Research Note

日本大学生産工学部研究報告B
2015 年 6 月 第 48 巻

Examining Corpus-based L2 Vocabulary Lists for Grade Level and Semantic Field Distribution

*Kiyomi CHUJO** *and Kathryn OGHIGIAN***

(Received February 6, 2015)

Abstract

Since classroom time is limited, identifying and prioritizing relevant target vocabulary is important. In Japan, four corpus-based high frequency vocabulary lists often used as core vocabulary sources for second language (L2) learners are the JACET List of 8,000 Basic Words, the Standard Vocabulary List, the BNC High Frequency Word List, and the 5,000 most frequently used words in the Corpus of Contemporary American English (COCA). This study explores how adequately the vocabulary of these lists was defined in terms of grade level and semantic category distribution. It was found that the selected words of each vocabulary list were at the appropriate grade level, however the semantic categories showed a marked tendency toward more adult concepts. It was also found that the addition of the COCA thematic vocabulary to the COCA high frequency list could complement the deficiency in semantic fields relevant to the developmental level of the students.

Keywords: Vocabulary List, Corpus-based, Thematic Vocabulary, Grade Level, Semantic Field Distribution

1. Literature Review

1.1 Vocabulary Lists Commonly Used in Japan in Second Language Learning

Thorndike and Lorge (1944)[1] and the General Service List (West, 1953)[2] have historically been used as the basis for major guidelines for compiling Japanese textbooks in secondary school systems (Ito, 1977)[3] and for reading materials such as graded readers for learners of English as a foreign or second language (Nation, 2004)[4]. These have gradually been replaced by corpus-based word lists developed from the British National Corpus (BNC)[5] such as the JACET List of 8,000 Basic Words (hereafter JACET) (JACET, 2003)[6], the Standard Vocabulary List (SVL) (ALC, 2001)[7], and the BNC High Frequency Word List (BNC HFWL) (Chujo, 2004)[8]. In addition to these, the 5,000 most frequently used words in the Corpus of Contemporary American English (COCA) became available in 2010 (Davies & Gardner, 2010)[9].

1.2 Evaluating Vocabulary Lists

One way to evaluate a vocabulary list is to measure text coverage, that is, to determine to what extent the vocabulary "covers" or includes the number of known words in a text. Meaningful input is generally defined at 95% coverage (Laufer, 1989)[10], or ideally, at 98% (Hu & Nation, 2000)[11]. In other words, a selected vocabulary list could be considered adequate if, once acquired, the reader is able to understand 95 to 98 words out of every 100. Text coverage is calculated by counting the number of the words known in the text, multiplying this number by 100 and then dividing by the number of tokens (total number of words) in the text (Chujo & Utiyama, 2005)[12]. Text coverage is based on frequency, i. e., the idea that word lists based on more frequently appearing words will provide more coverage. This idea has been used, for example, in Thorndike and Lorge (1944)[13] and Nation's fourteen 1,000-word-family lists (2006)[14]. However, there have been criticisms of high frequency word lists. Mackey (1965:183)[15] noted that "[e]ven though blackboard may not

*Professor, Department of Liberal Arts and Basic Sciences, College of Industrial Technology, Nihon University
**Adjunct Lecturer, Faculty of Science and Engineering, Waseda University

be a very frequent word elsewhere, it is a necessary word in the classroom" and "[s]uch words constitute the thematic vocabulary available for certain situations." Richards (1970:88)[16] questioned the usefulness of word-frequency lists such as those of Thorndike and Lorge because they did not include "soap, bath, cushion, chalk and stomach" in the first 2,000 words. Ishikawa (2005:44)[17] demonstrated that in the BNC, words such as "notebook, eraser, blackboard, pocket and chime" have a low frequency but are familiar to Japanese schoolchildren, and he concluded that the high frequency words derived from the BNC are weak in identifying familiar everyday vocabulary for children. In fact, the BNC has been shown to be inappropriate for using unchanged as the basis for syllabus design for EFL or ESL learners in primary or secondary schools because "[t]he BNC is predominantly a corpus of British, adult, formal, informative language, and most English learners in primary and secondary school systems are not British, are children, and need both formal and informal language for both social and informative purposes" (Nation, 2004:3-4)[18].

