Selecting Level-Specific BNC Applied Science Vocabulary Using Statistical Measures

Kiyomi Chujo* Masao Utiyama**
Nihon University*
National Institute of Information and Communications Technology**
chujo@cit.nihon-u.ac.jp mutiyama@nict.go.jp

The effectiveness of using statistical measures to extract multi-level applied science vocabulary from a large corpus was examined for the potential of being an easy-to-use tool for teachers and developers of pedagogical materials. Nine statistical measures were applied to a 7.37-million-word written 'applied science' component of the British National Corpus (BNC) to identify its domain-specific words, and an examination of the resulting vocabulary lists showed 1) that each statistical measure extracted a different level of domain-specific words by vocabulary level, grade level, and school textbook vocabulary coverage; 2) that specific measures produced level-specific words, i.e. beginning-level words were identified by cosine and complimentary similarity measures, intermediate-level words were produced by the log-likelihood ratio, the chi-square test, and the chi-square test with Yates correction, and advanced-level word lists were created with mutual information and McNemar's test. The authors conclude that the application of statistical measures can be an effective tool for identifying and selecting level-specific BNC applied science vocabulary for pedagogical purposes.

BACKGROUND

Because English is increasingly becoming a lingua franca for international technology and communications, there has been a growing interest in and necessity for English for Science and Technology (EST). According to "the tree of ELT" (English language teaching — Hutchinson and Waters, 1987), EST is categorized as one of the branches of English used for specific purposes (ESP), all of which are differentiated from general English. This kind of English is important not only in scientific and technological activities but also in universities, which, increasingly, find themselves responsible for providing EST-related English skills to an ever-expanding population of science and technology students.

Generally speaking, English classes at science and technology colleges in Japan consist of three types: English for General Purposes (EGP), EST, and, "semi-EST" courses. EGP courses are taught to freshman and sophomore students and are designed to further the student's abilities in using English as a communicative tool. Standard EFL college level textbooks are used in these courses. Usually EST courses are taught to seniors and graduate students and are designed to inculcate students with an ability to read and write the scientifically and technologically oriented English that they are likely to encounter in their professional careers. This goal has been greatly underscored by the rise of English as the recognized international language of science and technology. Consequently, technical articles from professional journals are used in lieu of a textbook. Semi-EST courses are taught to juniors and are meant to bridge the gap between the types of English used in EGP and EST classes; or, in other words, to supply a transition between EGP and EST classes.

In ESP, or EST, one characteristic of the linguistic knowledge needed to comprehend specialized texts is the heavy load of corresponding specialized vocabulary or "technical words that are recognizably specific to a particular topic, field, or discipline" (Nation 2001:198), for these words convey the import of the subject knowledge. In a previous study, we measured the graduations among vocabulary levels found within EGP, semi-EST, and EST materials used at a college of science and technology, as measured by the 13,994-word lemmatized "British National Corpus High Frequency Word List" (BNC HFWL) as a criterion. (This comparison procedure is detailed in Chujo and Genung, 2003.) The resulting study confirmed the existence of a large gap in the vocabulary level between the EGP and the EST teaching materials. The study also revealed that the texts used in the semi-EST classes have only a limited efficacy in bridging this big gap, but that when supplemented by a specialized EST vocabulary list, they can be more helpful in doing so.

We hypothesized that since the gap in vocabulary between the EGP and the EST teaching materials is so large that selecting and supplementing several levels of specialized EST vocabularies that would supplement semi-EST courses according to learners' proficiency levels in a graduated, step-by-step fashion could probably best improve the efficacy of the English. We know from Sutarsyah et al. (1994:48) that selecting multi-level beginning, intermediate, and advanced specialized vocabularies using the traditional vocabulary selection criteria of *frequency* and *range* is only partly successful in identifying the technical words. Because the focus of these measures is ranking general-purpose vocabulary in order of priority, separating technical vocabulary from general-purpose vocabulary is still labor-intensive, time-consuming, and heavily dependent on the selector's expertise in English education and specialist knowledge of the domain, which English teachers generally do not have. A means is clearly needed that provides easy-to-use tools that identify multi-level applied science vocabulary.

