It was found that *make* was the most frequently used collocate which occurs as collocation and *take* was the second most frequently used one, which was much more often used than other collocates.

Table 6. High-frequency collocates in collocations in the TIME corpus (five times or more)

Frequency	Collocates
75	make
60	take
24	give
19	get
12	do
11	win
10	pay
9	hold, keep, play
8	have
7	cause, use
6	provide, set
5	commit, conduct, express, find, follow, lose, open, raise, send, show, suffer, tell, write

Table 7. High-frequency collocations in the TIME corpus (10 times or more)

Frequency of collocations	Collocations
45	do thing
39	play role
24	do job
23	do work
21	find way
19	have sex, tell story, have trouble
18	ask question
17	take place
16	make decision, make sense
15	pay attention, have effect
14	lose job, make mistake, write song, take time, fight war
13	play game, raise money
12	take care, make choice, open door
11	use force, answer question, take step
10	send message, make money, make movie, take risk

Table 7 shows collocations which were used 10 times or more in the TIME corpus. They indicate that the number of high-frequency collocations is limited, in fact, only 31 collocations were used more than 10 times in the TIME corpus.

Table 8. Level of the nodes and collocates per high-frequency collocations in the TIME corpus

		L1	L2	L3	L4	L5	L6	L7	L8	L1+2
10 times or more	No.	**30**	**1**	0	0	0	0	0	0	**31.0**
	%	**96.8**	**3.2**	0.0	0.0	0.0	0.0	0.0	0.0	**100.0**
9 times or more	No.	**34**	**1**	0	0	0	0	0	0	**35.0**
	%	**97.1**	**2.9**	0.0	0.0	0.0	0.0	0.0	0.0	**100.0**
8 times or more	No.	**41**	**1**	0	0	0	0	0	0	**42.0**
	%	**97.6**	**2.4**	0.0	0.0	0.0	0.0	0.0	0.0	**100.0**
7 or more time	No.	**43**	**2**	0	0	0	0	0	0	**45.0**
	%	**95.6**	**4.4**	0.0	0.0	0.0	0.0	0.0	0.0	**100.0**
6 times or more	No.	**49**	**4**	0	0	1	0	0	0	**53.0**
	%	**90.7**	**7.4**	0.0	0.0	1.9	0.0	0.0	0.0	**98.1**
5 times or more	No.	**61**	**8**	1	1	1	0	1	0	**69.0**
	%	**83.6**	**11.0**	1.4	1.4	1.4	0.0	1.4	0.0	**94.5**
4 times or more	No.	**81**	**17**	1	2	1	2	1	0	**98.0**
	%	**77.1**	**16.2**	1.0	1.9	1.0	1.9	1.0	0.0	**93.3**
3 times or more	No.	**113**	**33**	4	9	1	3	1	0	**146.0**
	%	**68.9**	**20.1**	2.4	5.5	0.6	1.8	0.6	0.0	**89.0**
2 times or more	No.	**148**	**77**	13	21	2	7	3	0	**225.0**
	%	**54.6**	**28.4**	4.8	7.7	0.7	2.6	1.1	0.0	**83.0**

L = Level

A look at the frequency of collocations appearing in the TIME corpus and the level of the make-up of nouns and verbs revealed the importance of L1 and L2 words in collocations occurring in the TIME corpus (see Table 8). Table 8 shows that 95.6% of collocations appearing more than seven times consisted of L1 verbs and nouns, while those which consist of both L1 and L2 verbs and nouns reached 100%. Thereafter, the percentage of the L1 ratio of collocations became lower, however, seen in the ratio of collocations which consist of both L1 and L2. The data reveal that more than 80% of the collocations were made up of basic and simple-leveled words.

4.3. Features of collocations in the BNC and the TIME corpus

4.3.1. Which levels of words are included in the high-frequency verb-noun collocations?

A comparison of high-frequency collocations in the BNC and the TIME corpus brought two findings. The first is that high-frequency collocations were common in both corpora. Table 9 shows collocations which are frequently used in the TIME corpus and which are ranked within 100 in the BNC. Among 31 collocations which occurred 10 times or more in the TIME corpus, 25 collocations belonged to the 100 most frequent

collocations in the BNC. Thus, common high-frequency collocations in the BNC and the TIME corpus were ranked within 100 in this analysis.