Considering these criticisms, another way to objectively examine the appropriateness or inappropriateness of selected word lists is to investigate at what grade level these words would be understood. Chujo and Utiyama (2006)[19] used the Living Word Vocabulary (Dale & O'Rourke, 1981)[20] list to determine the grade level at which the central meaning of words from a corpus-based list such as the BNC could be readily understood. Chujo, Oghigian, Utiyama, and Nishigaki (2011)[21] used the Dale and O'Rourke list to evaluate corpus-based selected daily life vocabulary for elementary students from a corpus such as the Child Language Data Exchange System (CHILDES)[22]. Another method is to determine if the word lists include grade-appropriate concepts. Chujo et al. (2011)[23] also used the Longman Lexicon of Contemporary English (McArthur, 1981)[24] to examine the selected corpus-based daily life vocabulary for elementary students mentioned above. This resource classifies over 15,000 entries under a set of fourteen semantic fields such as life and living things, and people and the family.

2. Purpose of the Study

General trends in second language education are in using corpus-based vocabulary lists (Davies & Gardner, 2011)[25]. The purpose of this paper is to determine the appropriateness of the four corpus-based lists used in Japan (JACET, SVL, BNC HFWL and COCA). The specific research questions are as follows. 1) At what U. S. grade level are the selected words of each vocabulary list understood? Are they properly graded? 2) What semantic categories are represented, and how are these distributed? 3) What pedagogical applications are suggested by the results? The four lists are described in detail in the next section. In the Method section, a description is given on how the examined words were organized to allow comparisons, and the evaluation is described. The following section presents the results and discussion, and the final section provides the conclusions.

3. Four Vocabulary Lists

The four vocabulary lists were selected for the following reasons: (1) they were based on large-scale electronically-accessible corpora; (2) they were built in the 2000s; (3) they were compiled considering the educational purpose to a certain extent (i. e., for language learning rather than lexicography or translation); and (4) they were available and are currently used.

3.1 The JACET List of 8000 Basic Words (JACET)

JACET stands for the Japan Association of College English Teachers and the JACET word list contains the 8,000 basic words "designed for all English learners in Japan" in accordance with the frequency and the educational significance of each word (Uemura & Ishikawa, 2004)[26]. It is based on the BNC, and the JACET sub-corpus of approximately six million words is from American newspapers, magazines, TV programs, children's literature, Japanese high school English textbooks, and various English tests administered in Japan.

3.2 The Standard Vocabulary List 12000 (SVL)

The SVL is a list of 12,000 words specifically developed for Japanese learners of English by the publisher ALC. They emphasize high-frequency words for both native speakers' usefulness and their importance for Japanese learners. The SVL is based on various word lists and corpora including the BNC, along with a special consideration for Japanese learners of English. There are 12 levels of 1,000 words.

3.3 The British National Corpus High Frequency Word List (BNC HFWL)

The BNC HFWL is a list of 13,994 lemmatized words representing 86 million BNC words that occur 100 times or more (Chujo, 2004)[27]. It was created by: (a) using the CLAWS7 tag set to extract all base forms; (b) lemmatizing by inflectional form; (c) deleting any low frequency or unusual words (those appearing fewer than 100 times in this lemmatized list); and (d) identifying all proper nouns and numerals by their part of speech tags and deleting manually.

This vocabulary list was used in Chujo and Utiyama (2005, 2006)[28), 29)], and Chujo et al. (2011)[30)] as a reference list for extracting specialized words.

3.4 The Top 5,000 Lemmas in the Corpus of Contemporary American English (COCA)

The Corpus of Contemporary American English (COCA) is the largest freely-available corpus of American English. As of 2014, it contains more than 450 million words of text and is organized into spoken, fiction, popular magazines, newspapers, and academic text registers. It includes 20 million words each year from 1990-2012 and the corpus has been updated regularly. From this corpus, A Frequency Dictionary of Contemporary American English (Davies & Gardner, 2010)[31)] was published and from that, the top 5,000 lemmas were selected by taking into account both frequency and dispersion. In this paper, these 5,000 words will be referred to as "COCA" hereafter.

4. Method

4.1 Procedure for Reorganizing the Entry Data of Four Vocabulary Lists

Each examined vocabulary list used a slightly different notion of the concept of "words." For example, JACET included abbreviations such as ed. and etc.; the SVL included some proper nouns and numerals; and in the COCA, each part of speech was listed as a different word. In order to make the various entries of these lists comparable, they were reorganized using the BNC HFWL as a reference. In other words, in order to be comparable with the BNC HFWL, abbreviations were deleted from JACET; proper nouns and numerals were deleted from the SVL; and each part of speech variation in the COCA was listed as the same base word. The numbers of lemmas were decreased for each list as a result of these exclusions, but all four lists presented words (lemmas) in the same format.

Next, lists of lemmas were developed from each source that were organized alphabetically into groups of 1,000 lemmas: seven 1,000-lemma groups from the JACET, eleven 1,000-lemma groups from the SVL, four 1,000-lemma groups from COCA, and thirteen 1,000-lemma groups from the BNC. These lists are referred to hereafter as 1,000-lemma groups.

