Selecting level-specific specialized vocabulary using statistical measures

Kiyomi Chujo a, Masao Utiyama b, 1

a College of Industrial Technology, Nihon University, 2-11-1 Shin-ei, Narashino-shi, Chiba 275-8576, Japan
b National Institute of Information and Communications Technology, 3-5 Hikaridai, Seika-cho, Soraka-gun, Kyoto 619-0288, Japan

Received 11 April 2005; received in revised form 9 November 2005; accepted 19 December 2005

Abstract

To find an easy-to-use, automated tool to identify technical vocabulary applicable to learners at various levels, nine statistical measures were applied to the 7.3-million-word ‘commerce and finance’ component of the British National Corpus. The resulting word lists showed that each statistical measure extracted a different level of specialized vocabulary as measured by word length, vocabulary level, US native speaker grade level, and Japanese school textbook vocabulary coverage, and that these measures produced level-specific words; i.e., beginning-level basic business words were identified using Cosine and the complementary similarity measure; intermediate-level business words were extracted using log-likelihood, the chi-square test, and the chi-square test with Yates’s correction; and advanced-level business word lists were created using mutual information and McNemar’s test. We conclude that these statistical measures are effective tools for identifying multi-level specialized vocabulary for pedagogical purposes.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: Vocabulary; Vocabulary selection; Statistical measures; Specialized vocabulary; ESP; Corpus; Extraction; Multi-level

* Corresponding author. Tel.: +81 47 474 2825; fax: +81 47 473 1227.
E-mail addresses: chujo@ciith.niit.ac.jp (K. Chujo), mutiyama@nict.go.jp (M. Utiyama).
1 Tel.: +81 774 98 6835; fax: +81 774 98 6961.

0346-251X/ - see front matter © 2006 Elsevier Ltd. All rights reserved.
1. Introduction

Vocabulary expansion is essential for learners to gain proficiency in English (Nation, 1994) and empirical research has shown that having students use wordlists “play[s] an important role in speeding up lexical acquisition” (Beglar and Hunt, 2003, p. 9). To generate vocabulary lists for learners, earlier studies have used both objective measures such as frequency and/or range (Thordarson and Longe, 1944; Harris and Jacobson, 1972; Engels et al., 1981) and subjective selection principles such as ‘learnability’ (Mackey, 1965), ‘necessity’ (West, 1953), and ‘intuitions of teachers of English as a foreign language (EFL)’ (Hindmarsh, 1980). Despite being a widely used measure, frequency in particular has been criticized for its inability to extract low-frequency words, which often have high information content (Richards, 1970). Although some objective measures such as ‘coverage indices’ (Mackey and Savard, 1967) and ‘familiarity’ (Richards, 1974) have been proposed to compensate for this disadvantage, the issue is still unresolved. Regardless of methodology, researchers point out that it is important for teachers to be highly selective when choosing lexical items (Laufe et al., 2005).

Because English is increasingly becoming a lingua franca for international technology and commerce, the English for specific purposes (ESP) approach has been distinguished from general English in language teaching (Hutchinson and Waters, 1987; Dudley-Evans and St. John, 1998). One of the prominent characteristics of ESP is a heavy load of corresponding specialized vocabulary or “technical words that are recognizably specific to a particular topic, field, or discipline” (Nation, 2001, p. 198). To select specialized vocabulary, Sutarsyah et al. (1994, p. 48) found that using the criteria of frequency and range was 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. An automated means is clearly needed for creating technical vocabulary lists that are differentiated from high-frequency words and that make it possible “to generate word-lists which differentiate low frequency items from rare items” (McCarthy, 2002, p. 27).

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 wordlists/vocabulary in this article. Technical vocabulary means words specific to a field, and general vocabulary is a general base of English words.

2. Literature review

A number of corpus-based studies have used a range of statistical measures to identify collocations and technical terms. Kennedy (2003) used the mutual information (MI) measure to demonstrate the strength of the associations between adjectives and 24 selected amplifiers (or degree adverbs) and found the most frequently occurring amplifier collocations in the British National Corpus (BNC). For example, absolutely collocates most strongly with diabolical, and completely collocates most strongly with refitted. Scott (1997, pp. 236–245) defined a ‘key word’ as a ‘word which occurs with unusual frequency in a given text’ and proposed a method of identifying ‘key words’ in a text by using the chi-square and Yates’s correction (Yates) statistics, the latter of which is a version of chi-square for small samples. These statistics indicate whether a word is overused or underused in a specialized corpus compared with a corpus of general English.

