Practical English Studies, Vol.12, pp. 19-34, 2006.

Creating Spoken Academic Vocabulary Lists from the British National Corpus

Nihon University
Kiyomi Chujo
Chiba University
Chikako Nishigaki

English has increasingly become a global language of communication, trade and research. For ESL and EFL students, the need for English for academic purposes (EAP) skills — understanding academic textbooks and journal articles and attending or giving presentations — is also growing. Academic vocabulary development is essential in order to achieve a certain proficiency in EAP. Chujo and Utiyama (2004) and Utiyama et al. (2004) have established an easy-to-use tool using nine statistical measures to identify level-specific, domain-specific words (such as EAP) from a corpus. In this study, these measures were applied to a 1.63-million-word spoken educational/informative component of the British National Corpus composed of materials such as lectures, news commentaries and classroom interaction to produce nine word lists. We examined the top 500 most outstanding words of each list and confirmed that specific statistical measures produced level-specific lists of academic words. The selected spoken EAP words are grouped into three proficiency level sub-lists, which allow users to develop spoken EAP vocabulary lists on their own contexts.

1. Introduction

Because English has increasingly become a global language of communication, trade and research, the English for specific purposes (ESP) approach has been distinguished from general English in language teaching (Robinson, 1991; Hutchinson and Waters, 1987; Dudley-Evans and St. John, 1998). According to Robinson's "ESP family tree" (1991:3), a major distinction of ESP is often drawn between EOP (English for occupational purposes), involving work-related needs and training, and EAP (English for academic purposes), involving academic study needs. One of the prominent characteristics of ESP is a heavy load of corresponding specialized vocabulary or "technical words that are recognizably specific to a particular topic, field, or discipline" (Nation, 2001:198). In Chujo and Genung's 2004 study, we examined the former ESP, namely EOP, and reported on interesting and beneficial features of three business-oriented vocabularies both qualitatively and quantitatively.

Meanwhile for ESL and EFL students, the need for the latter ESP, namely EAP skills — understanding academic textbooks and journal articles and attending or giving presentations — is also growing (Coxhead, 2000; Hill, 2003; Takefuta and Suikou, 2005;

Evison and McCarthy, 2006; Yontz et al., 2006). Academic vocabulary development is essential in order to achieve a certain proficiency in EAP, and current research is now turning to developing EAP vocabulary and providing teachers and students with such word lists.

2. Review of the Literature

Coxhead (2000) developed an "Academic Word List (AWL)" for written academic English from a 3.5 million-word+ corpus of academic texts covering subject areas such as the arts, commerce, law, and science. The list contains 570 words generated by *range* and *frequency* and incorporates words not included in West's (1953) 2,000-word *A General Service List of English Words*. This written academic word list has been widely used in ESL and EFL settings (e.g. Hill, 2003; Parker and Allen, 2004; Huntley and Shaw, 2005; Grigorescu, Pena and Dwyer, 2006). Briggs and Lee (2002) developed a spoken academic database (Lexical Database of Academic Spoken English, or LDASE) from the Michigan Corpus of Academic Spoken English[1)]. It contains approximately 1.7 million words of academic speech collected between 1997 and 2001. Although it is spoken rather than written English, there is no clear means to "weed out non-academic/general spoken vocabulary" (Briggs and Lee, 2002) from academic vocabulary, and no spoken academic word list based on this database exists. Since the AWL is a written academic word list and the LDASE is a database rather than a spoken word list, apparently there seems to be a need for developing a spoken academic word list for pedagogic purposes.

Chujo and Utiyama (2004) and Utiyama et al. (2004) have established an easy-to-use tool employing nine statistical measures to identify level-specific, domain-specific words. Subsequently, Chujo and Utiyama (2005) created a list of written science vocabulary by applying those nine statistical measures to the 7.37-million-word written 'applied science' component of the British National Corpus (BNC). It was found that each measure extracted a different level of domain-specific words by vocabulary level, grade level, and school textbook vocabulary coverage and that specific measures produced level-specific words, i.e. *log likelihood ratio* (*LLR)* identified intermediate-level technical words, and *mutual information* (*MI)* identified advanced level technical words. These measures were effective in separating technical vocabulary from general-purpose vocabulary, and provide a useful template as a means of identifying spoken EAP vocabulary.

In 2000, the BNC became available outside EU countries. With more than 100 million words, it is considered "one of the largest and most representative corpora of a single variety of English currently available" (Kennedy, 2003:467). It contains about 10 million words of transcribed speech, covering a wide range of speech variation. Fortunately, a 1.63 million-word spoken component of the BNC comprises academic speech texts such as lectures and classroom interaction recorded in universities and schools. This sub-corpus of the BNC provides important academic spoken language data.