A number of corpus-based studies have used certain statistical measures to identify technical vocabulary. For example, Nelson (2000) used the *log-likelihood (LL)* statistic from WordSmith to find words that are statistically more frequently used in business English than in general English by comparing each word's frequency in the business English corpus with its frequency in the British National Corpus (BNC). He was able to generate a list of business-related words such as *business*, *market*, *customer*, *management*, *price*, and *bank*.

In a preliminary study, Chujo and Utiyama (2004) examined a range of statistical measures used in computational linguistics to identify technical vocabulary from a 100,000-word specialized corpus. Eight measures such as the *LL* and the *mutual information* (*MI)* were examined. They are statistics that indicate whether a word is overused or underused in a specialized corpus compared with a corpus of general English. Each resulting list was compared to an existing technical vocabulary control list, and the corresponding statistical measures were evaluated for their effectiveness by calculating the proportion of relevant candidates they produced. It was determined that all these measures effectively produce relevant technical vocabulary and that each measure creates a unique type of word list that can be specifically applied to student proficiency levels and lexicons. Our present study applies the same methods but to a much larger corpus, includes an additional statistical measure, and explores pedagogical applications based on BNC frequency, native speaker grade level, and Japanese textbook coverage.

Because of the lack of general agreement on how to define technical vocabulary (Justeson and Katz, 1995), we must clarify some terms. The rank-ordered lists produced by each statistical measure are called *specialized words/lists/vocabulary*. We have used the broad definitions of *technical vocabulary*; i.e., specialized lists contain three types of words: *technical vocabulary*, or words specific to a field, *sub-technical vocabulary*, or words more common in the specified field than elsewhere and less obviously technical vocabulary, and *general vocabulary*, which is a general base of English words.

RESEARCH QUESTIONS

We examined the effectiveness of using nine statistical measures to identify multi-level applied science vocabulary as easy-to-use tools for teachers and material writers. Specifically, the following questions were addressed:

1. What types of applied science words are extracted by each measure, and how are they ranked?
2. How frequently do the top (most prominently appearing) 500 words extracted by each method occur in the BNC?
3. At what U.S. grade level are the top 500 words extracted by each method understood?
4. What percentage of the top 500 words extracted by each method appear in Japanese high school (JSH) textbooks?

METHOD

Applied Science Master Word List

To extract applied science sub-lists, we needed to begin with one large master list of applied science terms. To create this kind of applied science-related master list, we began with the 'applied science' written component of the BNC. The 7.37-million words in this corpus were first lemmatized to extract all base forms using the CLAWS7 tag set. Then, for pedagogical application, all unusual or infrequent words were eliminated by deleting words which appear fewer than 100 times in the corpus. This created a list of 6,718 different words. All proper nouns and numerals were identified by their part of speech tags and deleted manually. Finally, this process yielded a 3,407-word applied science master list.

Control Lists

Three control vocabulary lists were used:

(1) The British National Corpus High Frequency Word List (BNC HFWL), a list of 13,994 lemmatized words representing 86 million BNC words that occur 100 times or more (compiling procedure is detailed in Chujo, 2004), was used for comparison to statistically determine if and how these applied-science-related words appear differently in a general corpus.

(2) *The Living Word Vocabulary* (Dale and O'Rourke, 1981) includes more than 44,000 items, and each has a percentage score that rates whether the word is familiar to students in U.S. grade levels 4 through 16. This list was used to determine the grade level at which the central meaning of a word can be readily understood.

(3) The junior and senior high school (JSH) textbook vocabulary list containing 3,245 different base words was compiled from the top selling series of JSH textbooks (the *New Horizon 1, 2, 3* series and the *Unicorn I, II* and *Reading* series) in Japan. Japanese high school students generally use these or similar books to study English before entering a university.

Statistical Measures

The measures examined were mutual information (*MI*) (Church and Hanks, 1989), the log-likelihood ratio (*LL*) (Dunning, 1993), the chi-square test (*Chi2*) and chi-square test with Yates's correction (*Yates*) (Hisamitsu and Niwa, 2001), the Dice coefficient (*Dice*) (Manning and Schütze, 1999), Cosine (*Cosine*) (Manning and Schütze, 1999), the complementary similarity measure (*CSM*) (Wakaki and Hagita, 1996), McNemar's test (*McNemar*) (Rayner and Best, 2001) and frequency (*Freq*). These statistics automatically identify prominent words by making comparisons between one specified list and another larger list. The formula for each measure and a detailed description of each measure can be found in Utiyama et al. (2004). The statistical score for the extent of each word's "outstanding-ness" (Scott, 1999) in frequency of occurrence is computed, and the words are sorted from the most outstanding to the least outstanding. Thus the words near the top are ranked as outstandingly prominent in terms of each statistical measure's criteria.