Table 9. Rank of high-frequency collocations in the BNC and the TIME corpus

Nodes	Collocates	Level(N+C)	**R. in BNC**	F. in BNC	Z-score	**R. in TIME**	F. in TIME
place	take	L1 + L1	**1**	12027	413.3	**10**	17
thing	do	L1 + L1	**2**	9961	116.7	**1**	45
effect	have	L1 + L1	**3**	7222	61.1	**13**	15
work	do	L1 + L1	**4**	5164	44.4	**4**	23
time	take	L1 + L1	**5**	4669	41.0	**15**	14
decision	make	L1 + L1	**6**	4451	198.2	**11**	16
job	do	L1 + L1	**7**	4330	78.6	**3**	24
question	ask	L1 + L1	**8**	4248	302.9	**9**	18
door	open	L1 + L1	**11**	3560	492.6	**22**	12
role	play	L1 + L1	**12**	3355	412.9	**2**	39
sense	make	L1 + L1	**17**	2818	124.3	**11**	16
way	find	L1 + L1	**19**	2742	64.1	**5**	21
step	take	L1 + L1	**21**	2643	177.7	**25**	11
care	take	L1 + L1	**23**	2609	140.0	**22**	12
question	answer	L1 + L1	**24**	2598	463.6	**25**	11
story	tell	L1 + L1	**28**	2054	188.1	**6**	19
mistake	make	L1 + L1	**31**	1968	199.6	**15**	14
game	play	L1 + L1	**32**	1956	237.9	**20**	13
trouble	have	L1 + L1	**34**	1891	27.6	**6**	19
attention	pay	L1 + L1	**43**	1707	258.5	**13**	15
money	make	L1 + L1	**51**	1533	35.9	**28**	10
money	raise	L1 + L1	**69**	1093	136.9	**20**	13
choice	make	L1 + L1	**71**	1085	55.9	**22**	12
risk	take	L2 + L1	**95**	845	49.3	**28**	10
job	lose	L1 + L1	**97**	839	92.1	**15**	14

R = rank, F = frequency

Table 10. Rank of six high-frequency collocations in the TIME corpus except collocations shown in Table 9

Nodes	Collocates	Level(N+C)	R. in BNC	F. in BNC	Z-score	R. in TIME	F. in TIME
sex	have	L2 + L1	-	-	2.5	6	19
message	send	L1 + L1	175	548	128.1	28	10
war	fight	L1 + L1	275	363	67.9	15	14
song	write	L1 + L1	352	288	57.4	15	14
force	use	L1 + L1	352	288	27.5	25	11
movie	make	L1 + L1	590	154	15.2	28	10

R = rank, F = frequency

As is seen in Table 10, among six collocations which occurred 10 times or more in the TIME corpus, *have sex* is not regarded as a collocational combination in the BNC,

because the z-score is 2.5, which is under the *three* needed as a collocational combination. In the other six collocations, *send message* is ranked 175th, *fight war* 275th, *write song* 352nd, *use force* 352nd, and *make movie* 590th in the BNC.

The other finding is that high-frequency collocations comprised basic words. The first 100 high-frequency collocations in the BNC consisted of 78.0% with L1 node and collocate combinations in Table 4 and 90.7% of collocations appearing more than six times in the TIME corpus consisted of L1 verbs and nouns in Table 8. Thus, high-frequency collocations in the BNC and the TIME corpus comprised basic words.

4.3.2. Are high-frequency collocations topic-oriented?

In order to examine the contexts used in high-frequency collocations in the TIME corpus, the topics in which collocations appeared and their frequency were grouped into four types: Social Sciences, Science & Technology, Art & Entertainment and Others (essays & opinions) as in Table 11. Collocations which occurred 15 times or more, occurred in all topic types. Then as their frequency of occurrence became lower, a few collocations appeared in only one or two topic types. However, among the collocations occurring 10 times or more there were none which were used in only one topic type. High-frequency collocations tend to be used regardless of topics.

A possible explanation for the result is that high-frequency collocations consist of basic and simple-leveled words, which can be widely used for many kinds of topics. As Aizawa (2005) indicates, topic-oriented and technical words tend to appear after 4000 level words in accordance with *JACET 8000* (2003). As section 4.1. shows that high-frequency words consist of the first 1000 and 2000 words, the result that the basic collocations tend to be used regardless of topics is reasonable.