4.2 Evaluating the Four Vocabulary Lists

In order to determine the pedagogical appropriateness for the four vocabulary lists, the words in each 1,000-lemma group from each vocabulary list were evaluated with regard to grade level, and semantic content and distribution. These procedures are detailed below.

4.2.1 Determining the grade level

To understand at what grade level these words would be understood by American native English speaking (NS) children, each 1,000-lemma group from the four vocabulary lists was compared to The Living Word Vocabulary (LWV) (Dale & O'Rourke, 1981)[32)] and The Basic Elementary Reading Vocabularies (Harris & Jacobson, 1972)[33)]. The Living Word Vocabulary includes more than 44,000 items and each presents a percentage score for those words or terms familiar to American students in grade levels 4, 6, 8, 10, 12, 13, and 16. (Grades 13 and 16 correspond to the university level.) The Basic Elementary Reading Vocabularies has 7,613 different words appearing in a selection of textbooks widely used in 1970 in grades one through six of the elementary school. This was used for determining the (U. S.) grade levels of reading vocabulary for the first, second and third grade levels. Using these control lists, we calculated the average grade level for each 1,000-lemma group. It should be noted that although the LWV is dated, it is the only such database available. For a more detailed justification of using this resource, see Hiebert (2005: 252-253)[34)] or Chujo et al. (2011)[35)].

4.2.2 Determining the semantic categories

Tom McArthur's Longman Lexicon of Contemporary English (1981)[36)] classifies over 15,000 entries under a set of fourteen semantic fields. These fourteen categories were used in this study to cluster words from the four vocabulary lists into groups so that semantic distribution could be compared. Some polysemous words, for example nail, were assigned to two semantic fields: "the body" and "substances, materials, objects, and equipment." Therefore the total number of semantic fields is larger than the number of words.

4.2.3 Determining the usefulness of supplementing the thematic list

In addition to the COCA 5,000 words, a second vocabulary list from COCA, which was not included in the 5,000 COCA vocabulary list, contains "31 thematic boxes" (Davies & Gardner, 2010)[37)] on various topics. From these, 943 words on 13 topics such as animals, body, clothing, colors, emotions, family, foods, materials, professions, sport and recreation, time, transportation, and weather were selected and compiled into a supplemental COCA thematic list (hereafter, COCA+). The semantic categories chosen were the same as the categories used previously (McArthur, 1981)[38)]. In order to see how well this supplemental COCA+ thematic vocabulary covered various activities if this was added to the main COCA 5,000 list, the distribution of the COCA+ sematic fields were compared to those of the main COCA list.

5. Results and Discussion

5.1 Evaluating Grade Level

The results of the comparison of the grade levels of the four vocabulary lists are shown in **Table 1**. The numbers indicate at what U. S. grade level the majority of NS students would readily understand the central meaning of each word in each of the 1,000-lemma groups.

A clear tendency for a steady increase in grade level corresponding to the lemma groups can be seen. The first 1,000 lemmas are generally understood by third grade students, although the words on SVL are aimed more at second grade students. The second 1,000 lemmas are generally understood by fourth or fifth grade students, and the third 1,000 lemmas are generally known by fifth or sixth grade students. The levels increase gradually: words from the fourth to seventh 1,000 lemma strata are generally known by seventh or eighth grade students, and from the eighth to 13th lemma strata by U. S. high school students.

This procedure identified an optimal number of words for a large working vocabulary list. In terms of practical application, the first and second lemma groups corresponded to the U. S. elementary school level, the next 3,000 to 7,000 lemmas were in line with the U. S. middle school level, and the 7,000 to 8,000 vocabulary was at the U. S. high school level. Interestingly, the JACET grade level increased steadily from third to fourth to fifth in the first 3,000 lemmas, but at 4,000, jumps to the eighth grade. Based on these results, a learner using the JACET material might have difficulty making a transition from fifth grade words to eighth grade words without supplemental vocabulary. In contrast, the SVL started from the second grade and increased reasonably to the middle school level and high school level. In each case, the difficulty level of the SVL was slightly easier compared to the other groups. As a matter of interest, Chujo, Nishigaki, Hasegawa, and Utiyama (2008: 63)[39] evaluated the Japanese junior high school English textbook levels as U. S. grade 2.6 and the Japanese senior high school English textbooks as U. S. grade 4.1.

5.2 Evaluating Semantic Content and Distribution

The distribution of semantic fields for the top 1,000, 2,000, 3,000 and 4,000 lemmas of the four vocabulary lists are shown in **Table 2** according to the McArthur's (1981)[40] order of the fourteen semantic fields. The numbers indicate what percentage each of the lemma groups include entries belonging to each of the fourteen semantic fields.