The goal of identifying specialized words by using these measures is to narrow down the number of candidates for the category of technical or sub-technical items, not to totally extract these items. Simply deleting the poor candidates would be a much simpler task for teachers and material writers than creating the entire list manually.

RESULTS AND DISCUSSION

1. What types of applied science words are extracted by each measure, and how are they ranked?

The top 50 words from each of the nine different measures in ascending order are shown in **Table 1**. Since the top 50 extractions made using *Freq* and *Dice*, *Cosine* and *CSM*, and *Chi2* and *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 50 words generated by each statistical measure.

Table 1. Top 50 Specialized Words in the BNC Applied Science Corpus Generated by Nine Measures

	Freq, Dice	Cosine, CSM	LL	Chi2, Yates	MI	McNemar
1	the	the	system	patient	client-server	adenoma
2	be	be	patient	system	biliary	antrum
3	of	of	software	software	silicon	malabsorption
4	and	in	user	user	lab	postprandial
5	to	a	computer	computer	hypertext	idiopathic
6	a	to	use	module	duct	hypertext
7	in	and	of	use	interoperability	luminal
8	it	use	the	file	endanger	value-added
9	have	system	module	technology	pixel	colonoscopy
10	that	for	be	database	high-end	distension
11	for	this	file	cell	reseller	percutaneous
12	with	with	technology	gastric	semiconductor	manometry
13	this	by	information	application	keyword	sclerotherapy
14	on	patient	application	of	motility	neuropathy
15	by	or	cell	information	sclerotherapy	connectivity
16	will	will	database	disease	metaplasia	proliferative
17	they	which	disease	the	manometry	duodenum
18	as	user	version	disk	adenoma	scalable
19	not	computer	product	version	mucosal	histology
20	or	from	network	network	endoscopic	recognizer
21	from	software	disk	be	colonic	cholecystectomy
22	can	information	package	interface	ileal	machine-readable
23	which	also	gastric	package	gastrin	radiocarbon
24	at	may	program	acid	distension	secretory
25	you	company	acid	server	antrum	habituation
26	use	can	process	product	filename	asynchronous
27	we	new	interface	program	motherboard	supplementation
28	but	study	study	environment	duodenal	ileum
29	he	program	server	process	endoscopy	lamina
30	system	show	environment	processor	cholangitis	perfusion
31	many	technology	this	workstation	reflux	motherboard
32	do	file	company	bile	plasminogen	interoperability
33	much	application	sun	sun	colonoscopy	autonomic
34	may	process	library	error	idiopathic	ulceration
35	all	product	error	ulcer	malabsorption	maximal
36	there	include	machine	study	ileum	radiological
37	if	as	window	input	resection	analyzer
38	I	cell	treatment	colitis	object-oriented	cirrhosis
39	also	result	result	hardware	shareware	endanger
40	year	number	concentration	digital	pepsinogen	insignia
41	time	module	scientist	library	gastric	lymph
42	other	each	processor	scientist	online	rectum
43	make	group	input	concentration	colitis	parietal
44	some	provide	electronic	dose	percutaneous	thesaurus
45	new	disease	workstation	electronic	luminal	pancreas
46	patient	many	in	machine	cholecystectomy	metaplasia
47	say	control	type	biopsy	ulcerative	cytoplasmic
48	only	version	available	mainframe	colorectal	byte
49	when	year	also	colonic	pancreatitis	gluten
50	into	water	code	desktop	pancreatic	lumen
Average Frequency	58279	48263	28659	24038	343	123
Average Word Length	3.2	4.9	6.5	6.7	9.0	9.6

