Corpus, ICT, and Language Education, (eds) G. Weir and S. Ishikawa.
University of Strathclyde Publishing, Glasgow, UK, 2010.

EVALUATING STATISTICALLY-EXTRACTED DOMAIN-SPECIFIC WORD LISTS

Kiyomi CHUJO
Nihon University
chuujou.kiyomi@nihon-u.ac.jp

Masao UTIYAMA
National Institute of Information and Communications Technology
mutiyama@nict.go.jp

Takahiro NAKAMURA
Shogakukan Inc
takahiro@shogakukan.co.jp

Kathryn OGHIGIAN
Waseda University
k_oghigian@aoni.waseda.jp

ABSTRACT

This study evaluates the efficacy of applying nine specific statistical tools (*Freq*, *Dice*, *Cosine*, *CSM*, *LLR*, *Chi2*, *Yates*, *PMI* and *McNemar*) to eight sub-domains of the BNC to create domain-specific single lexical vocabulary lists for particular proficiency levels, and examines the resulting lists for grade level, frequency level, and coverage for basic, function, and specialized words. It was demonstrated that it is possible to use particular measures to create lists for particular proficiency levels, that some measures extract a greater or lesser number of basic and function words, and that different measures extract a different number of specialized words.

1. INTRODUCTION

To become proficient in English, learners must expand their vocabulary (Nation, 2001). With advances in corpus linguistics, we now have large corpora and statistical tools to create our own, real-world vocabulary lists. We know that applying conventional *frequency* and *range* criteria to a corpus tends to extract general-purpose vocabulary (Sutarsyah et al., 1994) and is therefore of limited use in identifying technical words. Technical vocabulary appears outstandingly in a text or corpus of specific domains, but 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 a specialist knowledge of the domain (Zanchetta & Baroni, 2005; Chujo & Utiyama, 2006). Some automated means have been proposed to differentiate domain-specific vocabulary from high-frequency words, and low frequency items from rare items. What is needed at this stage is to evaluate how effectively these measures extract targeted-level, domain-specific vocabulary.

As suggested by Nation (2001, p. 18), 'one way of making a technical vocabulary is to compare the frequency of words in a specialized text with their frequency in a general corpus.' Some corpus processing tools such as WordSmith Tools (Scott, 1996) can identify words which occur with an unusually high frequency in a text or corpus when that text or corpus is compared with another corpus by using the *chi-square* (*Chi2*) or *chi-square with Yates's correction* (*Yates*) statistics. Other corpus processing tools are also equipped with these co-occurrence scoring tools. For example, Wmatrix (Rayson, 2005) uses *log-likelihood ratio* (*LLR*), and WordbanksOnline (HarperCollins Publishers Ltd, 2004) uses (*pointwise*) *mutual information* (*PMI*) and the *t* score. A number of corpus tools make use of statistics for the display of scores to build wordlists of genre-specific terms (Oakes, 1998, p. 196). As Lemay, L'Homme, and Drouin (2005, p. 232) pointed out, the tools cited here and studies cited in this paper are all based on the following two general assumptions. Firstly, '[t]he more frequently a unit appears in a corpus, the more likely it is to be significant in this corpus,' and secondly, 'a specialized corpus is a reflection of the use of lexical units in a given subject field, hence, "specific" units are likely to be terms.'

Putting these assumptions into practice, and using co-occurrence scoring tools, a number of corpus-based studies have proposed the use of certain statistical measures for comparing word frequency in technical and general corpora, and for identifying technical vocabulary (Chujo & Nishigaki, 2006; Nelson, 2000; Takefuta, 1981). Although these tools are often referred to as *automatic* term extraction tools, as Bowker and Pearson pointed out, 'these tools only suggest possible terms, and the list of proposed candidates must be verified by a [hu]man' (2002, p. 174). Oakes (1998, p. 177) suggested that simply deleting the poor candidates would be a much simpler task than creating the entire list manually. As more and more researchers use these tools (see for example Asmussen, 2005; Ooi et al., 2007), it is important to 'address the question of how to evaluate a number of co-occurrence-based scoring systems without relying on a '"feel" for the results' (Hardcastle, 2005, p. 1). Furthermore, it is important to clarify the efficiency of the extraction and to find the best statistical measure that gives a correct list of terms with the lowest rates of 'noise and silence' (Bowker & Pearson, 2002, pp. 231-232) by performing an evaluation of measures (Daille, 1995; Utiyama et al., 2004). While there have been natural language processing papers comparing statistical measures (Evert & Krenn, 2005; Wermter & Hahn, 2006), we present a pedagogical perspective with regard to grade level, frequency level, specialized word coverage, and type of domain-specific terms the statistical measures are extracting.

In summary, a review of the literature shows that it is possible to identify domain-specific vocabulary by using a wide range of statistical measures such as the *LLR* and *PMI*. However, for pedagogical application, it is important to identify the proficiency level of the resulting lists and to determine the effectiveness of these measures by comparing the extracted words to a large corpora and a greater number of existing control lists. In building on previous studies, this present study will compare statistically derived domain-specific wordlists from a large corpus, and seek to identify the grade level, frequency level, and specialist vocabulary coverage as well as basic vocabulary coverage in order to better understand the usefulness of each particular statistical measure. Specifically, the following questions were addressed:

1. In order to understand the proficiency level of the top 500 words produced by each measure, at what U.S. grade level are these words generally understood?
2. In order to understand the frequency level of the top 500 words produced by each measure, at what 1,000-word frequency band do they occur in a large general English corpus?
3. In order to understand how well each measure is able to extract domain-specific words, what percentage of the top 500 words from each measure appears in specialist dictionaries? What types of words do not appear in specialist dictionaries?