As for the collocations occurring 10 times or more, the following observations were made. First, constituents of collocations sharing the same semantic domain were treated in a different way. For example, in the business domain, *do job, lose job* and *do work* occurred in all four topic types, while *make money* did not appear in Art & Entertainment and *raise money* did not appear in Science & Technology and Art & Entertainment. In the war domain, *fight war* was used in all four topic types, while *use force* was not seen in Science & Technology and Art & Entertainment, and *take risk* was not found in Science & Technology. Second, collocations which were used in daily life did not appear in all the topic types. *Take care, open door* and *answer question*

occurred in two or three topic types. Especially, *open door* was used in two topic types: Science & Technology and Social Sciences, which were not related to a daily life. *Open door* may rather be used as metaphorical term. Last, *make movie* and *write song* seemed to be topic-oriented collocations. However, *make movie* was used only in Others and Art & Entertainment, while *write song* occurred even in Science & Technology, Social Sciences in addition to Art & Entertainment.

Table 11. Number of high-frequency collocations in the categorized four topics

Frequency of collocations	Covered topics	Collocations	No. of S&T	No. of SS	No. of O	No. of A&E
45	4	do thing	9	21	8	7
39	4	play role	17	12	6	4
24	4	do job	3	15	2	4
23	4	do work	3	15	3	2
21	4	find way	2	17	1	1
19	4	have sex	10	2	4	3
19	4	tell story	2	7	5	5
19	4	have trouble	4	10	3	2
18	4	ask question	1	12	4	1
17	4	take place	3	9	4	1
16	4	make decision	4	7	4	1
16	4	make sense	5	4	6	1
15	4	pay attention	1	9	3	2
15	4	have effect	7	2	2	4
14	4	lose job	1	9	3	1
14	4	make mistake	1	11	1	1
14	3	write song	1	2	**0**	11
14	4	take time	2	7	2	3
14	4	fight war	1	11	1	1
13	4	play game	1	6	3	3
13	2	raise money	**0**	11	2	**0**
12	3	take care	6	5	**0**	1
12	4	make choice	2	6	2	2
12	2	open door	1	11	**0**	**0**
11	2	use force	**0**	9	2	**0**
11	3	answer question	**0**	6	2	3
11	4	take step	3	4	3	1
10	4	send message	1	7	1	1
10	3	make money	2	6	2	**0**
10	2	make movie	**0**	**0**	3	7
10	3	take risk	**0**	5	3	2

S&T = Science & Technology, SS = Social Sciences, O = Others (essays & opinions), A&E = Art & Entertainment

4.4. Summary

The above findings were summarized in relation to the postulated research questions.

1. *What are high-frequency collocations in large corpora collected from native speakers of English?*

Based on the analyses of high-frequency collocations in the BNC and the TIME corpus, many high-frequency collocations overlapped within the rank 100 in both corpora, which can be interpreted as high-frequency collocations by native speakers of English. Among 31 collocations which occur more than 10 times in the TIME corpus, 25 were also ranked within 100 in the BNC. The extremely frequent collocations were in the order of frequency: *take place*, *do thing*, *have effect*, *do work*, *take time*, *make decision*, *do job*, *ask question*, *open door*, *play role*, *make sense*, *find way*, *take step*, *take care*, *answer question*, *tell story*, *make mistake*, *play game*, *have trouble*, *pay attention*, *make money*, *raise money*, *make choice*, *take risk* and *lose job*.

2. *What are features of those high-frequency collocations by native speakers of English?*

2a. *Which levels of words are included in the high-frequency verb-noun collocations, in the word list of basic words for Japanese learners of English?*

In addition to common collocations in the BNC and the TIME corpus referred to in research question 1, it was found that high-frequency collocations consisted of basic verbs and nouns as a result of the analyses of the BNC and the TIME corpus. This was seen among the 25 extremely high-frequency collocations. *Take risk* was the only one collocation which consisted of an L2 node and an L1 collocate.

2b*. Are high-frequency collocations of native-speaker English related to topics?*

The analysis of high-frequency collocations and the topic types where they occurred in the TIME corpus indicated that more than 15 time collocations which appeared occurred in all four topics set for this research. Therefore, it can be said that high-frequency collocations are used regardless of topics. Since the analysis is a small scale, however, more research is needed to solidify this conclusion.

