Extracting Level-Specific Science and Technology Vocabulary from the Corpus of Professional English (CPE)

Kiyomi Chujo,[1] Masao Utiyama[2] and
Takahiro Nakamura[3]

Abstract

With rapid advances in technology come rapid advances in the language of technology, or English for Science and Technology (EST). We have had success in our earlier research in devising a systematic means of extracting level- and domain-specific words from the British National Corpus. In this study, we apply a similar methodology to the Corpus of Professional English (CPE), a 20-million-word computerized database of English used by professionals in science and technology in twenty-two domains such as biology, chemistry, engineering, mathematics, medicine, and physics. This study describes the procedure for extracting multi-level EST vocabulary from the CPE by using nine statistical measures, and an examination of the top 500 most outstanding words produced by each statistical application by grade level and EST dictionary entry word coverage.

1. Introduction

With rapid advances in technology come rapid advances in the language of technology, or English for Science and Technology (EST). 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. EST courses are usually 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. Consequently, technical

[1] College of Industrial Technology, Nihon University
e-mail: chujo@cit.nihon-u.ac.jp
[2] National Institute of Information and Communications Technology
e-mail: mutiyama@nict.go.jp
[3] Shogakukan Inc.
e-mail: takahiro@shogakukan.co.jp

articles from professional journals are often used in lieu of a textbook.

One of the prominent aspects of the linguistic knowledge needed to comprehend specialized texts is the corresponding specialized vocabulary or "technical words that are recognizably specific to a particular topic, field, or discipline" (Nation, 2001: 198). EST vocabulary development is essential in order to achieve proficiency in EST, and current research is now turning to developing EST vocabulary lists for teachers, students, and researchers.

2. Review of the literature

2.1 EST word lists and corpora

Although various word lists, such as *A General Service List of English Words* (GSL) (West, 1953) and the *Academic Word List* (Coxhead, 2000) have been developed for assisting vocabulary learning, as Kuo (2007) notes, not many EST word lists have been compiled.

Flood and West (1953) devised a 'supplementary scientific and technical vocabulary' to the GSL. It is a definition list of 425 words "expressing all ordinary scientific and technical subjects within the limits of non-specialized study" (West, 1953: 583). Knight and Bethune (1972: 505) compiled a list of 277 'science words students know' to provide teachers "evidence of the recognition of selected science words by students in grades 7, 8, 9, 10, 11, and 12." Puangmali (1976) collected 500-550 words from ten engineering texts (totaling 5,202 words) to compile a 1,148-different-word engineering English vocabulary list.

As corpus linguistics has advanced, efforts to create larger EST corpora which cover more EST domains have been made. Some examples are the Jiaotong Daxue Computer Corpus of Texts in English for Science and Technology (JDEST), which "comprises 2000 texts of about 500 words each --- one million words in all" (Yang, 1985: 24-25), the Student Engineering English Corpus (SEEC) of nearly two million words built to "represent the engineering lexis encountered in English-language textbooks in basic engineering disciplines" (Moudraia, 2004: 139), and the Scientific Research Word List (SRWL) of 420 word families compiled from a corpus of 1.84 million words consisting of 360 journal articles in ten scientific research fields by Kuo (2007).

The largest current EST corpus is the Corpus of Professional English (CPE) which was compiled by the Professional English Research Consortium (PERC) established in 2002. The CPE consists of a 20-million-word computerized database of English used by professionals in science and technology in twenty-two domains such

as biology, chemistry, engineering, mathematics, medicine, and physics. Based upon the 2001 Journal Citation Reports, 3,739 files of 50,000 words per text were collected from the top 20 percent of the journals listed in terms of impact factor rating within each category[4]. The CPE corpus is still in development, but fortunately, the first phase trial version provides important language data.

2.2 Selecting EST words

We know from Sutarsyah *et al.* (1994) that selecting specialized vocabularies by using the traditional criteria of *frequency* and *range* is only partly successful. Because the focus of these measures is ranking general-purpose vocabulary in order of priority, separating specialized 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 may not have.

A number of corpus-based studies have used certain statistical measures to identify specialized vocabulary. For example, Nelson (2000) used the *log-likelihood* statistic from WordSmith Tools to find words that are statistically more frequently used in business English than in general English by comparing each word's frequency in a business English corpus with its frequency in the British National Corpus (BNC). Chujo and Utiyama (2004) and Utiyama *et al.* (2004) established an easy-to-use tool employing nine statistical measures to identify level-specific, domain-specific words from a corpus. 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 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. the *log likelihood ratio* (LLR) identified intermediate-level specialized words, and *mutual information* (MI) identified advanced level specialized words. These measures were effective in separating specialized vocabulary from general-purpose vocabulary, and provide a useful template as a means of identifying EST vocabulary. A close examination of the extracted words, however, shows that the applied science domain of the BNC gives much more weight to computer science technology and the medical field. Clearly we need a corpus representing a broader range of contemporary EST domains.