2. PROCEDURE

2.1 The Data

2.1.1 Creating a Master List from the BNC informative domain sub-corpora

The British National Corpus (BNC) is one of the largest electronically-accessible corpora consisting of over 100 million written and spoken words in British English (Leech et al., 2001). It consists of an approximately 90 million-word written component of informative and imaginative texts, and a 10 million-word spoken component. Seventy-five percent of the written component was drawn from informative texts, and the remaining 25 percent from imaginative texts. The informative domains shown in Table 1 are comprised of eight subject fields which are assumed to be specific for each domain.

Table 1. The British National Corpus informative domains

	Domain	Types	Tokens
1	Natural Science	2,334	2,806,175
2	Applied Science	3,357	5,548,470
3	Social Science	4,557	11,831,404
4	World Affairs	5,153	13,637,444
5	Commerce & Finance	2,973	5,883,249
6	Arts	3,017	4,993,728
7	Belief & Thought	1,677	2,258,188
8	Leisure	4,409	9,495,310

The procedure for preparing a domain-specific Master List for statistical application is as follows. To extract domain-specific vocabularies from each domain or sub-corpus, we first created a lemmatized list from each sub-corpus using the CLAWS7 tag set. We listed under a base form of a word or the same lemma, the inflectional variants and the declension of nouns, pronouns and adjectives. Next, all British spellings were changed to American spellings. In addition, if a word appeared fewer than 100 times in the sub-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) and 'they are of high frequency in particular texts but not in others,…and they could not be sensibly pre-taught because their use in the text reveals their meaning' (Nation, 2001, pp. 19-20). This process yielded each domain-specific Master List with the number of different words (types), representing the total number of words (tokens) in Table 1. It should be noted that this type of approach applied here will target only single-word lexical units (e.g. *sub-corpus*, *wordlist*) and variants such as compounds (e.g. *sub corpus* and *word list*) may be overlooked.

2.1.2 Control lists

While it has been established that it is possible to extract domain-specific words, their results can be evaluated from various viewpoints. From a pedagogical viewpoint, we are interested in knowing at what U.S. native speaker grade level they generally appear, and how frequently they occur in the entire BNC. Furthermore, we wanted to evaluate how effectively the extraction was done in terms of identified specialized words, and if so, to what extent. Each of these three methodologies, (grade level, frequency level, and technicality) is an important vocabulary selection criterion in the field of selecting vocabulary for designing a language syllabus. For the purpose of these comparisons, the following control vocabulary lists were used. These control lists are detailed below.

(1) The British National Corpus High Frequency Word List (hereafter BNC HFWL), is a list of 13,994 lemmatized words representing 86 million BNC words that occur 100 times or more. The BNC HFWL is the core of the BNC. It was used for comparison to statistically determine if and how each of the written domain-specific words in our master lists would appear differently from the words in a general corpus. It was created using the same procedures described above in 2.1.1. For the compiling procedure, see Chujo (2004).

(2) *The Living Word Vocabulary* (LWV) (Dale & O'Rourke, 1981) is useful for determining the (U.S.) grade level at which the central meaning of a word can be readily understood. Although somewhat dated, according to Hiebert (2005, p. 252) the LWV is the only comprehensive existing database on students' familiarity with word meanings. This wordlist includes more than 44,000 items and each presents a percentage score for those words or terms familiar to students in grade levels 4, 6, 8, 10, 12, 13, and 16. The *Basic Elementary Reading Vocabularies* (Harris & Jacobson, 1972), with 7,613 different words, is useful for determining the (U.S.) grade levels of reading vocabulary ranging from the first grade to the sixth grade. Since the data in the LWV begins from the fourth grade, the *Basic Elementary Reading Vocabularies* was used to determine the first, second, and third grade level correlations.

(3) *The Longman Business English Dictionary* (Pearson Education Limited, 2000) includes 'over 20,000 words and phrases' based on the analysis of millions of authentic business texts. In this study, we used only 'words' (totalling 4,565 entries) as an existing technical vocabulary control list to evaluate how effectively the statistical measures extracted business words. The inflectional and spelling variants were listed under a base form of a word, and all British spellings were changed to American spellings. Proper nouns and numbers were deleted.

(4) West's (1953) *A General Service List of English Words* (GSL) contains 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. It was compared to the extracted lists, as was the *Function Words* from Nation (2001, pp. 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. Since the GSL is based on its own word units and Nation's function words included contractions and conjugations for some verbs, and this study is based on 'lemma' units, some discrepancy might have occurred in counting and comparing words. Therefore the original GSL basic wordlist and Nation's function wordlist were reorganized into comparable lemma units and this reduced the lists to 1,867 basic words and 230 function words, respectively. As with the other resources, all British spellings were changed to American spellings.