It can be seen that the semantic fields containing the majority of these list words were: (a) "general and abstracts terms" such as *fact, event, risk,* and *matter*; (b) "thought and communication, language, and grammar" such as *mind, reason, analysis*, and *memory*; (c) "people and the family" such as *human, person, individual*, and *friend*; (d) "space and time" such as *world, space, history*, and *moment*; (e) "movement, location, travel, and transport" such as *moment, approach, remain*, and *arrive*; and (f) "numbers, measurement, money, and commerce" such as *figure, average, measure*, and *capacity*. On the other hand, the semantic fields containing the fewest of the four vocabularies were: (a) "food, drink, and farming"

Table 1 Each 1,000-lemma Group for the Four Vocabulary Lists and Average Grade Level

1,000-lemma Group	COCA	JACET	SVL	BNC HFWL
1 (1 - 1,000)	3	3	2	3
2 (1,001 - 2,000)	5	4	4	5
3 (2,001 - 3,000)	6	5	5	6
4 (3,001 - 4,000)	7	8	6	7
5 (4,001 - 5,000)		8	7	8
6 (5,001 - 6,000)		8	7	8
7 (6,001 - 7,000)		8	8	9
8 (7,001 - 8,000)			9	10
9 (8,001 - 9,000)			9	10
10 (9,001 - 10,000)			11	11
11 (10,001-11,000)			11	11
12 (11,001-12,000)				11
13 (12,001-13,000)				12

Table 2 A Comparison of Percentage of the Top 1,000, 2,000, 3,000, 4,000 Lemmas from Each Word List by Semantic Field

Semantic Field	Top 1,000 (1 - 1,000)				Top 2,000 (1 - 2,000)				Top 3,000 (1 - 3,000)				Top 4,000 (1 - 4,000)			
	COCA	JACET	SVL	BNC	COCA	JACET	SVL	BNC	COCA	JACET	SVL	BNC	COCA	JACET	SVL	BNC
life & living things	2.5	2.9	4.7	2.4	2.9	3.1	4.3	2.7	3.1	3.8	4.4	2.8	3.4	3.5	4.2	3.0
body	4.0	4.1	5.0	3.7	4.2	4.5	4.8	4.0	4.4	4.5	4.9	4.1	4.6	4.5	4.6	4.6
people & the family	11.4	9.6	7.6	10.6	11.1	10.2	9.0	10.7	11.6	10.1	9.6	11.0	11.6	11.2	9.8	11.6
buildings, houses, the home, clothes, belongings, and personal care	4.4	4.5	5.9	4.4	5.0	4.9	5.8	4.7	5.2	5.3	5.9	5.3	5.2	5.2	5.9	5.2
food, drink, and farming	1.9	2.5	5.5	1.8	3.0	3.1	4.5	2.5	3.3	3.4	4.3	2.9	3.4	3.1	4.3	3.0
feelings, emotions, attitudes, and sensations	5.3	7.0	6.2	6.0	6.1	7.2	6.5	6.4	6.9	7.6	7.2	6.9	7.1	7.3	7.8	7.4
thought & communication, language & grammar	12.5	12.4	9.5	13.1	11.9	11.5	10.4	12.0	11.6	11.2	10.4	11.9	11.2	11.6	10.6	11.7
substance, materials, objects, & equipment	4.6	4.2	5.6	4.3	5.6	6.0	7.0	5.7	6.4	6.8	7.2	6.3	6.8	6.7	7.0	6.6
arts & crafts, science & technology, industry & education	4.3	3.8	3.0	4.2	4.1	3.9	3.6	4.4	4.0	4.0	3.7	4.3	4.2	4.1	4.0	4.2
numbers, measurement, money, & commerce	8.3	7.5	6.3	9.6	8.2	7.3	6.8	8.7	7.5	6.7	7.3	8.2	7.3	7.6	7.1	7.6
entertainment, sports, & games	7.2	7.1	8.1	6.7	6.6	6.7	7.5	6.3	6.3	6.8	7.1	6.0	6.1	6.0	6.6	5.9
space & time	9.7	10.2	11.6	8.6	8.3	8.6	8.7	7.9	7.7	8.1	7.8	7.6	7.6	7.7	7.7	7.6
movement, location, travel, & transport	9.5	9.7	10.8	8.8	8.9	9.1	10.0	9.1	8.3	8.7	8.7	8.7	8.2	8.2	8.5	8.2
general & abstracts terms	14.4	14.4	10.0	15.8	14.1	14.0	11.0	14.8	13.7	12.9	11.5	14.1	13.2	13.3	12.0	13.7
SD	3.9	3.7	2.5	4.1	3.4	3.3	2.4	3.6	3.3	2.9	2.4	3.4	3.1	3.1	2.5	3.3