[4] The impact factor is a measure of the frequency with which an 'average article' in a journal has been cited in a particular year or period.
http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor/

In summary, a review of the literature shows that there is a need for an EST vocabulary list based on a more representative corpus. Secondly, it is possible to identify EST vocabulary by using statistical measures such as the *LLR* and *MI*. Finally, the 20-million-word CPE provides an excellent basis for the development of this kind of EST list.

3. Purposes of the Study

The purposes of this study are (1) to extract various levels of EST words by applying nine statistical measures to the 20-million-word CPE; (2) to identify the proficiency level of the EST words based on U.S. native speaker grade level; and (3) to determine the effectiveness of these measures by comparing the extracted words to an existing EST vocabulary list.

4. Procedure

4.1 The Data

4.1.1 CPE Master List

The CPE corpus has twenty-two sub-corpora shown in Table 1[5]. The procedure for preparing the CPE master list for statistical application is as follows.

Agriculture	Food Science
Biology	Forestry
Chemistry	General Science
Civil Engineering	Materials Science
Computer Science	Mathematics
Construction & Building Technology	Medicine
Earth Science	Metallurgy & Metallurgical Engineering
Electrical & Electronic Engineering	Nuclear Science & Technology
Engineering	Oceanography
Environmental Sciences	Physics
Fisheries	Telecommunications

Table 1: The Twenty-two CPE Domains

[5] Permission to use the first trial version of the CPE was granted by PERC. It is anticipated that these sub-corpora will be revised and expanded and the CPE corpus will be released in the near future.

We first created a lemmatized list from the CPE corpus using the CLAWS7 tag set[6] to extract all base forms. Next, if a word appeared fewer than 100 times in the corpus, it was deleted. Thus, all unusual or infrequent words were eliminated. Finally, all proper nouns and numerals were identified by their part of speech tags and deleted manually since statistical measures mechanically identify these words as technical words (Scott, 1999). This process yielded a 4,467-word CPE master list, representing 13,527,303 words.

4.1.2 Control Lists

We wanted not only to extract EST words but also to understand the proficiency level for these words, based on [U.S.] native speaker grade level. Furthermore, we wanted to evaluate how effectively the extraction was done in terms of identified EST words, and if so, to what extent. For the purpose of comparison, control vocabulary lists were created using the same procedures described above in 4.1.1. These control lists are described in detail below.

(1) The British National Corpus High Frequency Word List (hereafter BNC HFWL) is a list of 13,994 lemmatized words representing eighty-six million BNC words that occur 100 times or more. (For the compiling procedure, see Chujo, 2004). It was used for comparison to statistically determine how each of the domain-specific words in our CPE master list would appear differently from words in a general 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 [U.S.] students' grade levels four through sixteen. For supplementing grade levels one through three, 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 more recent data would be desirable, however, to our knowledge, no such data is available.

(3) *The Concise Illustrated Dictionary of Science and Technology* (Gibilisco, 1993: viii) includes "more than 5,500 of today's most commonly used scientific and technical terms." It is "designed especially for high school and college students who need accurate information on a broader range of subjects." In this study, we used 5,640 entries as an existing technical vocabulary control list to evaluate how effectively nine statistical tools extracted EST words. Using a larger (5,000-10,000

[6] http://www.comp.lancs.ac.uk/computing/users/eiamjw/claws/claws7.html

word) and more recent publication would be ideal; however, other publications, such as *Chambers Dictionary of Science and Technology* (Walker, 1999) containing over 50,000 entries, are too large for our limited resources. Within the confines of this study, it was not possible to create a control list from such a massive database.
(4) West's (1953) *A General Service List of English Words* containing the 2,000 most basic words from the *Interim Report on Vocabulary Selection* (Faucett *et al.*, 1936) that are considered necessary for learning English as a foreign language was compared to the extracted lists, as was the *Function Words* from Nation (2001: 430-431) containing 320 words. Function words express grammatical relationships with other words within a sentence. They may be prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles or particles.