2.2 Identifying Outstanding Domain-Specific Words

We used nine statistical measures: simple *frequency* (*Freq*), the *Dice coefficient* (*Dice*) and *Cosine* (*Cosine*) (Manning & Schütze, 1999), the *complementary similarity measure* (*CSM*) (Wakaki & Hagita, 1996), the *log likelihood ratio* (*LLR*) (Dunning, 1993), the *chi-square test* (*Chi2*) and *chi-square test with Yates's correction* (*Yates*) (Hisamitsu & Niwa, 2001), *pointwise mutual information* (*PMI*) (Church & Hanks, 1989), and *McNemar's test* (*McNemar*) (Rayner & Best, 2001). *Dice* and *Cosine* are statistics widely used to measure the similarity between collocations and between terms (Oakes, 1998). *CSM* is a similarity measure often used in optical character recognition. Different tools use different techniques in order to identify potential terms. A detailed description of each measure and their mathematical background, along with their strengths and weaknesses, 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), Oakes (1998) and Bowker and Pearson (2002). The formula for each measure is given in Appendix B.

In this paper, the statistically extracted rank-ordered lists produced by each measure are called domain-specific or specialized words/lists/vocabulary. We have used the broad definitions of technical vocabulary or specialist vocabulary; i.e. specialized lists contain three types of words: technical vocabulary, or words specific to a domain, sub-technical vocabulary, or words more common in a certain domain than elsewhere and less obviously technical vocabulary, and general vocabulary, which is a general base of English words. To illustrate this point, in the particular field of applied linguistics, the words *lemma* and *morpheme* would be examples of technical words; *type*, *token*, *range* and *frequency* would be sub-technical and *word* and *meaning* would be general words. (These examples are taken from Nation, 2001, p. 198.)

3. RESULTS AND DISCUSSION

3.1 Extracted Words Overview

In order to get an overview of what types of words each measure would extract, we began with just the top 15 words from each of the nine measures. Due to space limitations, here we show only the top 15 outstanding words from one of the eight sub-corpora, commerce/finance, in Table 2. As Asmussen (2005, p. 4) stated, 'generally, the designation of the domain proper is among the top fifteen types,' and these snapshots provide a clear and simple illustration of the types of the domain-specific words extracted by each measure. In addition, it should be noted that the top 15 extractions made using *Freq* and *Dice* and those for *Chi2* and *Yates* were almost the same, therefore this data appears as two columns (*Freq*/*Dice* and *Chi2*/*Yates*) rather than as four separate columns.

At first glance, the top 15 words shown in Table 2 clearly show the general tendencies inherent in the extraction of the nine measures. We can see that the lists are very different from one another even though they were extracted from the same master list. For example, words identified by *Freq* and *Dice* are general words that usually appear at the top of high frequency lists. In fact, all of the top 15 words extracted by *Freq* and *Dice* in Table 2 are function words. The top 15 words identified by *Cosine* and *CSM* include function words such as *the* and *of*, as well as some domain-specific words such as *company* and *market*. In the *LLR* list, we can see simple business words such as *market*, *company*, and *bank*. In *PMI* and *McNemar* lists, there are more complex words such as *lading*, *buyout*, *arbitrage*, *issuer*, and *subcontractor*. The bottom two rows of each column show the average frequency score and average word length of these 15 words. As we see from Table 2, the average frequency score decreases from left to right or from *Freq* to *McNemar*. On the other hand, the average word length increases from *Freq* to *McNemar*, ranging from 2.8 to 9.3 letters. Although we are aware that word difficulty may be influenced by many more factors than frequency and word length, the data in Table 2 suggests word difficulty increases as we move from left (*Freq*) to right (*McNemar*) for the top 15 words. This also suggests that specific statistical measures can be used to target specific grade level vocabulary.

To sum up, these nine measures seemed to predominantly extract not only some specific proficiency levels of words but also specific types of words. To confirm that this visual hypothesis holds true for the longer list, we will explore the grade level and frequency level of the top 500 words in sections 3.2 and 3.3 and we will examine the component distribution of the top 500 words in section 3.4. (For the justification for using 500 as a cut-off point, please see Utiyama et al., 2004.) Because the top 500 words for *Freq* and *Dice* and those for *Chi2* and *Yates* were almost the same, only seven columns are shown in the various tables.

Table 2. Top 15 commerce & finance domain-specific words

Rank	Freq/Dice	Cosine	CSM	LLR	Chi2/Yates	PMI	McNemar
1	the	the	the	market	market	lading	subcontractor
2	be	be	of	company	company	buyout	acquirer
3	of	of	be	bank	bank	long-run	payout
4	to	to	to	the	price	arbitrage	issuer
5	a	a	a	business	business	subcontractor	drafter
6	and	and	in	price	investment	stockmarket	no-arbitrage
7	in	in	will	rate	rate	offeror	long-run
8	that	for	for	cost	firm	drafter	shareholding
9	have	that	company	firm	cost	no-arbitrage	headhunter
10	it	have	market	tax	tax	shareholding	tax-free
11	for	will	by	investment	account	headhunter	buyout
12	they	company	or	account	the	payout	cross-border
13	on	market	business	share	profit	issuer	headhunting
14	will	it	this	profit	contract	liquidity	actuarial
15	this	by	may	contract	share	salesperson	stockmarket
Average frequency	152,838	148,062	128,069	42,795	42,795	203	134
Average Word length	2.8	3.1	3.4	5.6	5.6	8.9	9.3

3.2 Identifying the Grade Levels of the Top 500 Extractions

In order for these types of wordlists to be useful pedagogically, we wanted to determine their proficiency levels. This was estimated by investigating the number of known words at each grade level of the LWV and the reading grade word familiarity levels from Harris and Jacobson (1972).