Nelson (2000) used the LL 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 his one million-word business English corpus with its frequency in the BNC Sampler Corpus, which is a two million-word sub-corpus of the BNC. He was able to generate a list of business-related words such as business, market, customer, management, price, and bank. According to Oakes (1998, p. 174), LL “is a well-established statistical technique...and behaves well whatever the corpus size.” Tribble (2000, p. 81) showed that the ‘key-word’ function of WordSmith has the potential to provide important stylistic information. Hutton (2002, p. 68) stated that “many researchers find ‘keywords’ a useful starting point in investigating a special corpus.” Flowerdew (2005) also used the ‘key-word’ function of WordSmith to identify key lexicons that follow a problem-solution pattern. Chung and Nation (2004) used their own program to identify technical vocabulary by comparing the frequency of occurrence of words in an anatomy text with their frequency in a large general corpus and determined that it worked well but failed to identify words such as neck, chest, and skin, which were also common in the general corpus.

In a preliminary study (Chujo and Utiyama, 2004), we examined a range of statistical measures used in computational linguistics to identify technical vocabulary from a 100,000-word Test of English for International Communication (TOEIC) corpus, which is comprised of 16 practice tests. TOEIC is one of the most popular English certification tests in Japan. The measures examined were LL, MI, Ch2, Yates, the Dice coefficient (Dice), Cosine, complimentary similarity measures (CSM), and frequency. Dice and Cosine are statistics widely used to measure the similarity between collocations and between terms (Oakes, 1998, pp. 114, 184). CSM is a similarity measure often used in optical character recognition. Each measure was used to extract words that occur significantly more often in the TOEIC test corpus than in the BNC. Each resulting list was compared to an existing technical vocabulary control list (Chujo, 2003), and the corresponding statistical measures were evaluated for their effectiveness by calculating the proportion of relevant candidates they produced. We 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 average word length, BNC frequency, native speaker grade level, and Japanese textbook coverage.

3. Research questions

As noted earlier, it has been shown that specific statistics can be effectively used to identify specific types of words from a corpus. Our previous study focused on statistical application to a 100,000-word TOEIC corpus; this present study applies statistical measures to a 7.3 million word business and finance corpus to determine the effectiveness of chisquare (Ch2) statistic. He suggested that the procedure would provide guidance in identifying vocabulary items to be taught in EFL, and it is built into WordSmith Tools (Scott, 1996). With WordSmith Tools, users can choose between the log-likelihood (LL) and chi-square with Yates’s correction (Yates) statistics, the latter of which is a version of chi-square for small samples. These statistics indicate whether a word is overused or underused in a specialized corpus compared with a corpus of general English.
each of the nine statistics used in targeting appropriate vocabulary, and explores further issues relating to word length, vocabulary level, US native speaker grade level, and Japanese school textbook vocabulary coverage. Specifically, the following questions were addressed:

1. What are the differences and similarities in the specialized lists produced by each measure?
2. What types of business English words are extracted by each measure, and how are they ranked?
3. Do the measures extract words of different average lengths?
4. How frequently do the top (most frequently appearing) 500 words extracted by each method occur in the BNC?
5. At what US grade level are the top 500 words extracted by each method understood?
6. What percentage of the top 500 words extracted by each method appear in Japanese junior and senior high school texts?
7. What pedagogical applications are suggested by the extraction results?

4. Method

4.1. The data

4.1.1. Commerce and finance master word list

To extract business English specialized word sub-lists from a corpus, we needed to begin with one large master list of commerce and finance terms. To create this kind of business-related master list, we began with the 7.3 million word ‘commerce and finance’ written component of the BNC. This includes 284 texts from books in business and related fields such as accounting, advertising, banking, public relations, trading, and sales, and also business section articles from periodicals such as The Economist, The Guardian, and The Independent (see Burnard, 2000 for a list of excerpted works). The 7,257,533 words in this corpus were first lemmatized to extract all base forms using a tagging program (CLAWS7, 1996), which provides the possible base forms and parts of speech information for each word. (For example, finances, financing, and financed would be listed as finance.) This created a list of 154,669 different words. Secondly, if a word appeared fewer than 100 times in the corpus, it was deleted. Next, all proper nouns and numerals were identified by their part of speech tags and deleted manually because statistical measures mechanically identify these words as technical words (Scott, 1999) 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). Finally, this process yielded a 2597-word commerce and finance master list. It should be noted that the use of this type of statistical extraction will target only single-word lexical units (marketplace, stockmarket) and variants such as compounds (market place and stock market) may be overlooked.