4.2 Identifying Outstanding EST Words

4.2.1 Statistical Measures

We used nine statistical measures: simple *frequency* (*Freq*), the *Dice coefficient* (*Dice*) and *Cosine* (*Cosine*) (Manning and Schütze, 1999), the *complementary similarity measure* (*CSM*) (Wakaki and Hagita, 1996), the *log likelihood ratio* (*LLR*) (Dunning, 1993), the *chi-square test* (*Chi2*) and *chi-square test with Yates's correction* (*Yates*) (Hisamitsu and Niwa, 2001), *mutual information* (*MI*) (Church and Hanks, 1989), and *McNemar's test* (*McNemar*) (Rayner and Best, 2001)[7]. A detailed description of each measure can be found in Utiyama *et al.* (2004) and Chujo and Utiyama (2006), and the notation for these kinds of statistics can be found in Scott (1997).

4.2.2 Statistical Application

These statistical measures are widely used in computational linguistics. They automatically identify the most outstandingly frequent words in a given specified list (in this case, the CPE master list) by comparing each word's frequency with a control list (the BNC HFWL). In other words, 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 a word, 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 CPE word list with the frequency of the

[7] The formula for each measure is available on the web at http://www5d.biglobe.ne.jp/~chujo/ or http://www2.nict.go.jp/jt/a132/members/mutiyama.

same word in the BNC HFWL.

Using each measure, the statistical score for the extent of each word's "outstanding-ness" (Scott, 1999) in frequency of occurrence is computed. Since each measure uses a different formula, it gives a different score to each word. 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 EST word list, but it is not meant to be a definitive list. These statistical tools can help users to select technical vocabulary automatically without a great deal of specialist knowledge. With the extracted lists, users can easily manually delete irrelevant words.

5. Results and Discussion

5.1 Identifying the Vocabulary Levels of the Extracted Lists

5.1.1 Top 50 Extracted Words Overview Comparison

The top 50 words from each of the nine measures in descending order are shown in Table 2. Since the top 50 extractions made using *Freq* and *Dice* and for *Chi2* and *Yates* were almost the same, this data appears as two columns (*Freq/Dice* and *Chi2/Yates)* rather than as four separate columns. A glance at the top 50 words gives a brief overview of the general tendencies inherent in the extraction of the nine measures.

The lists in Table 2 are different from each other even though they were extracted from the same corpus. The top 50 words identified by *Freq* and *Dice* include general vocabulary such as *the* and *of* that usually appears at the top of high frequency lists, as well as some essential EST words more common in this field than others such as *result*, *cell*, *model*, and *data*. For *Cosine* and *CSM*, the top 50 extractions include many essential EST words such as *temperature, concentration* and *surface*. For *LLR*, *Chi2* and *Yates*, the top 50 extractions include a greater number of EST words. Both the *MI* and *McNemar* lists identify higher-level EST words. *MI* extracted important advanced level EST words such as *aqueous, oxidation, substrate* and *biomass*. McNemar extracted lower frequency words such as *seabed, deuterium*, and *angina*. Overall, word length and word difficulty appear to increase from left to right. This will be explored in the following sections.

	Freq / Dice	Cosine	CSM	LLR	Chi2 / Yates	MI	McNemar
1	the	the	the	of	cell	aqueous	seabed
2	of	of	of	cell	of	oxidation	deuterium
3	be	be	in	the	model	substrate	angina
4	and	and	be	model	data	biomass	high-performance
5	in	in	and	data	sample	latency	buckle
6	a	a	use	sample	temperature	lipid	hyperplasia
7	to	to	this	result	the	purify	canine
8	for	for	for	use	result	deformation	pipette
9	that	this	by	temperature	protein	wild-type	infestation
10	with	with	with	protein	concentration	anterior	dialysis
11	this	by	show	value	value	vesicle	pep
12	by	cell	cell	concentration	use	theorem	accretion
13	have	use	result	show	surface	spectral	aerodynamic
14	from	model	from	study	show	conductivity	estrogen
15	as	data	model	surface	study	computation	idealize
16	on	result	data	degree	parameter	tensile	enumeration
17	at	that	study	parameter	degree	fluorescence	catheter
18	use	from	value	observe	observe	droplet	plankton
19	it	show	increase	analysis	phase	incubate	annihilation
20	we	sample	sample	increase	decrease	posterior	globular
21	or	study	between	phase	analysis	algorithm	mackerel
22	can	value	high	in	experiment	decomposition	rout
23	not	temperature	effect	obtain	obtain	shear	suction
24	which	at	low	decrease	equation	axial	mononuclear
25	show	as	temperature	experiment	solution	exponential	unperturbed
26	they	concentration	each	method	increase	iteration	mimic
27	result	protein	system	solution	method	calibration	knockout
28	between	increase	surface	effect	function	amplitude	differentiated
29	cell	surface	degree	equation	acid	silica	surrogate
30	study	degree	analysis	function	effect	electrode	uppermost
31	also	analysis	concentration	low	low	modulation	dryness
32	may	effect	method	acid	particle	solute	prostate
33	model	observe	protein	sequence	respectively	acetate	elliptical
34	time	obtain	obtain	determine	interaction	diffusion	luminal
35	data	low	function	particle	sequence	kinase	woody
36	all	high	table	respectively	correspond	phenotype	crossover
37	than	between	observe	interaction	experimental	respiration	aforementioned
38	high	method	process	high	algorithm	oscillation	dyslexia
39	many	phase	test	correspond	determine	lattice	quadrant
40	increase	parameter	rate	indicate	density	viscosity	conceptually
41	will	on	also	table	in	selectivity	silt
42	value	experiment	present	reaction	component	activation	condense
43	much	solution	number	component	gene	coefficient	postoperative
44	system	function	solution	experimental	measurement	reactivity	shale
45	other	decrease	level	gene	reaction	dispersion	keyhole
46	but	each	than	algorithm	layer	precipitation	simulator
47	each	or	phase	density	indicate	spectroscopy	laterally
48	effect	equation	different	measurement	ratio	transverse	meteorite
49	only	system	experiment	flow	flow	simulation	graft
50	low	table	determine	test	distribution	denote	diurnal