Table 3 summarizes the average grade level of the top 500 extractions from each of the eight BNC domains. The average grade levels for each measure from the eight domains are shown at the bottom of the table and give us a general indication of proficiency level. Table 3 shows that the grade levels increase from *Freq* to *McNemar*, ranging from grade 3 to grade 10. This investigation of the grade level of each of these extracted vocabularies reveals that the potential exists for using specific statistical measures to target specific grade level vocabulary.

Table 3. Average grade level of the top 500 extractions for the eight BNC domains

Domain		Freq/Dice	Cosine	CSM	LLR	Chi2/Yates	PMI	McNemar
1	Natural Science	grade 3.5	6.0	5.5	7.2	8.2	10.2	10.6
2	Applied Science	3.3	5.2	5.2	7.5	8.3	11.6	12.0
3	Social Science	3.2	3.9	5.0	6.2	6.4	11.0	12.1
4	World Affairs	2.9	3.5	4.8	5.9	6.1	9.7	10.8
5	Commerce & Finance	3.3	4.3	5.0	6.1	6.5	9.0	10.2
6	Arts	2.7	3.4	4.2	5.8	6.1	7.7	8.7
7	Belief & Thought	2.9	4.2	4.7	5.7	5.8	6.8	7.3
8	Leisure	2.3	2.4	2.9	4.2	4.5	7.4	8.3
Average grade level per word through the eight BNC domains		grade 3.0	4.1	4.7	6.1	6.5	9.2	10.0

3.3 Identifying the Frequency Levels of the Top 500 Extractions

Furthermore, we examined the frequency distribution of the top 500 extracted words by using the BNC HFWL, which was divided into fourteen 1,000-word frequency bands of the most frequent words.

‘Vocabulary frequency level 1,000’ indicates ranks 1 to 1,000, ‘vocabulary frequency level 2,000’ indicates ranks 1,001 to 2,000, and so on.

Table 4 summarizes the average frequency level of the top 500 extractions from each of the eight BNC domains. As we can see from the table, each average frequency level produced by the measures for the eight domains varies slightly. In order to get a clear understanding of the graduation of the frequency levels, average frequency levels for the same measure from the eight domains are shown at the bottom of the table. Of course from Table 4, we can confirm the previous observation that the frequency levels increase from *Freq* to *McNemar*, ranging from the *Freq*/*Dice*, *Cosine* and *CSM* words under the 2,000-word level; *LLR* and *Chi2*/*Yates* words around the 3,000-word level; *PMI* at the 5,863-word level; and *McNemar* at the 6,916-word level. This observation across domains underscores the possibility that specific statistical measures can be used to target specific proficiency level vocabulary.

Table 4. Average vocabulary frequency level of the top 500 extractions from the eight BNC domains

Domain		Freq/Dice	Cosine	CSM	LLR	Chi2/Yates	PMI	McNemar
1	Natural Science	1,244	3,058	2,274	3,890	4,842	6,530	6,974
2	Applied Science	1,088	1,952	1,752	3,282	3,958	7,048	7,880
3	Social Science	1,022	1,152	1,366	2,046	2,200	6,532	8,166
4	World Affairs	1,046	1,254	1,678	2,534	2,710	6,750	8,744
5	Commerce & Finance	1,070	1,438	1,530	2,326	2,712	4,552	5,694
6	Arts	1,108	1,858	2,110	3,854	4,064	5,286	5,956
7	Belief & Thought	1,146	2,174	2,100	2,898	2,946	3,342	3,688
8	Leisure	1,052	1,364	1,700	2,930	3,276	6,866	8,228
Average frequency level per word through the eight BNC domains		1,097	1,781	1,814	2,970	3,339	5,863	6,916

3.4 Exploring the Technicality of the Top 500 Extractions

As we have seen in previous sections, the 500 domain-specific word extractions in the eight domains demonstrated a similar tendency according to each statistical measure in grade levels and frequency level. In this section, a closer investigation of the content of each domain-specific word extraction was conducted for the commerce/finance domain, in order to attempt to characterize the nature of each statistical tool for identifying domain-specific words.

The simplest way to implement this is to take each extracted domain-specific wordlist and compare it with three types of vocabularies: function words, basic words, and specialist words. Although there could arguably be one more category between basic words and specialist words, i.e. sub-technical words, there is currently no standardized means to make a clear distinction, and no clear agreement in the literature on defining a sub-technical word. (For a discussion on this, see Baker, 1988; Hutchinson & Waters, 1987; Justeson & Katz, 1995; Nation, 2001; or Utiyama & Chujo, 2007.)

The approach taken here was to count the number of overlapping words between the top 500 words extracted by each measure and each of the three different vocabularies as a certain indicator of technicality. For the specialist words, dictionary entries were used from the *Longman Business English Dictionary* for the commerce/finance domain. As mentioned in 2.1.2, we used the most cited basic vocabulary, *The General Service List of English Words* (GSL) (West, 1953), for basic words. For function words, we used the function words listed in Nation (2001).