This review of the literature indicates the following conclusions. Firstly, there seems to be a need for a spoken EAP vocabulary list for pedagogic purposes. Secondly, it is possible to identify spoken EAP vocabulary by using statistical measures such as the *LLR* and *MI*. Finally, the 1.63-million word academic spoken language component of the BNC provides an excellent basis for the development of this kind of EAP list.

3. Purposes of the Study

The purposes of this study are (1) to extract various levels of spoken EAP words by applying nine statistical measures to the 1.63-million-word educational/informative component of the BNC; (2) to verify the proficiency level as measured by US native speaker grade level, and Japanese high school (JSH) textbook vocabulary coverage; and (3) to create beginner, intermediate and advanced level spoken EAP lists.

4. Procedure

4.1 The data

4.1.1 Spoken EAP Master List

In order to extract spoken EAP sub-lists, we needed to begin with one large master list of spoken EAP terms. To create this kind of spoken EAP master list, we began with the educational/informative spoken component of the BNC. This includes 169 spoken texts of monologues and dialogues in three subcategories: (1) lectures, talks, and educational demonstrations recorded within universities and schools, (2) news commentaries from national and regional broadcasting companies, and (3) classroom interactions including home tutorials (see Burnard, 2000:15).

The 1.63 million words in this corpus were first lemmatized to extract all base forms using the CLAWS7 tag set[2)]. (For example, *communicate*, *communicates*, *communicated*, and *communicating* are forms of the same word and were listed under a base word *communicate* with a frequency of four occurrences.) Secondly, if a word appeared fewer than 10 times in the corpus, it was deleted. Next, all proper nouns and numerals were identified by their part of speech tags and deleted manually because statistical measures mechanically identify these words as technical words (Scott, 1999). Finally, this process yielded a 3,839-word spoken EAP master list.

4.1.2 Control Lists

We wanted not only to extract spoken EAP words but also to know if these words appear generally in English at what [US] native speaker grade level. In addition, we wanted to know if these extracted EAP words are learned by Japanese students in the course of their junior and senior high school years, and if so, to what extent. For these reasons, three control vocabulary lists were created by using the same procedures

described above in 4.1.1, and these are described in detail below:
(1) The British National Corpus Spoken High-Frequency Word List (hereafter BNC SHFWL) is a list of 5,862 base words representing 7.5 million BNC spoken words that occur 10 times or more in the BNC spoken business, public/institutional, and leisure components (i.e. general topics). It was used for comparison to statistically determine how the spoken EAP words in our master list would appear differently from words in a general spoken corpus.
(2) *The Living Word Vocabulary* (Dale and O'Rourke, 1981) is a list that includes more than 44,000 items, and each has a percentage score to correlate word familiarity to [US] students' grade levels 4 through 16. For supplementing grade levels 1 through 3, reading grades from *Basic Elementary Reading Vocabularies* (Harris and Jacobson, 1972) were used. These lists were used to determine the grade level at which the central meaning of a word can be readily understood. Of course, using speaking grade data instead of reading grade data would be desirable, however, to our knowledge, no such data exist.
(3) The junior and senior high school (JSH) textbook vocabulary list containing 3,245 different base words compiled from the top selling series of junior high school textbooks (*The New Horizon 1, 2, 3* series) and senior high school textbooks (*The Unicorn I, II* and *Reading* series) in Japan was used to determine the percentage of EAP vocabulary that is covered in JSH textbooks. Japanese junior and senior high school students generally use these or similar books to study English before entering a university.

4.2 Identifying Outstanding Spoken EAP Words

4.2.1 Statistical Measures

In this study, we used nine statistical measures: simple *frequency* (*Freq*), the *Dice coefficient* (*Dice*), *Cosine* (*Cosine*), the *complementary similarity measure* (*CSM*), the *log likelihood ratio* (*LLR*), the *chi-square test* (*Chi2*), *chi-square test with Yates's correction* (*Yates*), *mutual information* (*MI*), and *McNemar's test* (*McNemar*)[3]. The formula for each measure is available on the web[4]. A detailed description of each measure can be found in Utiyama et al. (2004) and Chujo and Utiyama (forthcoming 2006) and the notation for these kinds of statistics can be found in Scott (1997).

4.2.2 Identifying Outstanding Spoken EAP Words

These statistical measures are widely used in computational linguistics. They automatically identify outstanding words in frequency of occurrence by making comparisons between one specified list (in this case, the spoken EAP master list) and another larger list (the BNC SHFWL). These statistics indicate whether a word is overused or underused in a specified list compared with a list of general English. We want to determine those words that 'stand out.' The statistical score of word *X*, i.e. the extent of the dissimilarity between two word lists, is calculated by comparing the patterns of the frequency of each word in the spoken EAP word list with the frequency of the same word

in the BNC SHFWL.