Table 2: A Comparison of the Top 50 Words for Each Measure

5.1.2 Top 500 Word Grade Level Comparisons

Next, in order for these word lists to be useful pedagogically, we wanted to determine their proficiency levels. This was done by investigating at what U.S. grade level these

words would be understood by native English speaking (NS) children. In 1981, Dale and O'Rourke published *The Living Word Vocabular*y 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, Table 3 shows the average grade level at which NS students would readily understand the central meaning of each word for the top 500 extractions produced by the statistical measures. To the best of our knowledge, there is no similar data available for grades one through three, so for this comparison, we used reading grade word familiarity levels from Harris and Jacobson (1972). When we calculated the average grades of all 500 words, any words not appearing in either resource were labeled '17th grade'.

Measures	Average Grade
Freq	4.5
Dice	4.5
Cosine	6.2
CSM	6.1
LLR	7.6
Chi2	7.8
Yates	7.8
MI	12.0
McNemar	12.6

Table 3: U.S. Grade Level Based on Word Familiarity

In looking at the table, we can see that the top 500 words from *Freq* and *Dice* are generally understood by fourth grade students on average; those of *Cosine* and *CSM* are generally understood by sixth grade students; those of *LLR*, *Chi2* and *Yates* are generally known by seventh grade students; those of *MI* and *McNemar* are generally known by twelfth grade students.

In terms of practical pedagogical application, we inferred from this result, in addition to several of our previous similar studies (Chujo and Utiyama, 2005; Chujo and Utiyama, 2006; Chujo, Utiyama and Oghigian, 2006; Chujo *et al*., 2007) that (1) the EST words extracted by *Freq*, *Dice*, *Cosine* and *CSM* might be most useful for selecting EST words for beginner level EST students; (2) the *LLR*, *Chi2* and *Yates* lists might be most useful for intermediate level EST learners; and (3) the *MI* and *McNemar* vocabulary might be most appropriate for advanced level EST learners.

5.1.3 EST Dictionary Entry Word Overlap and Basic Vocabulary Overlap

Finally, to evaluate how effectively these tools extracted EST words, they were compared to an existing EST vocabulary control list. For this we used the 5,640 entry words in the *EST Dictionary*. In Figure 1, the black bar in the graph shows the overlap in percentage between the EST dictionary entries and the top 500 words. We also compared the top 500 words with the most often cited basic vocabulary, i.e. the GSL, and this overlap is indicated with grey bars. The overlap with function words is shown with white bars.

We can see the nine tools effectively produced relevant EST vocabulary. In particular, we see that there is about a 67 percent overlap between the *LLR, Chi2* and *Yates* extractions and the dictionary entries, which are written for students at the junior or senior college level. Considering the conciseness and limited scale of the EST dictionary entries we used in comparison with the divergent distribution of words in the twenty-two domains of CPE master list, an overlap of 67 percent is reasonable[8].