Words on these three types of control lists partly overlap and are not mutually exclusive. For example, *company*, *market*, *business*, and *bank* in the top 15 commerce/finance extracted list belong to both the basic words and the specialist words categories. In addition, most function words belong to the basic words category. We wanted to clarify how the proportion of each word category (function, basic, and specialist) was distributed across the top 500 words extracted by each of the measures for this domain. In order to avoid double-counting the results, we considered that a specialist word is one that is recognizably specific to a particular topic, field or discipline (Nation 2001, p. 198). Therefore, we eliminated all the function words and basic words from the specialist words included in the business dictionary entries. For example, *company*, *market*, *business*, and *bank* only belong to the basic words. Then we discarded all the function words from the basic words category. At the end of this procedure we had three exclusive wordlists; a 3,865-word specialized list of business dictionary entries, a 1,732-word list of basic words, and a 230-word list of function words.

Next we compared the top 500 extractions with these three category wordlists to calculate their coverage percentage. Figure 1 shows how these three categories of words distribute across the top 500 commerce/finance extracted words. The overlap with function words is shown with dark grey sections at the lowest part of the bars. The next section on the bars shows the overlap with basic words. The next section, shown in grey, shows the overlap in percentage between the dictionary entries and the top 500 words. The percentages of words not appearing in any of these three categories are denoted by ‘other’ and are shown in the white section at the top of each bar. We will look at the coverage of each category of words individually in the following sections.

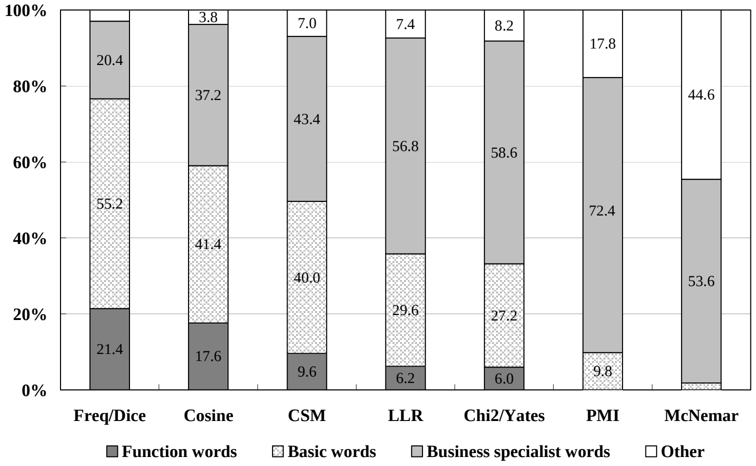


Fig 1. The proportion of the three categories of words in the top 500 extractions of the commerce/finance domain

3.4.1 Function word distribution

Asmussen (2005, p.5) used *LLR* for identifying domain-specific words for updating *The Danish Dictionary* and observed ‘frequent function words appearing so saliently ranked in our domain-specific vocabularies.’ The coverage percentage of function words extracted by each measure in Figure 1 decrease from left to right or from *Freq* to *McNemar*; e.g. 21.4 % for *Freq*; 17.6% for *Cosine*; 9.6% for *CSM*; 6.2% for *LLR*; and 6.0% for *Chi2/Yates*. No function words were extracted by *PMI* or *McNemar*. This data gives us a general idea of the percentage of function words that each statistical measure might extract from a corpus.

3.4.2 Basic word distribution

The coverage percentages of basic words decrease from left to right or from *Freq* to *McNemar*: 55.2% for *Freq*/*Dice*; 41.4 % for *Cosine*; 40.0% for *CSM*; 29.6 % for *LLR*; 27.2 % for *Chi2*/*Yates*; 9.8 % for *PMI*; and 1.8% for *McNemar*. Approximately three fourths of the *Freq* top 500 extractions belonged to either function words or basic words. About half of the *Cosine* and *CSM* top 500 extractions belonged to function words or basic words. The percentage of basic words extracted by *PMI* and *McNemar* was less than 10 percent. Seeing the high degree of basic and function words extracted by some measures, and few or none extracted by others, sheds light on the relationship of the extracted words in both grade level and frequency ranges. Of course it makes sense that statistics that extract a higher level of basic and function words would be more highly represented in lower grades and higher frequencies.

3.4.3 Specialist word distribution

Finally, to evaluate how effectively these measures extracted specialized words, the extracted commerce/finance list was compared to the existing specialized word control list. For this we used the 3,865 entry words in the *Longman Business English Dictionary*. Unlike the basic and function words, the nine statistical measures produced relevant specialized words, but not in the same regular pattern. The number of specialized words increases from left to right or from *Freq* to *PMI* but declines slightly for *McNemar*: 20.4% for *Freq*/*Dice*; 37.2 % for *Cosine*; 43.4% for *CSM*; 56.8% for *LLR*; 58.6% for *Chi2*/*Yates*; 72.4 % for *PMI*; and 53.6% for *McNemar*. The *PMI* obtained the highest overlap with the dictionary entries, followed by *Chi2*/*Yates*, *LLR*, and *McNemar*. For *PMI* in particular, there is about a 72 percent overlap with business words.

In addition to investigating the commerce/finance domain vocabulary, the same procedures were used to examine applied science vocabulary. We compared the top 500 BNC applied science extractions with a specialized list of science and technology (EST) dictionary entries (Walker, 1999). For *PMI* in particular, there is about a 49 percent overlap with EST words. Considering the gap in overlap between the two domains, the domain-specific vocabulary for commerce/finance is perhaps rather less divergent than for the applied science or EST domain. Moreover, the EST dictionary is written for students at the junior or senior college level. Also considering the conciseness and limited scale of the EST dictionary entries we used in comparison with the divergent distribution of words in the broad applied science domain, an overlap of 49 percent is reasonable. In a previous study, we compared the equivalent top 500 extractions from the *Corpus of Professional English* (CPE), a 20-million-word English corpus used by professionals in science and technology in twenty-two domains and which is probably more concentrated than the BNC applied science vocabulary, with this EST dictionary entry. We obtained an approximate 59 percent overlap for *PMI* (Chujo et al., 2007). The data showing specialist word distribution for the applied science domain is given in Appendix A.