Using each measure, the statistical score for the extent of each word's "outstanding-ness" (Scott, 1999) in frequency of occurrence is computed as follows: (1) four variables 'a, b, c, d' ('the frequency of word *X* in the spoken EAP list,' 'the frequency of word *X* in the BNC SHFWL,' 'the number of running words in EAP not involving word *X*' and 'the number of running words in BNC SHFWL not involving word *X*') are computed for each word. (2) The variables are applied to each formula to yield each word's "outstanding-ness" score. Since each measure uses a different formula, it gives a different score to each word[5]. Finally, (3) the words are sorted from the most outstanding to the least outstanding by their statistical ranking. Thus the words near the top are ranked as very outstanding in terms of each statistical measure's criteria.

The goal in using these measures is to narrow the number of candidates for the spoken EAP word list, but it is not meant to be a definitive list. These statistical tools can help users to select technical vocabulary automatically without specialist knowledge. By using extracted lists, users can easily manually delete irrelevant words.

4.3 Verifying the Vocabulary Levels of the Extracted Lists

All the extracted lists were initially examined for an overview comparison of the top 30 extracted words[6]. Next, the 500 most outstanding words of each list were studied to determine their potential for pedagogic applications from grade level based on word familiarity[7] and number of words covered by the JSH textbook vocabulary.

4.4 Developing Tri-level Spoken EAP Lists

From each of the top 500 words created by the nine statistical measures, three levels of spoken EAP lists were systematically created as follows: (1) the JSH words were subtracted from each EAP list to produce a core of spoken EAP words that would be new to high school graduates; and (2) the core words are classified into three proficiency level groups based on grade level (described in 4.3 above).

5. Results and Discussion

5.1 Verifying the Vocabulary Levels of the Extracted Lists

5.1.1 Top 30 Extracted Words Overview Comparison

The top 30 words from each of the nine measures in descending order are shown in **Table 1**. Since the top 30 extractions made using *Freq/Dice* and *LLR/Chi2/Yates* were almost the same, they are shown in the same column. The bottom two rows of each column show the average frequency score and average word length of the top 30 words generated by each statistical measure. A glance at the top 30 words gives a brief overview, and exhibits general tendencies inherent in the extraction of nine measures.

The lists in **Table 1** are very different from each other even though they were extracted

from the same data. The top 30 words identified by *Freq/Dice* and *Cosine* are general spoken vocabulary that usually appears at the top of high frequency lists in both small and large corpora. For *CSM*, the top 30 extractions include some EAP words such as *minus* and *write*. For *LLR/Chi2/Yates*, the top 30 extractions include some important EAP words that are used in certain academic subjects such as mathematics *(minus*, *equal*, *square*, *function*, and *fraction)*, science *(acid* and *carbon)*, and history (*peasant*, *revolution*, *communist*, and *reform).* The *MI* and *McNemar* lists identify more academic EAP words such as *reactor*, *theology*, *perimeter*, *immune*, and *nucleus*.

Table 1 A Comparison of the Top 30 Words for Each Measure

Ranking	Freq/Dice	Cosine	CSM	LLR/Chi2/Yates	MI	McNemar
1	be	be	the	may	reactor	legacy
2	the	the	of	of	novel	ambition
3	and	of	a	peasant	myth	beet
4	you	a	in	minus	theology	diphtheria
5	to	and	so	okay	legacy	biographical
6	a	to	may	the	summation	perimeter
7	that	that	this	university	immune	disruption
8	it	you	okay	acid	atom	coup
9	of	it	which	which	monetary	frequency
10	I	in	to	kind	critically	translate
11	have	have	very	equal	nucleus	sacred
12	in	they	as	science	therapeutic	warfare
13	do	I	right	square	quantitative	overthrow
14	they	we	if	function	butane	backslash
15	not	do	by	fraction	anthem	hemisphere
16	we	this	and	so	particle	clash
17	will	so	many	carbon	reproduction	resistor
18	this	may	much	revolution	agrarian	motorist
19	what	not	or	a	enlighten	descriptive
20	so	will	way	sense	translation	underline
21	can	what	how	very	christianity	lab
22	get	can	sort	communist	statistic	tolerate
23	on	if	from	by	numerical	crescendo
24	there	there	kind	reform	critic	suffix
25	he	but	write	lecture	civilization	symbolic
26	but	right	minus	ego	ambition	demon
27	for	which	about	in	composer	ethanol
28	go	as	thing	graph	sulphate	skeletal
29	if	on	because	between	beet	cent
30	think	okay	some	differentiate	differ	compartment
Average Frequency	25,221	24,513	12,575	7,268	28	10
Average Word Length	2.8	2.9	3.5	5.4	8.0	7.7

As we see from the data in the bottom two rows of **Table 1**, the average frequency score of each list decreases from left to right or from *Freq* to *McNemar*. Inversely, the average word length increases from left to right, ranging from 2.8 to 8.0 and 7.7 letters. As Chujo, Utiyama, and Nishigaki (forthcoming 2006) have shown, difficulty levels increase with increasing word length. Although we are aware that word difficulty may be influenced by factors other than frequency and word length, this might support the possibility that

specific statistical measures can be used to target specific grade level vocabulary. This will be explored in the following sections.