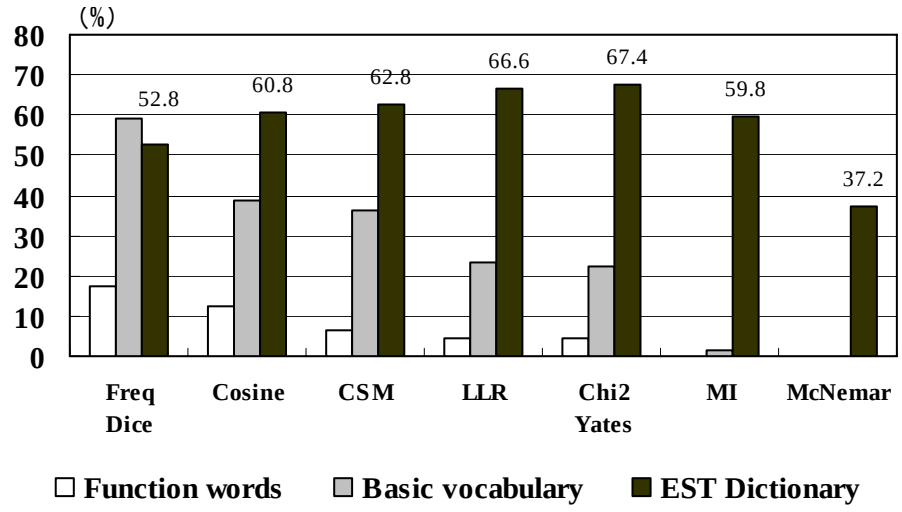

Figure 1: The Overlap of the Top 500 Extractions with the EST Dictionary Entries, Basic Vocabulary, and Function Words

Let's take a detailed look at the *Chi2/Yates* top 500 extractions, which show the highest overlap (67.4 percent). In Table 4**,** some of the examples from the top 500 words are shown. Words that are in the EST dictionary are underlined, function words

[8] In a previous study, when we compared the equivalent top 500 extractions from the BNC commerce component, which is less divergent than EST vocabulary, with business dictionary entries, we obtained an approximate 80 percent overlap (Chujo *et al*., 2007).

are in parentheses, and words in the GSL are italicized[9]. Essential EST words included in both the top 500 and the EST dictionary are words often used in ordinary scientific and technical subjects such as *absorb, absorption, abstract, accumulation, acid, activation, activity, addition, adsorption, algorithm, allele, alloy, amino, amplitude, analysis, angle, anion, anterior, antibody, antigen, application, approximation, aqueous, assay, atom, average, axial,* and *axis*. On the other hand, examples not covered by the concise dictionary entries are *abundance, activate, additional, analytical, analyze, approximate, approximately, associate,* and *assume*. Some of these words are derivations of the above-mentioned words. Others, particularly verbs such as *associate* and *assume*, are not included probably because of the size of the dictionary.

(*above*), absorb, absorption, abstract, abundance, accumulation, acid, activate, activation, activity, addition, additional, adsorb, adsorption, algorithm, allele, alloy, (*also*), amino, amplitude, analysis, analytical, analyze, (*and*), *angle* , anion, anterior, antibody, antigen, application, approximate, approximately, approximation, aqueous, assay, associate, assume, atom, *average* , axial, axis
bacterial, bacterium, *band* , bandwidth, *base* , baseline, (*be*), *beam* , behavior, (*between*), *bind* , biomass, blot, bond, (*both*), *boundary* , buffer, (*by*)
calculate , calculation, calibration, carbon, *case* , catalyst, cation, cavity, cell, cellular, *chain* , channel, characteristic, characterize, chemical, clinical, cluster, coefficient, column, combustion, *compare* , comparison, complex, component, composite, composition, compound, compression, computation, compute, concentration, *condition* , conductivity, configuration, *consider* , consist, consistent, constant, constraint, *contain* , *content* , continuous, *control* , core, correlate, correlation, correspond, criterion, crystal, *current* , *curve* , cycle, cytokine
data, decomposition, *decrease* , define, deformation, degradation, *degree* , demonstrate, denote, density, dependence, deposition, depth, derivative, derive, *describe* , detect, detection, detector, *determine* , deviation, diameter, differ, *difference* , *different* , diffraction, diffusion, dimension, dispersion, displacement, distribution, domain, dose, droplet, (*during*), dynamic
(*each*), *effect* , *elastic* , electrode, electron, element, embryo, emission, encode, energy, enzyme, equation, equilibrium, error, estimate, estimation, evaluate, excitation, exhibit, *experiment* , *experimental* , exposure, expression, extract

Table 4: An Excerpt of *Chi2/Yates* Top 500 Extractions

In this top 500 list, 4.4 percent are function words (shown in parentheses) and when removed from the above list, the remaining list seems to comprise essential

[9] Since the GSL is based on 'word families' units, and this study is based on 'lemma' units, there might be some discrepancy in counting and comparing words.

basic EST words. From this result, we might conclude that the tools extracted most commonly used EST words as well as basic EST words also used in general domains. The *Yates* is only one instance, however, and a similar breakdown of the extracted words was observed for the *Cosine*, *CSM*, and *LLR* extractions.