3.4.4 ‘Other’ distribution

Among the top 500 words, some did not appear in any of the three categories of words and these are denoted as ‘other’ in Figure 1. The number of ‘other’ words increased from left to right or from *Freq* to *McNemar*: 3.0% for *Freq*/*Dice*; 3.8 % for *Cosine*; 7.0% for *CSM*; 7.4% for *LLR*; 8.2% for *Chi2*/*Yates*; 17.8% for *PMI*; and 44.6% for *McNemar*.

We closely examined the content of the ‘other’ lists for the commerce/finance domain and roughly categorized these words into three groups. While the original business dictionary entries contained ‘words and phrases,’ we only used ‘word’ entries to create a control list within the confines of this study. Thus a certain percentage of these words were parts of these ‘phrases’ such as ‘*equitable*’ in *equitable mortgage* and ‘*discretionary*’ in *discretionary costs*, which should have been included in the specialized words category. These comprise the first group. The second group contains derivational forms of the ‘word’ entries such as *deregulation* and *entrepreneurial*. The third are ‘intuitively less typical domain-specific words [which] often occur in many different domains’ (Asmussen, 2005, p.7). For example, the top 500 words for *Freq*/*Dice*, *Cosine*, *CSM*, *LLR*, and *Chi2*/*Yates* included some less typical entries such as *achieve*, *approach* and *generally*, and top 500 *PMI* and *McNemar* words included some candidates that introspectively do not seem typical for the commerce domain such as *day-to-day*, *peat*, *pest* and *postwar*. The percentages of the third type of words are 1.8% for *Freq*/*Dice*; 1.8% for *Cosine*; 4.0% for *CSM*; 3.2% for *LLR*; 3.0% for *Chi2*/*Yates*; 6.4% for *PM*I; and 22.8% for *McNemar*. 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.

4. PEDAGOGIC IMPLICATIONS

The goal of this study has been to examine the effectiveness of the statistical measures to target a particular proficiency level and to gage whether or not the lists generated in fact are specialist words when compared to a general corpus. Educators wishing to produce their own lists can find value in the application of various statistics to their own corpora even though the lists generated are not necessarily definitive. In using these measures, it is possible to narrow the number of candidates to create a wordlist at a targeted proficiency level.

Using the pedagogical findings of this study, we were able to develop a web-based vocabulary-learning program for Japanese college students for teaching intermediate-level business vocabulary, utilizing extracted *LLR* lists from the 7.12-million-word commerce/finance (written business) component and the 1.32-million-word spoken business component of the BNC. A description of the creation of this resource from systematically selecting the vocabulary to developing the program is provided in Chujo, Oghigian, Nishigaki, Utiyama, and Nakamura (2007). We found that the ease and speed of selecting targeted spoken and written business words saved both time and cost. We also created an English for Academic Purposes (EAP) web-based vocabulary-learning material from the 1.27-million-word educational spoken component of the BNC based on the selected beginner, intermediate, and advanced level spoken EAP words utilizing these nine measures. These web programs are available at http://www5d.biglobe.ne.jp/~chujo/.

5. CONCLUSION

Although it has been shown that specific statistic can extract certain types of words from a large corpus, little has been known about what types and levels of domain-specific terms the statistical measures are actually extracting. This study was conducted to answer this question. We were able to obtain a stable result showing that a specific measure tends to extract words within a similar grade level, frequency level, and type of vocabulary. These results support those of prior studies based on either a smaller 40,000-word Kyoto tourism corpus (Chujo, Utiyama & Oghigian, 2006) and 100,000-word TOEIC corpus (Chujo & Genung, 2005), or a larger 17-million-word corpus of professional English (CPE) for science and technology (Chujo, Utiyama & Nakamura, 2007). This 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 and domains are different.

From a practical pedagogical perspective, the potential for using corpora to create real-world vocabulary lists to target specific proficiency levels and specific domains is certainly worth exploring. In this study, we have found that it is possible to use particular statistical measures to extract certain types of words from a corpus. More specifically, we now know that *Freq*, *Dice*, *Cosine* and *CSM* tend to extract a greater number of basic and function words, and that not surprisingly, these words tend to be understood at lower grade levels and are therefore more appropriate for beginner level students. *LLR*, *Chi2* and *Yates* tend to extract fewer basic and function words and more intermediate level words, and *PMI* and *McNemar* tend to extract few or no function or basic words and tend to extract words more applicable to higher level students. These findings are supported by the appearance of these types of words in higher, mid-range and lower frequency bands, respectively.

When comparing the extracted domain-specific lists to control lists, i.e. comparing our extracted specialized vocabulary to similar-domain specialist dictionaries, we were able to obtain a maximum overlap of 72 percent for commerce/finance. That means generally, culling a corpus for a particular type of vocabulary might only net half of the possible items available. One of the main reasons for this is that the study was based on extracting only single-word units, e.g. *company*, or *management*, rather than multi-word units such as *equitable mortgage* and *discretionary cost*, and the control lists (dictionaries) included derivational forms that were not included in the lemmatized lists. Both of these factors can be addressed in a future study.