5.1.2 Top 500 Word Grade Level Comparisons

Once we were able to create lists of extracted words, we wanted to investigate at what US grade level these words would be understood by native English speaking (NS) children. In 1981, Dale and O'Rourke published *The Living Word Vocabulary* which is "an inventory of the written words known by children and young people in grades 4, 6, 8, 10, 12, 13, and 16" (1981:vii). Based on this data, **Figure 1** shows at what grade level the majority of NS students (80%) would readily understand the central meaning of each word for the top 500 extractions produced by the statistical measures[8]. To the best of our knowledge, there is no similar data available for grades 1 through 3, so for this comparison, we used reading grade (as opposed to written grade) word familiarity levels from Harris and Jacobson (1972). Note that in **Figure 1**, 'N/A' denotes those words not appearing in either the written or reading resources.

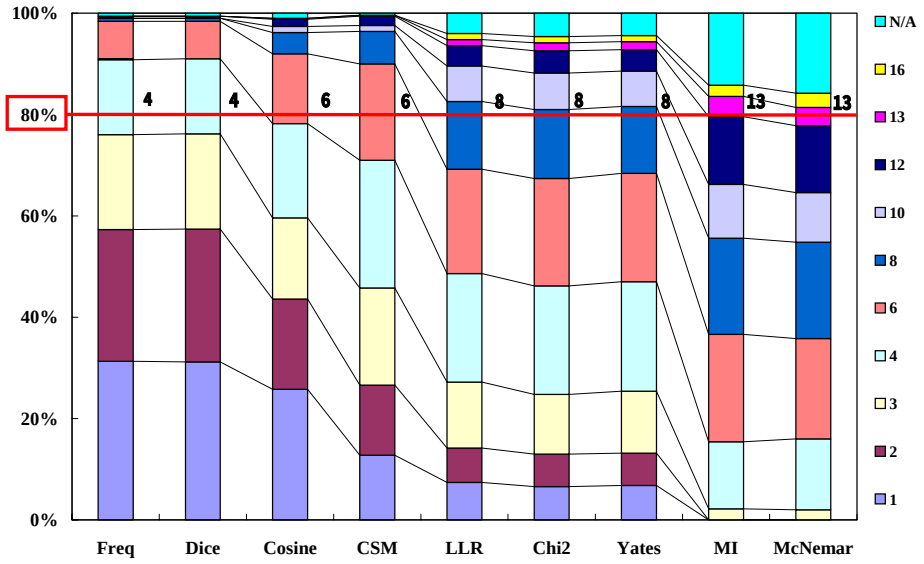

Figure 1 US Grade Level Based on Word Familiarity

In looking at the horizontal line and corresponding grade levels for each bar in the graph, we can see that 80% of the top 500 words from *Freq* and *Dice* are understood by 4th grade level students; those of *Cosine* and *CSM* are understood by 6th grade students; those of *LLR*, *Chi2* and *Yates* are known by 8th grade students; those of *MI* and *McNemar* are known by 13th grade students. In other words, NS elementary school children understand most of the top 500 words from *Freq/Dice/Cosine/CSM*, NS secondary school students understand those of *LLR/Chi2/Yates*, and NS college freshmen understand those of *MI/McNemar*.

5.1.3 Japanese High School Textbook Vocabulary Coverage and Implications

As educators in Japan, our primary interest is in how this extracted EAP vocabulary compares to what Japanese students may or may not already have studied and therefore how useful these lists might be. To do this, we compared the top 500 extractions to the vocabulary representing what most college students have studied before entering university. This list, comprised of 3,245 different words, was compiled from the top selling series of junior and senior high school (JSH) textbooks in Japan from the 7th through 12th grades. (See 4.1.2 for a detailed description of this control list.) **Figure 2** shows both what percentage of the top 500 extractions do appear in JSH textbooks in the lower section of the graph, and what percentage do not appear, in the upper section of the graph. For EFL teachers and learners in Japan, having a general idea of the vocabulary level of EAP words and knowing whether or not these are covered in JSH textbooks is important information. We see from **Figure 2** that while only 5 percent of the *Freq/Dice* top 500 extractions are not covered in the JSH school textbooks; 16 percent of the *Cosine*, 20 percent of the *CSM* extractions; 45 percent of the *LLR*, 48 percent of the *Chi2* and 47 percent of *Yates* are not covered; and 86 percent of the *MI* and 85 percent of *McNemar* extractions are not covered in the JSH school textbooks. The data in **Figure 2** again verifies that the nine different statistical measures extract quite different grade levels of words.