As for *Freq/Dice*, although basic EST words such as *acid, activation, activity, age, algorithm, analysis, animal, antibody, application, area, atom,* and *average* were identified, more function words such as *a, about, above, after, all, along, also, although, among, and, another, any, as,* and *at*, and more basic vocabulary such as *account, add, age, allow, amount, animal, appear, apply,* and *average* are also identified. As mentioned earlier, the goal in using these measures is to narrow the number of candidates for the EST word list. By using these extracted lists, users can easily manually delete irrelevant words. Although *Freq/Dice* have extracted beginner level proficiency EST words, removing basic vocabulary and function words might be time consuming.

Looking at the right most bars in Figure 1, we can see *MI* includes very few basic words and *MI* and *McNemar* include no function words. Bearing in mind the results from Tables 1 and 2, and Figure 1, *MI* seems appropriate for extracting higher proficiency level EST words. A close examination of the top 50 *McNemar* words in Table 2 turns up some puzzling entries such as *mackerel, unperturbed, knockout* and *woody*. This underscores the idea that the extracted lists are not meant to be definitive and that educators can use these as a starting point for crafting lesson-specific vocabulary. It also shows the importance of identifying words by domain, which will be discussed in the next section. We'll need a more thorough investigation of these higher or more advanced level extractions.

6. Conclusions and Pedagogical Applications

In this study, EST words were identified from the CPE using nine statistical measures. The identified EST words are grouped into three proficiency levels by [U.S.] grade level and were examined for overlap with EST dictionary entries. Users can further refine the candidates in the extracted lists based on their own appropriate contexts and levels.

Further research is aimed at developing these specialized vocabularies into e-learning materials for vocabulary building. Once these types of EST lists are created and refined, concordancing programs can be used either by educators or students to confirm the meaning of words and find real phrases containing the words. PERC is preparing to release the CPE using an online multi-functional search service called the

Shogakukan Corpus Network, which is also adopted for the online BNC search service and the online CobuildDirect search service developed by the Shogakukan multimedia department. Once available, concordance lines and collocation tables can be obtained easily and examples are shown in Figures 2 and 3.

One pedagogical application might be to create an intermediate EST vocabulary list from the *LLR* or *Chi2/Yates* data, and an advanced level EST vocabulary list from the *MI* data. Samples of words and phrases with Japanese translations for the intermediate and advanced levels are shown in Tables 5 and 6, respectively. In this way, students are presented with vocabulary in chunks, or partial contexts, rather than just having a list of words.