It could be argued that teaching multi-word units of meaning are more effective for students than single-unit vocabulary lists. Once we understand the nature of statistical extraction for single-word units, an important next step would be exploring the extraction of multi-word units. In the meantime, this work shows that statistical extraction of specific vocabulary from corpora can produce at least a workable starting point for educators. As studies continue, we hope the effectiveness of this approach will prove to be a useful tool.

REFERENCES

Asmussen, J. (2005). Automatic detection of new domain-specific words, using document classification and frequency profiling. *Proceedings from the Corpus Linguistics Conference Series*, *1* (1). Retrieved December 7, 2006 from http://www.corpus.bham.ac.uk/PCLC/

Baker, M. (1988). Sub-technical vocabulary and the ESP teacher: An analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language*, *4*(2), 91-105.

Bowker, L., & Pearson, J. (2002). *Working with specialized language*. London: Routledge.

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

Chujo, K., & Genung, M. (2005).Utilizing the British National Corpus to analyze TOEIC tests: The quantification of vocabulary-usage levels and the extraction of characteristically used words. *TOEIC Research Report, 3,* 1-20. Retrieved June 1, 2010 from http://www.toeic.or.jp/toeic_en/media/pdf/KiyomiChujo_E.pdf

Chujo, K., & Nishigaki, C. (2006). Creating spoken academic vocabulary lists from the British National Corpus. *Practical English Studies*, *12*, 19-34.

Chujo, K., & Utiyama, M. (2006). Selecting level-specific specialized vocabulary using statistical measures. *System, 34* (2), 255-269.

Chujo, K., Utiyama, M., & Oghigian K. (2006). Selecting level-specific Kyoto tourism vocabulary using statistical measures. In L. Yiu-nam, M. Jenkins, & H. Chung-shun (Eds.), *New aspects of English language teaching and learning* (pp. 126-138). Taipei, Taiwan: Crane Publishing Company Ltd. Retrieved June 1, 2010 from http://www5d.biglobe.ne.jp/~chujo/eng/list.html.

Chujo, K., Oghigian, K., Nishigaki, C., Utiyama, M., & Nakamura, T. (2007). Creating e-learning material with statistically-extracted spoken & written business vocabulary from the British National Corpus. *Journal of the College of Industrial Technology, Nihon University*, *40*, 1-12. Retrieved June 1, 2010 from http://www5d.biglobe.ne.jp/~chujo/data/BNCbiji.pdf.

Chujo, K., Utiyama, M., & Nakamura, T. (2007). Extracting level-specific science and technology vocabulary from the corpus of professional English (CPE). *Proceedings of the Corpus Linguistics 2007 Conference*, Birmingham, UK. Retrieved December 1, 2007 from http://www.corpus.bham.ac.uk/corplingproceedings07/

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

CLAWS7. (1996). [Computer software]. Lancaster: Lancaster University. Retrieved from February 1, 2010 http://ucrel.lancs.ac.uk/claws7tags.html

Daille, B. (1995). Combined approach for terminology extraction: Lexical statistics and linguistic filtering. *UCREL Technical Papers*, Department of Linguistics, University of Lancaster, *5*, 1-67.

Dale, E., & 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.

Evert, S., & Krenn, B. (2004). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language*, *19*(4), 450-466.

Faucett, L., Palmer, H. E., West, M., & Thorndike, E. L. (1936). *Interim report on vocabulary selection*. London: PS King.

Hardcastle, D. (2005). Using the distributional hypothesis to derive co-occurrence scores from the British National Corpus. *Proceedings from the Corpus Linguistics Conference Series*, *1* (1). Retrieved December 7, 2006 from http://www.corpus.bham.ac.uk/PCLC/

HarperCollins Publishers Ltd. (2004). Wordbanks Online. Retrieved May 7, 2010 from http://www.collins.co.uk/books.aspx?group=154

Harris, A. J., & Jacobson, M. D. (1972). *Basic elementary reading vocabularies*. New York: Macmillan.

Hiebert, E. H. (2005). In pursuit of an effective, efficient vocabulary curriculum for elementary students. In E. Hiebert, & M. Kamil (Eds.), *Teaching and learning vocabulary* (pp. 243-263), Mahwah, US: Lawrence Erlbaum Associates, Inc., Publishers.

Hisamitsu, T., & Niwa, Y. (2001). Topic-word selection based on combinatorial probability. *NLPRS-2001*, 289-296.

Hutchinson, T., & Waters, A. (1987). *English for specific purposes*. Cambridge, UK: Cambridge University Press.

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

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

Lemay, C., L’Homme, M.-C., & Drouin, P. (2005). Two methods for extracting “specific” single-word terms from specialized corpora: Experimentation and evaluation. *International Journal of Corpus Linguistics*, *10* (2), 227-255.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: The MIT Press.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, UK: 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.

Oakes, M. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Ooi, V., Tan, P., & Chiang, A. (2007). Mapping undergraduate culture using the personal blog: A corpus-based approach. *Book of Abstracts Corpus Linguistics 2007, University of Birmingham,* 49.

Pearson Education Limited. (2000). *Longman business English dictionary*. Harlow: Pearson Education Limited.

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

Rayson, P. (2005). Wmatrix: A web-based corpus processing environment. Computing Department, Lancaster University. Retrieved January 2, 2007 from http://ucrel.lancs.ac.uk/wmatrix/

Scott, M. (1996/1999/2004). Wordsmith tools. (http://www.lexically.net/wordsmith/).