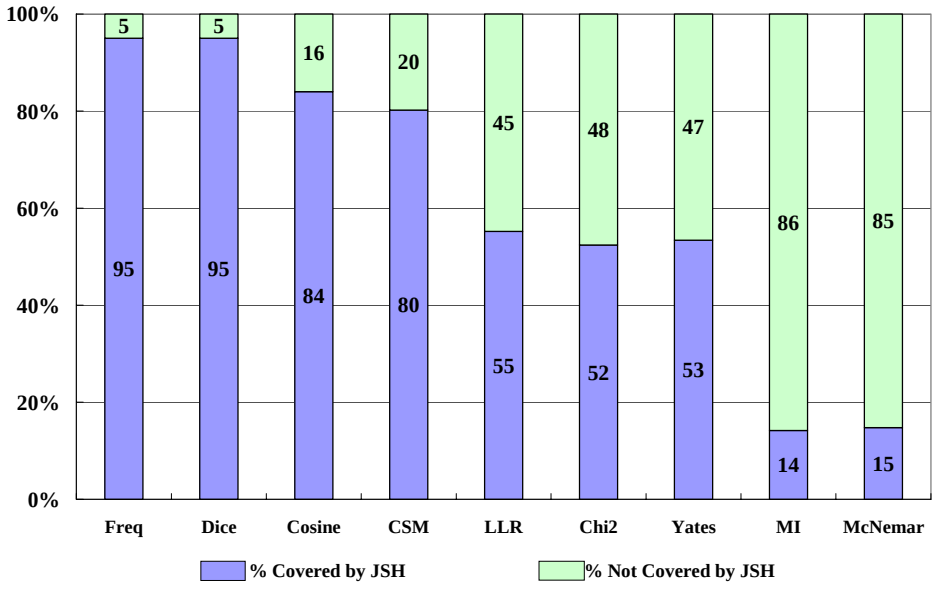

Figure 2 Percentage of Top 500 words Covered / Not Covered by JSH Textbooks

The results of these analyses support the finding that certain statistical measures tend to extract EAP vocabulary belonging to certain grade level. In terms of practical pedagogical application, we inferred from both grade level familiarity data and JSH text coverage data

in this study, in addition to several similar previous studies (Chujo and Utiyama, 2004; 2005; 2006), that (1) the spoken EAP words extracted by *Freq/Dice/Cosine/CSM* would be good for beginner level Japanese EFL learners; (2) the *LLR/Chi2/Yates* lists would be suitable for intermediate level Japanese EFL learners; and (3) *MI* and *McNemar* would be appropriate for advanced level Japanese EFL learners.

5. 2 Developing Beginner, Intermediate and Advanced Level Spoken EAP Lists

To create three levels of spoken EAP lists from each of the top 500 words created by the nine statistical measures, we first subtracted all the words taught in junior and senior high school by comparing each list to the JSH vocabulary list. That gave us 648 spoken EAP core words that would be new to high school graduates. Only 57 of these [spoken] words appeared in Coxhead's [written] AWL. From these 648 words, we next designated the 116 words appearing in at least one of the top 500 *Freq/Dice/Cosine/CSM* lists as beginner-level spoken EAP words. Next, 157 words appearing in at least one of the top 500 *LLR/Chi2/Yates* lists were designated as intermediate-level spoken EAP words. Finally, the 375 words extracted by *MI* and *McNemar* were categorized as advanced-level spoken EAP words.

In order to understand the individual items in the spoken EAP lists, we need information about the discipline/academic division in which they occurred. Such information can be obtained by searching the BNC educational/informative spoken component on the web[9)]. We examined the subject areas of the source texts from which each of the extracted 648 words was derived and tried to verify each EAP word's domain or subject area. **Table 2** shows some examples of beginner, intermediate, and advanced level spoken EAP words under the subject areas that they mainly occurred. In each subject area, the five highest frequency words are shown. We can see the level differences among the three proficiency level words by comparing the extracted words within the same subject. For example if we look at 'education' we can see that the beginner level words *print*, *process*, *pupil*, *quarter* and *suggest* are relatively simple, and the intermediate level words *achievement*, *concentration*, *curriculum*, *literacy*, and *tutorial* are more difficult than the beginner words. The advanced level words *convergence*, *literal*, *profound*, *stanza*, and *talented* seem to be the most difficult when compared with the other two levels. Since the JSH words have already been subtracted from these lists, the gap in proficiency among three levels is reduced, but still we can recognize the overall differences in proficiency among these levels.