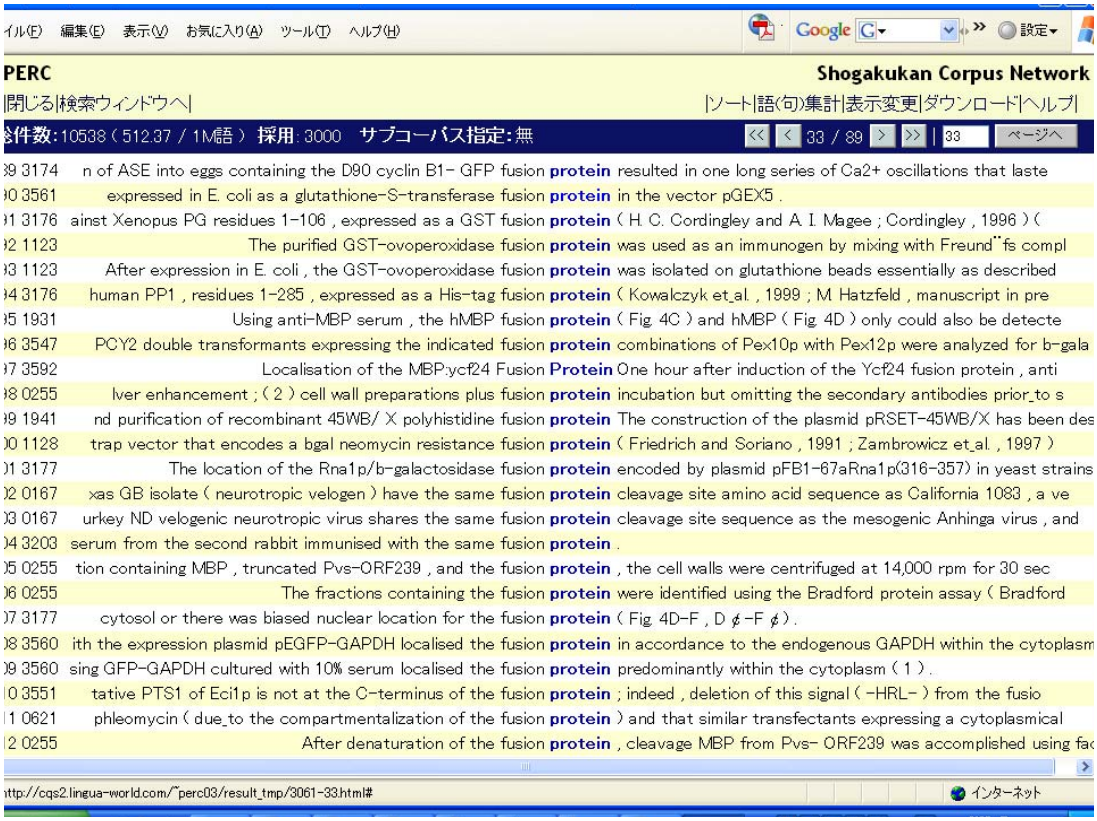

Figure 2: Concordance Lines Showing 'Protein'

Figure 3: Collocation Table Showing 'Protein'

word		Phrase 1		Phrase 2	
equation	式	differential equations	微分方程式	an integral equation	積分方程式
gene	遺伝子	gene products	遺伝子産物	a gene family	遺伝子族
density	密度	the energy density	エネルギー密度	cell density	細胞集密度
layer	層	the surface layers of the metal	金属の表層	the boundary layer	境界層
measurement	測定(値)	temperature measurements	温度測定	a direct measurement of the winds	風の直接測定
membrane	膜	the cell membrane	細胞膜	the inner membrane	内膜
parameter	パラメータ	parameter values	パラメータ値	model parameter	モデルパラメータ
particle	粒子	metal particles	金属粒子	the average particle size	平均粒度
protein	タンパク質	protein synthesis	タンパク質合成	protein structure	タンパク質構造
ratio	比率	a signal/noise ratio	信号対雑音比	sex ratio at birth	誕生時の性比

Table 5: An Example of Intermediate Level EST Vocabulary

word		Phrase 1		Phrase 2	
activation	活性化	he activation energy	活性化エネルギー	cell activation	細胞の活性化
algorithm	アルゴリ ズム	an efficient algorithm for solving the problem	問題解決の効率的なアルゴリ ズム	algorithm development	ｱﾙｺﾞﾘｽﾞﾑ開発
coefficient	係数	the correlation coefficient	相関係数	the coefficient of variation	変動係数
gel	ゲル	silica gel	シリ カゲル	collagen gel	コラーゲンゲル
gradient	勾配，傾き	a temperature gradient	温度勾配	a pressure gradient	圧力勾配
neuron	ニューロン	sensory neurons	感覚ニューロン	motor neurons	運動ニューロン
oxidation	酸化	oxidation reactions	酸化反応	the oxidation level	酸化レベル
simulation	シミュレーション	simulation results	シミュレーション結果	computer simulation	計算機シミュレーション
substrate	基質	substrate binding	基質結合	substrate recognition	基質の認識
thermal	熱の	thermal dynamics	熱力学	thermal energy	熱エネルギー

Table 6: An Example of Advanced Level EST Vocabulary

Furthermore, it is possible to provide additional information such as in which domain a particular word most often appears (see Table 7), or to develop lists according to the domain (Table 8). It is clear that there exist a number of potential pedagogical applications, which is important when we consider that there are many kinds of learners. At the university level in Japan, the reality is that although students master a senior high school vocabulary of approximately 2,000 to 3,000 words, many freshmen and sophomore students need to review and practice to consolidate their knowledge and to bridge the gap between their understanding of these words and their successful usage of these words. Thus, some institutes may wish to include basic words with a gradual introduction of domain-specific words, where others might find it more useful to provide a cross-section of domains. In any event, the CPE has shown itself to be an invaluable resource, and the applications of various statistics as demonstrated in this study provide a powerful tool for educators in mining the CPE for specialized vocabulary. The addition of concordancing and collocation tools provides an important context and allows learners to understand how these types of words appear in and function in the language.