The specialized lists in **Table 1** are very different from each other even though they were extracted from the same data. We can see that the BNC applied science corpora includes various areas such as engineering, medicine, and chemistry. The top 50 words identified by *Freq* and *Dice* are general vocabulary that usually appears at the top of high frequency lists in both small and large corpora. For *Cosine* and *CSM*, the top 50 extractions are some important 'basic EST words' that are all high-frequency words and have particular technical uses in applied science, for example *system*, *application*, *module*, *cell*, *patient* and *disease*. The *LL*, *Chi2*, and *Yates* seem to be well-suited to identifying sub-technical words in applied science such as *interface*, *server*, *network*, *digital*, and *code* in computer science and also *gastric*, *acid*, *ulcer*, *bile*, *colitis* and *biopsy* in medical science, and they appear to bridge the gap between general English and technical words. The *MI* and *McNemar* lists identify EST technical words such as *hypertext*, *pixel*, and *semiconductor* related to computer science and technology, *biliary*, *antrum*, *duct*, *postprandial*, *motility*, *histology*, and *endoscopy* in the medical field.

As we see from the data in the bottom two rows of **Table 1**, the average frequency score of each list, ranging from 58279 to 123, decreases from left to right or from *Freq* to *McNemar*. Inversely the average word length of lists increases from left to right, ranging from 3.2 to 9.6. As Takefuta et al. (1994) have shown, difficulty levels in words increase with increasing word length. This suggests that each measure identified different difficulty levels of words as its outstanding words and this supports the possibility that specific statistical measures can be used to target specific grade-level vocabulary.

2. How frequently do the top 500 words extracted by each method occur in the BNC?

The BNC represents present-day general vocabulary usage. We examined the frequency distribution of the top 500 extracted words by using the BNC HFWL, which was divided into 14 1000-word frequency bands of the most frequent words. 'BNC frequency bands 1000' indicates ranks 1 to 1000, 'frequency bands 2000' indicates ranks 1001 to 2000, etc. The percentages of the 500 words of each list that belong to each frequency band are shown in **Table 2**; a blank space indicates that no words belonged to that band. Because the scores for *Freq* and *Dice* were identical, and those of *Chi2* and *Yates* were almost the same, only seven columns are shown.

Table 2. Frequency Distribution of Top 500 Extractions

BNC Frequency Bands	Freq, Dice	Cosine	CSM	LL	Chi2, Yates	MI	McNemar
1,000	92.4	65.8	59.8	34.8	29.8	2.2	
2,000	6.6	12.0	20.8	16.0	13.8	4.2	
3,000	0.8	6.8	10.2	13.0	11.6	7.2	
4,000	0.2	5.0	5.0	9.4	9.0	8.0	
5,000		3.4	2.2	6.0	6.4	10.0	3.6
6,000		3.4	1.8	7.0	7.6	12.2	30.2
7,000		2.4	0.2	6.0	6.8	12.2	20.8
8,000		0.6		3.0	4.6	10.8	12.2
9,000		0.4		2.6	4.0	12.4	12.4
10,000				1.0	2.8	8.2	8.2
11,000				1.0	2.0	4.6	4.6
12,000		0.2		0.2	1.4	3.8	3.8
13,000						3.0	3.0
13,994					0.2	1.2	1.2

🟪 > 20% ▩ 10-20% 🟦 2-10%

Table 2 shows clear graduations of frequency levels. Looking across the table from *Freq* to *McNemar*, the top 500 words belong to increasingly lower frequency bands. The majority of the top 500 words from *Freq* and *Dice* belong to the top 1000 BNC. About 80% of *Cosine* and *CSM* belong to the top 2000 BNC. About 80% of *LL*, *Chi2*, and *Yates* words belong to top 5000 BNC or top 6000 BNC. Uniquely, *MI* extracts words evenly from all the frequency bands of BNC HFWL words and about 80% belong to the top 9000 BNC. Interestingly, *McNemar* extracts words only from ranks of lower frequency than BNC rank 5001, and 80% are extracted from ranks 5001 to 10000. The nine statistical measures clearly extract different outstanding levels of applied science words.