such as *food*, *water, chicken*, and *oil*; (b) "life and living things" such as *animal*, *bird*, *dog*, and *cat*; (c) "body" such as *arm*, *hair*, *eye*, and *heart*; and (d) "arts and crafts, science and technology, industry and education" such as *make*, *produce*, *school*, and *classroom*. This indicates that the easiest vocabulary on the four lists generally relate to abstract concepts belonging to semantic fields appropriate to adults rather than school children or EFL/L2 learners focused on basic communication.

In Table 2, the figure in the bottom row shows the standard deviation (SD) among the percentage scores of each lemma group which can explicitly describe the degree of variability among the distribution of words belonging to the fourteen categories. Looking at the SDs of the top 1,000, 2,000, 3,000 and 4,000 lemmas of the SVL, they were 2.5, 2.4, 2.4, and 2.5, respectively, while those of other three lists varied widely from 2.9 to 4.1. This indicates that the semantic distribution of the SVL is more balanced and proportionate among the fourteen semantic fields than the other three lists.

Fig. 1 offers a visual representation of the distribution of semantic fields for the top 1,000-lemma groups of the four lists. The percentage of top 1,000 lemmas classified into each of the semantic fields is shown by a radar chart. For example, the round dots show the distribution from the SVL.

Looking at the radar graph, it can be seen that the zigzag lines of the top 1,000 lemmas of COCA, JACET and BNC lists fluctuate almost in unison, corresponding to the semantic fields, while that of the SVL, which is based on various word lists and corpora including the BNC, along with a special consideration for Japanese learners of English, has the smallest fluctuation. The fact that all three lines demonstrate the same pattern also indicates there is some correlation among the comparisons.

To see the relationship between the semantic category distributions between sets of two lists, Pearson's correlation was calculated. The four top 1,000-lemma lists highly correlated with each other. The values indicated a strong correlation between the COCA and the JACET ($r=.980$, $\rho=.0000$); the COCA and the SVL ($r=.793$, $p=.0007$); between the COCA and the BNC HFWL ($r=.985$, $\rho=.0000$); the JACET and the SVL ($r=.846$, $\rho=.0001$); between the JACET and the BNC HFWL ($r=.971$, $\rho=.0000$); and the SVL and the BNC HFWL ($r=.729$, $\rho=.003$). Note that all the p values for these correlations were less than 0.01, so the correlation was significant at the 1% significance level. It is not surprising that the COCA and the BNC HFWL have the highest correlation because they were created from the highest frequency words with no manual corrections. On the other hand, the SVL had less correlation with the other three lists and this could be attributed to the fact that it was largely changed from the original high frequency lists which tended to have markedly adults concepts to less adults concepts and more concepts appropriate developmental level of the students.

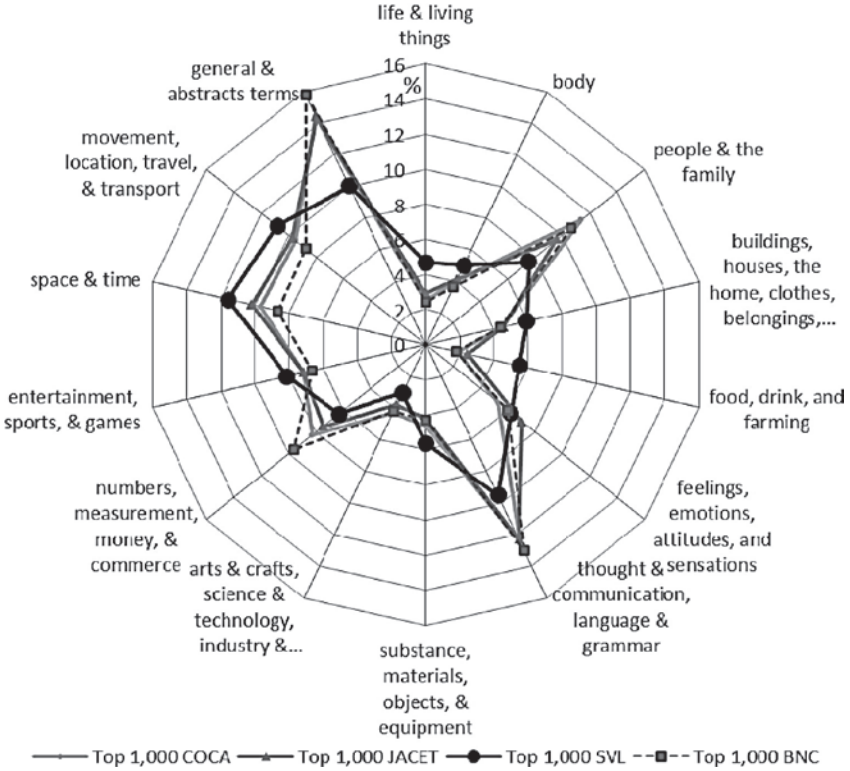

Fig. 1 A Comparison of Percentages of the Top 1,000 Lemmas of the Four Lists by Semantic Field