Word	Appearing Frequently in These Domains		
activation	① Biology	② Medicine	③ Food Science
algorithm	① Computer Science	② Electrical & Electronic E.	③ Telecommunications
coefficient	① Construction & Building T.	② Mathematics	③ Civil Engineering
gel	① Food Science	② Biology	③ Materials Science
gradient	① Forestry	② Oceanography	③ Construction & Building T.
neuron	① Medicine	② Biology	③ Agriculture
oxidation	① Chemistry	② Metallurgy	③ Earth Science
simulation	① Forestry	② Telecommunications	③ Nuclear Science & T.
substrate	① Materials Science	② Electrical & Electronic E.	③ Food Science
thermal	① Materials Science	② Nuclear Science & T.	③ Earth Science

Table 7: An Excerpt of Words and Domains

	Biology	Electrical Engeneering	Food Science	Forestry	Materials Science	Medicine	Metallurgy
1	activation	approximation	amino	basal	axial	antigen	alloy
2	anterior	cavity	ethanol	biomass	compression	baseline	binary
3	assay	excitation	extraction	estimation	conductivity	dilute	deformation
4	nucleotide	interact	fatty	flux	degradation	elevate	diffusion
5	cleavage	neuron	gel	gradient	dye	inhibitor	displacement
6	hybridization	propagation	glucose	nutrient	mesh	mediate	fracture
7	incubate	quantum	lipid	physiological	modulus	neuron	geometry
8	induction	sensor	metabolism	regression	radius	parasite	lattice
9	intracellular	spectral	peptide	simulation	shear	subunit	simulate
10	peptide	wavelength	purify	uptake	substrate	vaccine	transient

Table 8: An Excerpt of Domain Specific Words

For more information on the various statistics used in this study, or for word lists created in previous studies, please see http://www5d.biglobe.ne.jp/~chujo/ and/or http://www2.nict.go.jp/jt/a132/members/mutiyama. For more information on Shogakukan, Inc., please see http://www.corpora.jp/.

References

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 M. Utiyama (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.

Chujo, K. and M. Utiyama (2005) Selecting Level-Specific BNC Applied Science Vocabulary Using Statistical Measures, in Selected Papers from the Fourteenth International Symposium on English Teaching, pp. 195–202. Taipei: English Teachers' Association/ROC.

Chujo, K. and M. Utiyama (2006) 'Selecting level-specific specialized vocabulary using statistical measures'. *SYSTEM, 34* (2), 255–69.

Chujo, K., M. Utiyama and K. Oghigian (2006) Selecting Level-Specific Kyoto Tourism Vocabulary Using Statistical Measures, in New Aspects of English Language Teaching and Learning, pp. 126–138. Taipei: Crane Publishing Company Ltd.

Chujo, K., K. Oghigian, C. Nishigaki, M. Utiyama and T. Nakamura (2007) 'Creating e-learning material with statistically-extracted spoken and written business vocabulary from the British National Corpus'. *Journal of the College of Industrial Technology Nihon University 40*, 1–12.

Church, K. W. and P. Hanks (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), 213–38.

Dale, E. and J. O'Rourke (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.

Faucett, L., H. E. Palmer, M. West, and E. L. Thorndike (1936) Interim Report on Vocabulary Selection. London: PS King.

Flood, W. E. and M. West (1953) Supplementary Scientific and Technical Vocabulary, in M. West (ed.) A General Service List of English Words, pp. 583–88. London: Longman.

Gibilisco, S. (1993) The Concise Illustrated Dictionary of Science and Technology. Blue Ridge Summit, PA: TAB Books.

Harris, A. J. and M. D. Jacobson (1972) Basic Elementary Reading Vocabularies. New

York: Macmillan.

Hisamitsu, T. and Y. Niwa (2001) ‘Topic-word selection based on combinatorial probability’. *NLPRS-2001*, 289–96.

Knight, D. W. and P. Bethune (1972) ‘Science words students know’. *Journal of Reading 15*, 504–506.

Kuo, Chih-Hua (2007) ‘Constructing a scientific research word list’. *JALT CALL 2007* presentation abstracts, p. 25, Waseda University.

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

Moudraia, Olga. (2004) ‘The student engineering English corpus’. *ICAME Journal 28*, 139–43.

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, Manchester: University of Manchester.

Puangmali, S. (1976) ‘A study of engineering English vocabulary’. *RELC Journal 1*, 40–52.

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

Scott, M. (1997) ‘PC analysis of key words and key key words’. *System 25*(2), 233–45.

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

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

Utiyama, M., K. Chujo, E. Yamamoto, and H. Isahara (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), 165–97.

Wakaki, M. and N. Hagita (1996) ‘Recognition of degraded machine-printed characters using a complementary similarity measure and error-correction learning’. *IEICE Trans. Inf. and Syst. E79-D*, 5.

Walker, M. B. (1999) Chambers Dictionary of Science and Technology. Edinburgh: Chambers Harrap Publishers Ltd.

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

Yang, H. (1985) ‘The JDEST computer corpus of texts in English for Science and Technology’. *ICAME News 9*, 24–25.