Table 2 Examples of Tri-level Spoken EAP Words by Subject-Area Classification

Beginner

Education	History	Psychology	Economics	Biology	Chemistry	Mathematics
print	communist	abuse	agriculture	acid	calcium	angle
process	landlord	analysis	correlation	compound	carbonate	fraction
pupil	peasant	ego	protectionism	mechanism	gas	function
quarter	reform	self	urban	muscle	metal	graph
suggest	revolution	sex	variable	position	oxygen	multiply

Intermediate

Education	History	Psychology	Economics	Biology	Chemistry	Mathematics
achievement	communism	anxiety	consumption	membrane	alkali	cube
concentration	egalitarianism	consciousness	differential	molecule	electron	curve
curriculum	feudalism	identification	inefficient	organic	hydroxide	infinity
literacy	industrialization	repression	statistic	protein	neutron	inverse
tutorial	revolutionary	superego	textile	toxin	oxide	probability

Advanced

Education	History	Psychology	Economics	Biology	Chemistry	Mathematics
convergence	agrarian	altruism	contentious	enzyme	anaerobic	discontinuity
literal	authoritarian	cognitive	diagnostic	molecular	inorganic	displacement
profound	propaganda	evolutionary	empirical	optical	nitrate	fluctuation
stanza	redistribution	incest	migrate	receptor	ore	geometry
talented	tyrant	psychoanalytic	regression	selective	soluble	quantum

In searching for each word's subject area, we became aware that some words occurred often in a limited number of specific subject areas, and others occurred in a wider range of subject areas. The words shown in **Table 2** belong in the former category, although the latter category words are also very important for ESL and EFL learners who intend to engage in academic study in English because they are academic words that "occur reasonably frequently over a very wide range of academic texts" (Nation, 2001:17). Because there may be some variation in what users will select, and in order to ensure the spoken EAP words are useful for all learners, we attached the beginner, intermediate, and advanced level top 100 spoken EAP lists in **Appendix 1**, **2** and **3**. Each list's 100 high frequency words are shown in alphabetical order. We also recognize and ask readers to note that this method for extracting domain-specific vocabulary is useful and effective in that it reduces the number of relevant candidates to a manageable list (500 from an original corpus of 1.63 million) from which educators or students can select appropriate candidates based on the appropriate context. Clearly casting a wide net over such a large corpus will invariably produce some extraneous entries.

6. Conclusion

In this study, spoken EAP words were carefully selected from the real speech data of the

BNC educational/informative component using nine statistical measures. The selected spoken EAP words are grouped into three proficiency level lists, which allow users to further refine the candidates based on their own appropriate contexts and level. These lists are useful to teachers for college or university EAP courses both in Japan and abroad, and are also useful to students who are interested in developing spoken EAP vocabulary on their own.

Further research is aimed at developing these spoken EAP words into e-learning materials for vocabulary building. Our goal is also to determine how to incorporate Coxhead's written EAP list with these spoken EAP lists so that students will be able to become competent both in written and spoken academic vocabulary.

Notes

* Part of this study is based on a presentation given at the Inaugural International Conference on the Teaching and Learning of English in Asia, November 17, 2005, in Penang, Malaysia.

1) Michigan Corpus of Academic Spoken English (MICASE) Web Site: http://www.lsa.umich.edu/eli/micase/index.htm.
2) CLAWS7: http://www.comp.lancs.ac.uk/computing/users/eiamjw/claws/claws7.html.
3) References for each measure are as follows: *Dice* (Manning and Schütze, 1999), *Cosine* (Manning and Schütze, 1999), *CSM* (Wakaki and Hagita, 1996), *LLR* (Dunning, 1993), *Chi2* and *Yates* (Hisamitsu and Niwa, 2001), *MI* (Church and Hanks, 1989), and *McNemar* (Rayner and Best, 2001).
4) The formulas are available at http://www2.nict.go.jp/jt/a132/members/mutiyama and http://www5d.biglobe.ne.jp/~chujo/eng/index.html.
5) For example, the scores of *MI* and *Dice* are obtained by the following formulas:
   $MI = \log (a (a+b+c+d) / ((a+b)(a+c)))$ $Dice = 2a / (2a+b+c)$
6) Based on findings from our previous studies, we judged that the top 30 words show the general tendency of the overall extractions.
7) From our previous studies, we can say that the top 500 words exhibit the characteristic of each measure's extraction.
8) Note that grades 13 through 16 denote four years at the college or university level.
9) In order to retrieve the source text we used the Shogakukan Corpus Network: http://scn02.corpora.jp/~sakura03/. There is a charge for access to this website.

References

Briggs, S. and Lee, D. (2002) "Developing a Lexical Database of Academic Spoken English (LDASE) for Language Testing: Problems and Prospects." Paper presented at the 4th North American Symposium on Corpus Linguistics and Language Teaching,

Indianapolis, Indiana.