Scott, M. (1997). PC analysis of key words and key key words. *System*, *25*(2), 233-245.

Sinclair, J., & Renouf, A. (1988). A lexical syllabus for language learning. In R. Carter, & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 140-160), London and New York: Longman.

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

Takefuta, Y. (1981). *Kompyuuta no mita gendai-Eigo (Computational analysis of contemporary English)*. Tokyo: Educa. (In Japanese).

Utiyama, M., Chujo, K., Yamamoto, E., & 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.

Utiyama, M., & Chujo, K. (2007). Linking word distribution to technical vocabulary. *Journal of the College of Industrial Technology, Nihon University*, *40*, 13-21.

Wakaki, M., & 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.

Walker, M. B. (1999). *Chambers dictionary of science and technology*. Edinburgh: Chambers Harrap Publishers Ltd.

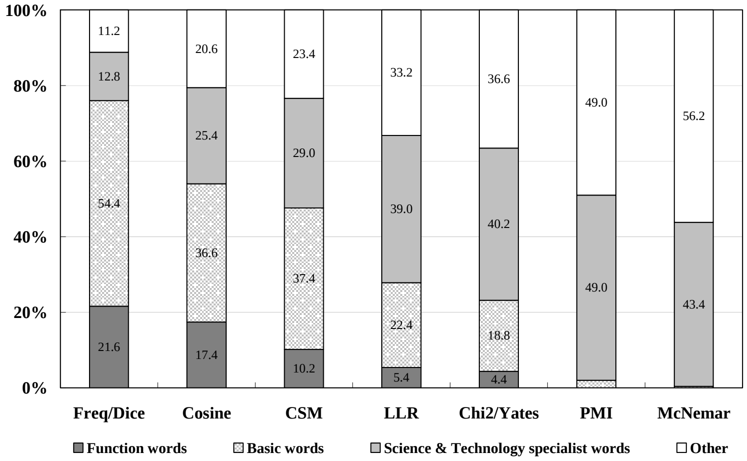
Wermter, J., & Hahn, U. (2006). You can't beat frequency (unless you use linguistic knowledge): A qualitative evaluation of association measures for collocation and term extraction. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, 785-792, Sydney, July 2006.

West, M. (1953). *A general service list of English words*. London: Longman.

Zanchetta, E., & Baroni, M. (2005). Morph-it! A free corpus-based morphological resource for the Italian language. *Proceedings from the Corpus Linguistics Conference Series, 1* (1). Retrieved December 7, 2006 from http://www.corpus.bham.ac.uk/PCLC/

APPENDIX A

The Proportion of the Three Categories of Words in the Top 500 Extractions of the Applied Science Domain



Note that the overlap with function words is shown with dark grey sections at the lowest part of the bars. The next section on the bars shows the overlap with basic words. The next section, shown in grey, shows the overlap in percentage between the *Concise Illustrated Dictionary of Science and Technology* entries and the top 500 words. The percentages of words not appearing in any of these three categories are denoted by 'other' and are shown in the white section at the top of each bar. As seen with the commerce/finance 500 extractions, there is a decrease in the number of function and basic words and an increase in specialized words as we move from left to right, but declines slightly for *McNemar*. The *PMI* obtained the highest overlap with the EST dictionary entries, followed by *McNemar, Chi2/Yates,* and *LLR*.

APPENDIX B

The Formulas for the Nine Statistical Measures

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 Commerce /Finance word list with the frequency of the same word in the BNC HFWL.

The program computes *a*, *b*, *c*, *d*, and *N*, and cross-tabulates these:

a stands for the frequency of word *X* in the Commerce/Finance word list
b stands for the frequency of word *X* in the BNC HFWL
c stands for the number of running words in Commerce/Finance not involving word *X*
d stands for the number of running words in BNC HFWL not involving word *X*
N denotes *a* + *b* + *c* + *d*

	Commerce/Finance	BNC HFWL
X	a	b
not X	c	d

$$LLR_0 = a\log\left(an/\left((a+b)(a+c)\right)\right) + b\log\left(bn/\left((a+b)(b+d)\right)\right)$$

$$+\, c\log\left(cn/\left((c+d)(a+c)\right)\right) + d\log\left(dn/\left((c+d)(b+d)\right)\right)$$

$$Chi2_0 = \left(n(ad-bc)^2\right)/\left((a+b)(c+d)(a+c)(b+d)\right)$$

$$Yates_0 = n\left(|ad-bc| - n/2\right)^2 / \left((a+b)(c+d)(a+c)(b+d)\right)$$

Correction of the above three measures：

$$LLR = sign(ad-bc) \times LLR_0$$

$$Chi2 = sign(ad-bc) \times Chi2_0$$

$$Yates = sign(ad-bc) \times Yates_0$$

$$sign(z) = \begin{cases} +1 & if\ z > 0 \\ -1 & otherwise \end{cases}$$

$$Dice = 2a/(2a+b+c)$$

$$Co\sin e = a/\sqrt{(a+b)(a+c)}$$

$$CSM = (ad-bc)/\sqrt{(a+c)(b+d)}$$

$$PMI = \log\left(an/\left((a+b)(a+c)\right)\right)$$

$$McNemar = \frac{(b-c)^2}{b+c}$$

$$Freq = a$$