3. At what U.S. grade level are the top 500 words extracted by each method understood?

To understand grade level definitions for these extracted words, we examined the top 500 extractions for word familiarity by native English speaking children. Using *The Living Word Vocabulary* (Dale and O'Rourke, 1981), which is "an inventory of the written words known by children and young people in grades 4, 6, 8, 10, 12, 13, and 16," (p. vii) we determined at what grade level the majority of native English speaking students would readily understand the central meaning of each word in the top 500 extractions produced by the nine statistical measures. The results are shown in **Figure 1**. 'N/A' denotes the words not appearing in *The Living Word Vocabulary*. Most of N/A words are technical words such as *colonic*, *ulcerative*, and *mucosal*.

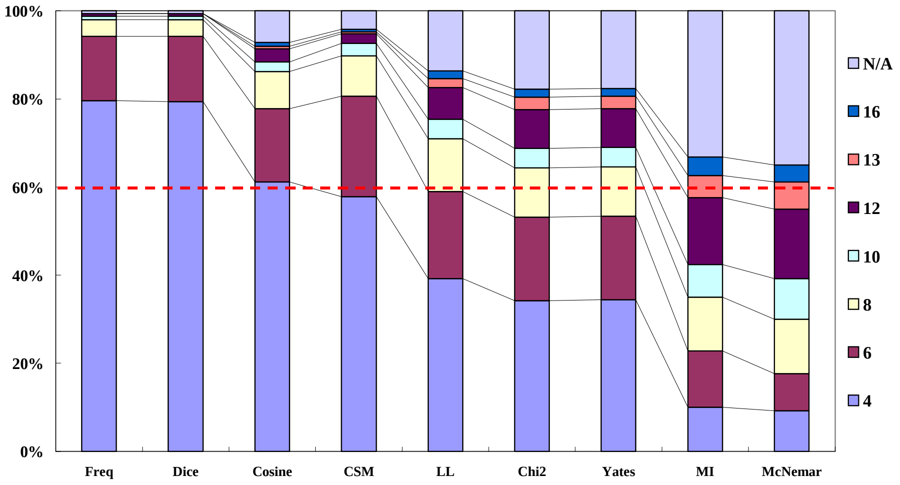

Figure 1. U.S. Grade Level Based on Word Familiarity

The grade at which 60% of extracted words are understood can be used to clarify the grade level comparisons: 60% of the top 500 words from *Freq*, *Dice* and *Cosine* are understood by 4th grade students, those of *CSM* are understood by 6th grade students, those of *LL*, *Chi2* and *Yates* are known by 8th grade students, those of *MI* and *McNemar* by 13th grade students. This confirms that each statistical measure identifies different grade levels of words.

4. What percentage of the top 500 words extracted by each method appear in Japanese junior and senior high school textbooks?

For this study to be meaningful in an EFL context, we must compare the vocabulary of the top 500 words to an EFL standard. We compared the extracted words to the vocabulary learned by Japanese students. JSH text coverage is one way to obtain an accurate estimate of the vocabulary level of each extraction, which is crucial information to EFL learners. For EST learners who want to acquire applied science vocabulary, the percent of words that were not covered by the JSH textbook vocabulary, represented by the upper section of each bar in **Figure 2**, may be important information. **Figure 2** graphically illustrates that while only 11% of the *Freq/Dice* top 500 extractions are not covered in the JSH school textbooks, 32% of the *Cosine*, 36% of the *CSM*, 58% of the *LL*, and 63% of the *Chi2/Yates* extractions are not covered, and 94% of the *MI* and 97% of *McNemar* extractions are not covered in the JSH school textbooks. Overall the bar graph covered by the JSH textbook is similar to the bar graph of U.S. 4th grade in **Figure 1**. The data in **Figure 2** again show that the nine different statistical measures extract words of quite different grade levels.