Burnard, L. (2000) "Reference Guide for the British National Corpus (World Edition)." http://www.natcorp.ox.ac.uk/World/HTML/thebib.html.

Chujo, K. and Genung, M. (2004) "Comparing the Three Specialized Vocabularies Used in 'Business English,' TOEIC, and British National Corpus Spoken Business Communications." *Practical English Studies*, *11*, 1-15.

Chujo, K. and Utiyama, M. (2004) "Toukeiteki shihyou wo shiyoushita tokuchougo chuushutsu ni kannsuru kenkyuu [Using Statistical Measures to Extract Specialized Vocabulary from a Corpus]." *KATE Bulletin*, *18*, 99-108.

——— (2005) "Selecting Level-Specific BNC Applied Science Vocabulary Using Statistical Measures." *Selected Papers from the Fourteenth International Symposium on English Teaching*, Taipei: English Teachers' Association/ROC, 195-202.

——— (forthcoming 2006) "Selecting Level-Specific Specialized Vocabulary Using Statistical Measures." *System*, *34*, 2.

Chujo, K., Utiyama, M. and Nishigaki, C. (forthcoming 2006) "Towards Building a Usable Corpus Collection for the ELT Classroom." In E. Hidalgo, L. Quereda, and J. Santana (Eds.), *Corpora in the Foreign Language Classroom*. Amsterdam: Rodopi.

Church, K. and Hanks, P. (1989) "Word Association Norms, Mutual Information, and Lexicography." *Proceedings of ACL-89*, 76-83.

Coxhead, A. (2000) "A New Academic Word List." *TESOL Quarterly*, *34*, 2, 213-238.

Dale, E. and O'Rourke, J. (1981) *The Living Word Vocabulary*. Chicago: World Book-Childcraft International, Inc.

Dudley-Evans, T. and St. John, M. J. (1998) *Developments in English for Specific Purposes: A Multi-Disciplinary Approach*. Cambridge: Cambridge University Press.

Dunning, T. E. (1993) "Accurate Methods for the Statistics of Surprise and Coincidence." *Computational Linguistics*, *19*, 1, 61-74.

Evison, J. and McCarthy, M. (2006) "Features of Academic Discourse and EAP." Paper presented at a meeting of the 40th TESOL Convention, Tampa, FL, 3/15/2006.

Grigorescu, C., Pena, J. and Dwyer, E. (2006) "A K-12 Academic Word List: an Update." Paper presented at a meeting of the 40th TESOL Convention, Tampa, FL, 3/18/2006.

Harris, A. J. and Jacobson, M. D. (1972) *Basic Elementary Reading Vocabularies*. New York: The Macmillan Company.

Hill, M. (2003) "Academic and Professional Vocabulary Learning Online." *GLOBALED Internet Journal*, University of Melbourne. http://www.globaled.com/articles/HillMonica2003.pdf (retrieved 11/1/2005)

Hisamitsu, T. and Niwa, Y. (2001) "Topic-Word Selection Based on Combinatorial Probability." *NLPRS-2001*, 289-296.

Hutchinson, T. and Waters, A. (1987) *English for Specific Purposes*. Cambridge: Cambridge University Press.

Huntley, H. and Shaw, D. (2005) "Active Learning of Academic Vocabulary." Paper presented at a meeting of the 39th TESOL Convention, San Antonio, TX, 4/1/2005.

Ichikawa, Y., Yasuyoshi, I., Hestand, J.R., Shiokawa, H., Kobayashi, C., and Ishizuka, K.

(2002) *Unicorn English Course I & II*. Tokyo: Bun'eido.

Ichikawa, Y., Yasuyoshi, I., Hestand, J.R., Shiokawa, H., Kobayashi, C., and Hagino, T. (2003) *Unicorn English Reading*. Tokyo: Bun'eido.

Kasajima, J., Asano, H., Shimomura, Y., Makino, T., Ikeda M. et al. (2002) *New Horizon English Course 1, 2, & 3*. Tokyo: Tokyo Shoseki.

Kennedy, G. (2003) "Amplifier Collocations in the British National Corpus: Implications for English Language Teaching." *TESOL Quarterly*, *37*, 3, 467-487.

Manning, C. D. and Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*. Cambridge: The MIT Press.

Nation, I. S. P., (2001) *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Parker, M. and Allen, E. (2004) "Individualizing Academic Vocabulary Acquisition in the Classroom." Paper presented at a meeting of the 38th TESOL Convention, Long Beach, CA, 3/31/2004.

Rayner, J. C. W. and Best, D. J. (2001) *A Contingency Table Approach to Nonparametric Testing*. New York: Chapman & Hall/CRC.

Robinson, P. (1991) *ESP Today: A Practioner's Guide*. New York: Prentice Hall.

Scott, M. (1997) "PC Analysis of Key Words and Key Key Words." *System*, *25*, 2, 233-245.