5.3 Evaluating the Usefulness of Supplementing the Thematic List

In order to determine the usefulness of supplementing the COCA+ thematic list to the main COCA list, a calculation on how the addition of this thematic vocabulary could supplement the semantic distribution of the concepts to the original word list was done and the results are shown in **Fig. 2**. The percentage of the original COCA 1,000 lemmas classified into each semantic field is shown by grey bars; and the percentage of COCA+ 1,000 lemmas supplemented by a thematic vocabulary list is shown with black bars. Whereas the original COCA 1,000 did not include concepts germane to young learners and second language students such as life and living things (*egg*, *cow*, *elephant*, *butterfly*, *mosquito*, *snake*, and *whale*), or food (*carrot*, *asparagus*, *mushroom*, *hamburger*, *sandwich*, *bread*, *butter*, *soup*, *pudding*, *dessert*, *lunch*, and *breakfast*), it can be seen from Figure 2 that these categories are supplemented by the thematic COCA+ vocabulary. In fact, the supplemental vocabulary applies to several categories that would be useful for L2 learners such as food, drink and farming (improved from 1.9% to 4.6%); the body (improved from 4.2% to 5.2%); buildings, houses, home and clothes (improved from 4.4% to 5.0%); substances and objects (improved from 4.6% to 5.6%); feelings and emotions (improved from 5.3% to 6.1%); and entertainment, sports and games (improved from 7.2% to 7.6%). The thematic vocabulary is an important supplement, although there would be benefit from further improvements.

6. Conclusion

In this study, four vocabulary lists used in Japan in second language learning were evaluated for grade level and semantic category. When the words from each list were organized into comparable lemma and sorted into 1,000-word high frequency lemma groups, there was a clear linear progression of grade level for the four vocabulary lists such that the first and second lemma groups corresponded to the U. S. elementary school level, the next 3,000 to 7,000 words corresponded with the U. S. middle school level, and the 7,000 to 8,000 vocabulary was at the U. S. high school level. In addition, it was noted that the words on the SVL list were slightly easier (i. e., understood by younger students in each group) and the words from JACET jumped remarkably from the fifth grade to the eighth grade, suggesting a supplemental list might be required for students using only this resource.

A comparison of semantic categories showed that the concepts represented by the vocabulary in all four vocabulary lists (but slightly less so in the SVL), were abstract and thus not as appropriate or beneficial to young or second language learners, if one accepts the idea that concrete items (*mother*, *dog*, *banana*, *tree*) are more readily understood by this population than abstract terms (*moment*, *fact*, *history*, *reason*, *remain*). It was also found that the addition of the COCA