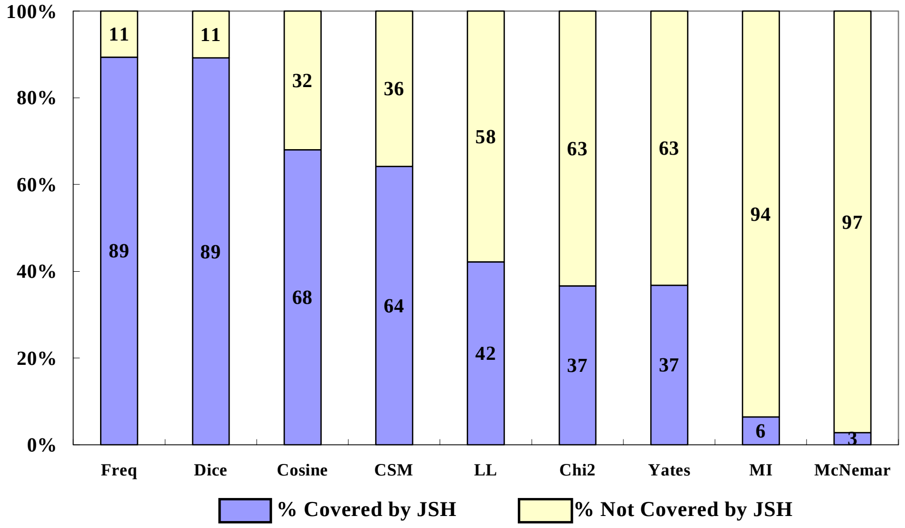

Figure 2. Percentage of Top 500 Words Covered / Not Covered by the JSH Textbook

CONCLUSION

In this study we applied nine statistical measures to the BNC applied science corpus and all of the data show that the statistical measures we used tend to extract specialized vocabulary belonging to certain frequency bands and grade levels. Our results were similar to those of prior studies based on a 100,000-word specialized corpus (Chujo and Utiyama, 2004, and Utiyama et al., 2004). Our study in combination with these previous studies shows that the results of the statistical measures on corpora are quite similar even if the examined corpora sizes are different.

The obvious pedagogical implication is that these statistical tools can very effectively be used to extract various types of specialized lists that can be quickly and accurately targeted to learners' proficiency levels. For example, we can infer that the basic applied science words extracted by *Cosine* and *CSM* would be good for beginning-level scientific English learners, the *LL/Chi2/Yates* lists would be suitable for intermediate-level scientific English learners, *MI* and *McNemar* would be appropriate for advanced-level scientific English learners, and *Freq* and *Dice* might be useful for applied science students who need to consolidate JSH vocabulary while learning basic scientific words. Thus we can use multi-level applied science lists that would supplement and bridge the big gap between EGP and EST in a graduated, step-by-step fashion.

These statistical tools can help teachers and material writers to select sub-technical or technical vocabulary without much specialist knowledge. Using extracted lists, they can easily manually delete less relevant candidates. Further research will include developing these extracted specialized lists into e-learning materials for vocabulary building.

REFERENCES

Chujo, K. and Genung, M. (2003). Vocabulary-level assessment for ESP texts used in the field of industrial technology. *English Teaching, 58*(3), 259–274.

Chujo, K. (2004). Measuring vocabulary levels of English textbooks and tests using a BNC lemmatised high frequency word list. In: J. Nakamura, N. Inoue, and T. Tabata (Eds.), *English corpora under Japanese eyes* (pp. 231–249). Amsterdam: Rodopi.

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.

Church, K. W. and Hanks, P. (1989). Word association norms, mutual information, and lexicography. *Proceedings of ACL-89*, 76–83.

Dale, E. and O'Rourke, J. (1981). *The living word vocabulary*. Chicago: World Book-Childcraft International, Inc.

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

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.

Justeson, J. and Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering, 1*, 9–27.

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.

Nelson, M., (2000). A corpus-based study of business English and business English teaching materials. Unpublished Ph.D. thesis, University of Manchester, Manchester, England.

Rayner, J. C. W. and Best, D. J. (2001). *A contingency table approach to nonparametric testing*. New York: Chapman & Hall/CRC.

Scott, M. (1996/1999). WordSmith Tools [Computer software]. Oxford: Oxford University Press.

Sutarsyah, C., Kennedy, G., and Nation, P. (1994). How useful is EAP vocabulary for ESP? A corpus-based study. *RELC Journal*, *25*, 34–50.

Takefuta, Y., Hasegawa, S., and Chujo, K. (1994). Goi list 'Gendaieigo no Keyword' no nin'chi level niyoru kubun no datousei (Validity of cognitive level grading for Keyword System 5000). *Working Papers in Language and Speech Science*, *4*, 53–63.

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.

Wynne, M. (1996). A post-editor's guide to CLAWS7 tagging. Retrieved August 31, 2005, from http://www.comp.lancs.ac.uk/computing/users/eiamjw/claws/claws7.html.