——— (1999) "WordSmith Tools Manual [Computer software]." Oxford: Oxford University Press.

Takefuta, Y. and Suikou, M. (2005) *Kore kara no daigaku eigo kyouiku [College English Education in the Future]*. Tokyo: Iwanami Shoten.

Utiyama, M., Chujo, K., Yamamoto, E. and Isahara, H. (2004) "Eigokyouiku no tameno bunya tokuchou tango no sentei shakudo no hikaku [A Comparison of Measures for Extracting Domain-Specific Lexicons for English Education]." *Journal of Natural Language Processing*, *11*, 3, 165-197.

Wakaki, M. and Hagita, N. (1996) "Recognition of Degraded Machine-Printed Characters Using a Complementary Similarity Measure and Error-Correction Learning." *IEICE Trans. Inf. & Syst*. E79-D, 5.

West, M. (1953) *A General Service List of English Words.* London: Longman, Green & Co.

Yontz, R., Cortes, V., Reppen, R. and Simpson-Vlach, R. (2006) "English for Specific Purposes: Genre and Corpora in EAP Classrooms." Paper presented at a meeting of the 40th TESOL Convention, Tampa, FL, 3/15/2006.

Acknowledgements

This study was funded by a Grant-in-aid for Scientific Research (No.17520401) from the Ministry of Education, Science, Sports and Culture. It was also supported in part by Shogakukan Inc. We thank Dr. Masao Utiyama, National Institute of Information and Communications Technology, for his contribution. We are grateful to the anonymous reviewers for detailed comments on an initial draft of this article.

Appendix 1 Most Frequently Appearing Beginner-Level Spoken EAP Words

absolute	carbonate	expectation	naught	quarter
absolutely	chemistry	extent	normally	range
abuse	claim	fraction	notion	ratio
acceleration	comma	function	novel	reform
acid	communist	gas	obviously	reproductive
actual	compound	gene	occur	resistance
agricultural	concept	general	oxygen	revolution
agriculture	constant	gradient	peasant	sample
aid	correlation	graph	physics	self
alright	cos	historian	pizza	sex
analysis	council	input	police	sodium
angle	couple	integrate	policy	sort
argue	diagram	junction	position	suggest
association	differentiate	landlord	potential	total
awkward	domestic	magistrate	presumably	towards
basically	ego	mechanism	print	tutor
bracket	equation	metal	process	unit
calcium	essentially	moral	property	urban
calculator	evidence	multiply	psychology	variable
cancel	exam	muscle	pupil	whereas

Appendix 2 Most Frequently Appearing Intermediate-Level Spoken EAP Words

abstract	conviction	genetic	laboratory	radical
achievement	critic	gram	lecturer	rational
adjacent	cube	gulf	marginal	reader
alkali	cultural	handout	mathematical	repression
anxiety	curriculum	handset	membrane	revolutionary
apostrophe	curve	hostage	methyl	select
axis	decimal	hydrochloric	moderate	sibling
bandage	differentiation	hydroxide	molecule	slope
behave	distinguish	identification	monoxide	so-called
biological	dominant	imply	motivation	statistic
casualty	dyslexic	industrialization	neutron	straightforward
centimeter	egalitarianism	inequality	nitrogen	structural
circuit	electron	infinity	origin	subsistence
coefficient	evaluation	insight	oxide	superego
communism	extreme	intellect	parameter	textile
concentration	feudalism	intellectual	potassium	toxin
conductance	functional	interaction	presidency	tutorial
consciousness	furnace	interval	probability	universe
constitution	galaxy	ion	protein	zero
consumption	gamma	isomer	psychological	zinc

Appendix 3 Most Frequently Appearing Advanced-Level Spoken EAP Words

absent	diagnostic	in-service	physiological	sling
acute	displacement	latent	plural	soluble
agrarian	empirical	limitation	poisonous	spontaneous
altruism	equilibrium	literal	positional	subjective
analyst	evolutionary	magnesium	preservation	sulfuric
archaeology	excess	mammal	primal	suspicion
bakery	fantasy	managerial	profound	taboo
bodily	feudal	manifest	progression	therapeutic
butane	fingerprint	microprocessor	propaganda	transference
classical	formation	molecular	proton	translation
clinical	geometry	newly	psychoanalytic	traumatic
cognitive	gravis	nitrate	quantitative	tremendously
collectivization	handwriting	node	receptor	typically
compose	harmonic	optical	reef	tyrant
compute	hearsay	ore	regression	unchanged
consciously	helium	paradox	repress	undergraduate
consume	hopeful	passion	restorationist	universal
convergence	idiomatic	persuasive	reveal	uranium
coral	incest	pessimistic	rigid	vivid
cursor	infinite	philosophical	semantics	vowel