4.5. Discussion

There are three interesting points arising from the corpus data of verb-noun collocations in the BNC and the TIME corpus.

First, high-frequency collocations were high ranked in both the BNC and the TIME corpus. This was contrary to the present writer's expectation because the sources of these two corpora were different: the BNC is extracted samples of British English while the TIME corpus is extracted samples of mainly North American English. The total tokens and types were also different: about 100 million tokens were in the BNC and about 453 thousand tokens were in the TIME corpus.

Second, high-frequency collocations consisted of basic-level words, according to the results of analyzed data extracted from the BNC and the TIME corpus. In fact, collocations composing L1 and L2 verbs and nouns made up around 85% of all the occurring collocations in the TIME corpus. In the BNC, the coverage of L1 verb-noun collocations reached 78%, and the coverage of L1 and L2 verb-noun collocations reached 98% in the first 100 high-frequency collocations. These findings were desirable for Japanese upper secondary school students because they are expected to develop their four skills comprehensively using textbooks with a very limited number of vocabulary. Thirteen hundred words are targeted for 10th graders, but as they are calculated in the word-form system in which headwords, inflectional forms, reduced forms and derivative forms are respectively counted, these 1300 words will be in fact smaller in number, if they are calculated as one word.

Third, among collocations extracted from the TIME corpus, some which are related to specific topic types and tend to be ranked lower in the BNC, although they occurred 10 times or more in the TIME corpus. For example, among six collocations which occur 10 times or more in the TIME corpus, but which are not ranked within the 100 in the BNC *fight war* (275th) and *use force* (352nd) are on the Iraqi issues and *write song* (352nd) and *make movie* (590th) are on entertainment. TIME American version tends to reflect current domestic issues such as presidential election and the US related issues such as Iraqi war. They may have ranked lower in the more general corpus.

5. Conclusion

In this paper high-frequency collocations used by native speakers of English were

examined in the runup to selection of basic collocations for Japanese learners of English. This examination is indispensable because one's collocational competence is best reflected in its native speaker's ability in establishing or confirming rules of the grammar and the usage of language (Crystal, 1992). Moreover, Leech et al. (2001), Nation (2001), Schmitt (2000) and Schmitt & McCarthy (1997) point out that frequency is the main criterion in analyses of collocations in corpus study. The findings showed that high-frequency collocations in the BNC and the TIME corpora were fairly common, consisted of basic verbs and nouns in reference to *JACET 8000* and tended to be used regardless of the topics.

However, native speaker's high-frequency collocations are not necessarily equal to what is expected as the basic collocations for Japanese learners of English. It is because general corpora such as the BNC tend to (a) lack daily words which are popular in lower and upper secondary English textbooks, (b) choose words related to current affairs such as political issues and economic issues, many vulgar words, and slang words, and (c) disregard words for the beginning level. Therefore, native speaker's high-frequency collocations cannot be adopted as basic collocations of Japanese learners of English.

A pedagogical viewpoint should be taken to think what collocations are basic for Japanese learners of English who need to learn English for General Purposes (EGP). According to the government guidelines for foreign language teaching issued by MEXT (2003), they need to learn English for General Purposes (EGP) to develop basic English skills. Therefore, more words for the beginning level and more daily words which are popular in lower and upper secondary English textbooks should be taken into consideration for basic collocations for Japanese learners of English. Leech, Rayson and Wilson (2001) and the editors of *JACET 8000* support this pedagogical viewpoint for Japanese learners of English and emphasize both the use of frequency data as the reality of collocation use by native speakers of English and the selection from the frequency-collocations of native speakers of English for educational purposes. Based on these viewpoints, basic collocations should be identified scientifically and educationally for Japanese learners of English.

Notes

This paper is partially based on the Chapters 5 and 6 of the author's Ph.D dissertation submitted to the Graduate School of Education, Waseda University, in 2005, under the title "The Acquisition of Basic Collocations by Japanese Learners of English."