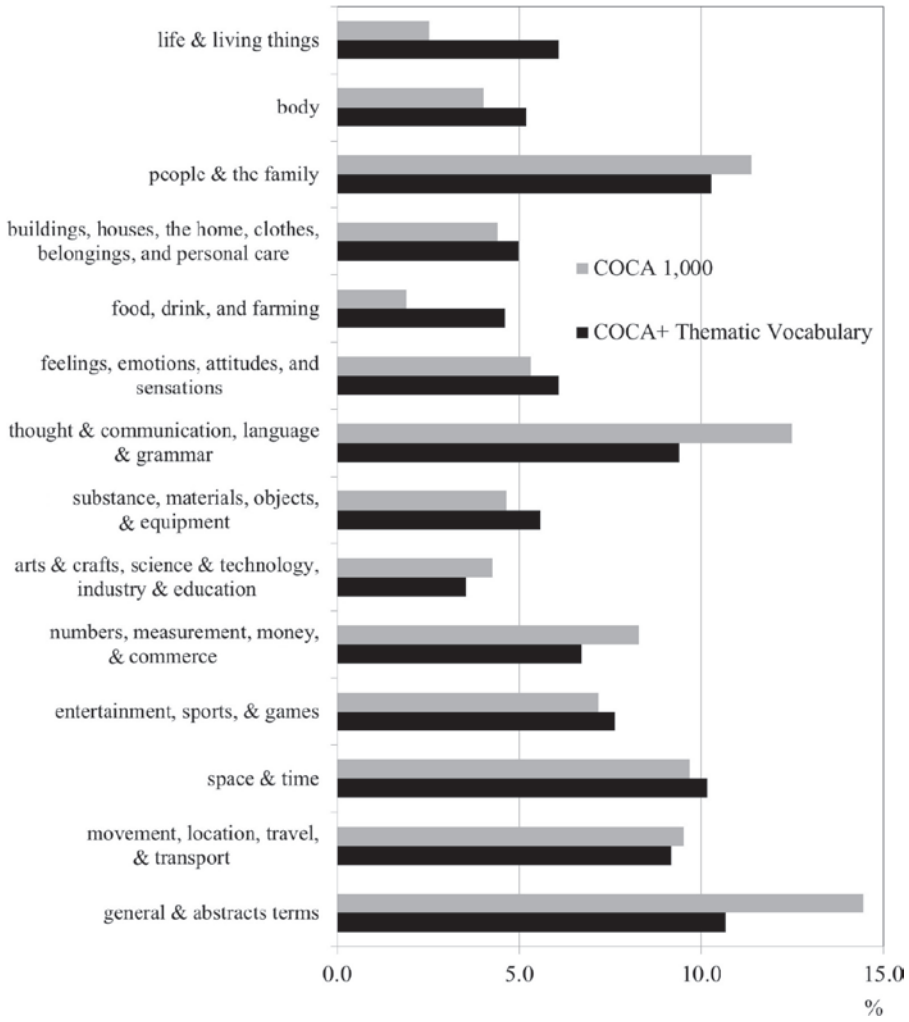

Fig. 2 A Comparison of Percentages of the Top 1,000 Lemmas of the COCA with/ without Thematic Vocabulary by Semantic Field

"thematic vocabulary" to the COCA high frequency list could complement the deficiency in semantic fields relevant to the developmental level of the students. It is hoped that the findings of this study will allow users of these vocabulary lists to be more aware of their applications and limitations.

Acknowledgements: Part of this research was funded by a Grant-in-aid for Scientific Research (21320107; 25284108) from the Japan Society for the Promotion of Science and the Ministry of Education, Science, Sports and Culture.

References

1) Thorndike, E. L., & Lorge, I. *The First 1,000 Words: The Teacher's Word Book of 30,000 Words*. New York: Bureau of Publications Teachers College, Columbia University, 1944.
2) West, M. *A General Service List of English Words*. London: Longman, Green & Co, 1953.
3) Ito, K. Monbushou Houmonki. [A Report from Visiting MEXT] *The English Teachers' Magazine,* 26, 6, 1977, 42.
4) Nation, P. A Study of the Most Frequent Word Families in the British National Corpus. In Bogaards, P. & Laufer, B. (eds.), *Vocabulary in a Second Language*. Amsterdam: John Benjamins Publishing Company, 2004, 3-13.
5) *British National Corpus*, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. 〈http://www.natcorp.ox.ac.uk/〉 (1 June 2014).
6) JACET (The Japan Association of College English Teachers) Kihongo Kaitei Committee. *JACET List of 8000 Basic Words*. Tokyo: JACET, 2003.
7) ALC. *Standard Vocabulary List* (SVL) 12000, 2001, URL:http://www.alc.co.jp/eng/vocab/svl/
8) Chujo, K. Measuring Vocabulary Levels of English Textbooks and Tests Using a BNC Lemmatized High Frequency Word List. In Nakamura, J., Inoue, N. & Tabata, T. (eds.), *English Corpora under Japanese Eyes*. Amsterdam: Rodopi, 2004, 231-249.
9) Davies, M., and Gardner, D. *A Frequency Dictionary of Contemporary American English: Word Sketches, Collocates, and Thematic Lists*. London and New York:

Routledge, 2010.
10) Laufer, B. What Percentage of Text Lexis Is Essential for Comprehension? In Lauren, C. & Nordman, M. (eds.), *Special Language: From Humans to Thinking Machines*. Clevedon: Multilingual Matters, 1989, 316-323.
11) Hu, M. and Nation P. Unknown Vocabulary Density and Reading Comprehension. *Reading in a Foreign Language*, 13, 1, 2000, 403-430.
12) Chujo, K., and Utiyama, M. Understanding the Role of Text Length, Sample Size and Vocabulary Size in Determining Text Coverage. *Reading in a Foreign Language*, 17, 1, 2005, 1-22.
13) Thorndike, E. L., and Lorge, I. (1944).
14) Nation, P. How Large a Vocabulary Is Needed for Reading and Listening? *The Canadian Modern Language Review*, 63, 1, 2006, 59-82.
15) Mackey, W. F. *Language Teaching Analysis*. London: Longmans, 1965.
16) Richards, J. C. A Psycholinguistic Measure of Vocabulary Selection. *IRAL*, 8, 2, 1970, 87-102.
17) Ishikawa, S. Frequency and Familiarity in Compiling the English Word List for Children. *IEICE Technical Report*, TL 25, 2005, 43-48.
18) Nation, P. (2004).
19) Chujo, K., and Utiyama, M. Selecting Level-specific Specialized Vocabulary Using Statistical Measures. *System*, 34, 2, 2006, 255-269.
20) Dale, E., and O'Rourke, J. *The Living Word Vocabulary*. Chicago: World Book-Childcraft International, Inc., 1981.
21) Chujo, K., Oghigian, K., Utiyama, M., and Nishigaki, C. Creating a Corpus-based Daily Life Vocabulary for TEYL. *Asian EFL Journal*, 49, 2011, 30-59.
22) Child Language Data Exchange System. (n. d.) URL: http://childes.psy.cmu.edu/
23) Chujo, K., Oghigian, K., Utiyama, M., and Nishigaki, C. (2011).
24) McArthur, T. *Longman Lexicon of Contemporary English*. Essex: Longman, 1981.
25) Davies, M., and Gardner, D. (2010).
26) Uemura T.,and Ishikawa, S. JACET 8000 and Asia TEFL Vocabulary Initiative. *Journal of Asia TEFL*, 1, 1, 2004, 333-347.
27) Chujo, K. (2004).
28) Chujo, K., and Utiyama, M. (2005).
29) Chujo, K., and Utiyama, M. (2006).
30) Chujo, K., Oghigian, K., Utiyama, M., and Nishigaki, C. (2011).
31) Davies, M., and Gardner, D. (2010).
32) Dale, E., and O'Rourke, J. (1981).
33) Harris, A. J., and Jacobson, M. D. *Basic Elementary Reading Vocabularies*. New York: The Macmillan Company, 1972.
34) Hiebert, E. H. In Pursuit of an Effective, Efficient Vocabulary Curriculum for Elementary Students. In Hiebert, E. H. & Kamil, M. L. (eds.), *Teaching and Learning Vocabulary.* Mahwah, NJ/London: Lawrence Erlbaum Associates, Publishers, 2005, 243-263.
35) Chujo, K., Oghigian, K., Utiyama, M., and Nishigaki, C. (2011).
36) McArthur, T. (1981).
37) Davies, M., and Gardner, D. A Frequency Dictionary of Contemporary American English. In Newman, J., Baayen, H. & Rice, S. (eds.), Corpus-based Studies in Language Use, *Language Learning, and Language Documentation*. Amsterdam/New York: Rodopi, 2011, 283-297.
38) McArthur, T. (1981).
39) Chujo, K., Nishigaki, C., Hasegawa, S., and Utiyama, M. Yutori Kyouiku Jidai no Koukou Eigo Kyoukasho wo Kangaeru: 1980 Nendai to 2000 Nendai no Koukou Kyoukasho Goi no Hikaku Bunseki kara no Kousatsu. [The Impact of Yutori Kyouiku: A Comparative Study of 1988 and 2006 High School Textbook Vocabulary], *English Corpus Studies*, 15, 2008, 57-79.
40) McArthur, T. (1981).

学年レベルと意味領域の分布に基づくコーパス準拠語彙リストの調査

中條清美，キャサリン・オヒガン

概　要

教育用基本語彙の一般的な作成方法には，大規模コーパスから出現頻度の高い語を選定するという方法が用いられる。本研究では，現在我が国で利用されている４種のコーパス準拠の大規模語彙リストの基本語彙としての有効性を「学年レベル」と「意味領域の分布」という２つの観点から調査し，比較した。調査した語彙リストは，JACET8000（JACET List of 8,000 Basic Word），SVL12000（Standard Vocabulary List 12000），BNCHFWL（British National Corpus High Frequency Word List），COCA5000（The 5,000 most frequently used words in the Corpus of Contemporary American English）である。結果，４種の語彙リストを構成する語彙の学年レベルは，ほぼ適切なものであった。一方，意味領域の分布については，大人向けの分野（抽象，人間，思考）で高く，子供向けの分野（生物，身体，飲食物）における割合が低いことが明らかになった。COCA5000 には 31 のテーマ別語彙リストが付属しており，これらを追加利用すれば，不足している子供向けの分野の語彙が補充可能であることを明らかにした。

キーワード：語彙リスト，コーパス準拠，テーマ別語彙，学年レベル，意味領域の分布