* The reason these four collocation dictionaries were used in this analysis is that they were the most representative collocation dictionaries, whether they were corpus-based or non corpus-based dictionaries. *COBUILD English Collocations on CD-ROM* (1995) and *Oxford Collocations Dictionary for Students of English* (2002) are corpus-based dictionaries, with examples taken from the *Bank of English*, which shows high frequent word combinations used in the daily life of native speakers of English. *The BBI Dictionary of English Word Combinations* (1997) is however, based on the native speakers' intuition, and is not corpus-based. *The Kenkyusha Dictionary of English Collocations* (1995) has been one of the major collocation dictionaries in Japan since it was first published in 1939 and the present edition contains 380,000 word combinations. Thus, these four dictionaries were used in order to select well-balanced collocations based on both corpus and the intuition of native speakers of English.

References

Aizawa, K. (2005, December). *Mind the gap: Discrepancies between difficulty of words and their frequencies.* Presentation given at JACET English vocabulary group second conference on researching, learning and teaching second language vocabulary, Tokyo, Japan.

Anderson, J. R. (1985). *Cognitive psychology and its implications*. New York: W. H. Freeman.

Alexander, R. J. (1984). Fixed expressions in English: Reference books and the teacher. *ELT Journal, 38*(2), 127-134.

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System. 21*(1), 101-114.

Barnbrook, G. (1996). *Language and computers: A practical introduction to the computer analysis of language.* Edinburgh: Edinburgh University Press.

Benson, M., Benson, E., & Ilson, R. (1997). *The BBI dictionary of English word combinations.* Amsterdam: John Benjamins.

Berry-Rogghe, G. L. M. (1973). The Computation of collocations and their relevance in lexical studies. In A. J. Aitken, R. W. Bailey, & N. Hamilton-Smith (Eds.), *The Computer and Literary Studies* (pp. 103-112). Edinburgh: Edinburgh University Press.

Caroli, M. T. (1998). *Relating collocations to foreign language learning.* Unpublished master's thesis. University of Reading, Reading, United Kingdom.

Crystal, D. (1992). *A dictionary of linguistics and phonetics*. London: Andre Deutsch.

Ellis, N. C. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33-68). Cambridge: Cambridge University Press.

Hill, J. (2000). Revising priorities: from grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching collocation* (pp. 28-46). Hove: Language Teaching Publications.

Ichikawa, H. et al. (Eds). (1995). *The Kenkyusha dictionary of English collocations* (2nd ed.). Tokyo: Kenkyusha.

Jones, S., & Sinclair, J. McH. (1974). English lexical collocations. *Cahiers de Lexicologie, 24*, 15-61.

Korosadowicz-Struzynska, M. (1980). Word collocations in FL vocabulary instruction. *Studia Anglica Posnaniensia*, *12*, 109-120.

Koya, T. (2004). Collocation research based on corpora collected from high school English textbooks in Japan. In Y. Watanabe, I. Nagano, & A. Morita (Eds.), *Collection of papers in honor of Professor Yoshiaki Shinoda*, (pp. 99-113). Tokyo: Nanundo.

Koya, T. (2005) *The Acquisition of Basic Collocations by the Japanese Learners of English.* A dissertation presented to the Graduate School of Waseda University.

Lea, D. et al. (Eds). (2002). *The Oxford collocations dictionary for students of English.* Oxford: Oxford University Press.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English*. Harlow: Longman.

Lewis, M. (Ed.). (1993). *The lexical approach.* Hove: Language Teaching Publications.

Lewis, M. (Ed). (2002). *Implementing the lexical approach.* Hove: Language Teaching Publications.

Lewis, M. (Ed.). (2000). *Teaching collocation.* Hove: Language Teaching Publications.

McCarthy, M. (1984). A new look at vocabulary in EFL. *Applied Linguistics*, *5*(1), 12-22.

McCarthy, M. (2004, August). *Collocation in vocabulary teaching and learning*. Lecture given at the meeting of JACET summer seminar program, Gunma, Japan.

Murata, M. et al. (Eds.). (2003). *The JACET list of 8000 basic words*. Tokyo: The Japan Association of College English Teachers.

Nattinger, J., & 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*(2), 223-242.

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

Sinclair, J. McH. et al. (Eds). (1995). *The COBUILD English collocations on CD-ROM.* London: Harper Collins.

Yorio, C. A. (1980). Conventionalized language forms and the development of h communicative competence. *TESOL Quarterly, 14*(4), 433-442.