4.1.2. Control lists

We wanted not only to extract business-related words but also to know if these words appear generally in English, at what frequency, and at what (US) native speaker grade level. In addition, we wanted to know if these extracted business terms are learned by Japanese students in the course of their junior and senior high school years, and if so, to what extent. For these reasons, three control vocabulary lists were used:

1. The British National Corpus High-Frequency Word List (BNC HFWL), a list of 13,994 lemmatized words representing 86,123,934 total words in the BNC that occur 100 times or more which was created using the same procedure as for the creation of the master list describing in Section 4.1.1 (compiling procedure is detailed in Chuo, 2004). It was used for comparison to statistically determine if and how these business-related words appear differently in a general corpus. The BNC is “one of the largest and most representative corpora of a single variety of English currently available” (Kennedy, 2003, p. 467), and the BNC HFWL is its core.

2. The LinG 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 (US) 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 authors created a junior and senior high school (JSH) textbook vocabulary list containing 3098 different base words. These were compiled from the top selling series of JSH textbooks (the New Horizon 1, 2, 3 series and the Unicon 1, 2 and Reading series) in Japan (Asano et al., 2000; Suenaga et al., 2000). Japanese high school students generally use these or similar books to study English before entering a university.

4.2. Statistical measures

The measures examined were mutual information (MI) (Church and Hanks, 1989), the log-likelihood ratio (LL) (Dunning, 1993), the chi-square test (Ch2) and chi-square test with Yates’s correction (Yates) (Hisamitsu and Niwa, 2001), the Dice coefficient (Dice) (Manning and Schütze, 1999), the 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). The first seven of these were used by Chuo and Utyayama (2004) and Utyayama et al. (2004), and all eight statistical measures are affinity or similarity measures that are widely used in computational linguistics. They automatically identify prominent words by making comparisons between one specified list (in this case, the commerce and finance master list) and another larger list (the BNC HFWL). In addition to these eight statistical measures, the simple frequency measure was included and used for comparison. The formula for each measure is given in Appendix.

To understand the word lists obtained, it is important to understand the concept of “outstandingness” (Scott, 1999). We want to determine what words each measure will identify and in order to be able to compare the word lists for each measure, we must determine not just those words which each measure will identify but those words that “stand out”, or are the most prevalent. The statistical score for the extent of each word’s “outstandingness” in frequency of occurrence is computed as follows: (1) four variables a, b, c, d (‘the frequency of word X in the Commerce word list’, ‘the frequency of word X in the BNC HFWL’, ‘the number of running words in Commerce not involving word X’ and ‘the number of running words in BNC HFWL not involving word X’), are computed for each word. (2) The variables are applied to each formula to yield each word’s “outstandingness” (Scott, 1999).
score. Since each measure uses a different formula, it gives a different score to each word. A detailed description of each measure can be found in Uitijama et al. (2004) and the notation for these kinds of statistics can be found in Scott (1997). Finally, (3) the words are sorted from the most outstanding to the least outstanding by the statistical ranking. 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 items, not to totally extract these items. Because there may be some variation in what any particular teacher or material writer will select, simply deleting the poor candidates from a more encompassing automated list would be a much simpler task than creating the entire list manually.

4.3. Understanding the meaning of the extracted specialized lists

All the extracted word lists were examined for:

(1) Agreement with the other statistical measures
(2) Top 50 specialized words overview comparison

In addition, the 500 most outstanding words of each list were studied to examine:

(3) Average word length
(4) Distribution of the BNC HFWL frequency bands
(5) Grade level based on word familiarity
(6) Number of words not covered by the JSH textbook vocabulary

5. Results and discussion

5.1. Agreement with the other statistical measures

To quantify the degree of similarity or difference of the lists generated by each measure, we compared their rank-ordered output for the same data and numerically expressed their agreement with each other by using Kendall’s rank correlation coefficient. This coefficient was used because the data used were ordinal and were ranked by statistical scores. The correlation is shown in Table 1, which provides a broad profiling of the rank-ordered output of the different statistical measures. A correlation of 1.0 or 0.9 is very strong. The correlations indicate that LL, Yates, and Chi2 produce results that are similar to each other, and CSM is also quite similar to the three measures; in addition, the similarity in output of Freq and Dice is very strong. On the other hand, MI, Cosine, and McNemar show low correlations with all other measures. McNemar has a particularly low correlation to other measures and has a marginally close correlation with MI.

5.2. Top 50 specialized word comparisons

The top 50 words from each of the nine different measures in descending order are shown in Table 2. Since the top 50 extractions made using Freq and Dice were virtually identical, they are shown in the same column, and because those of Chi2 and Yates were the same, they are listed in the same column. These similarities meet our expectations based on the above correlation observations. The bottom three rows of each column show the average frequency score, average word length, and percentage of function words (Nation, 2001) of the top 50 words generated by each statistical measure.

The specialized lists in Table 2 are very different from each other even though they were extracted from the same data. Words identified by Freq and Dice are general vocabulary words that usually appear at the top of high frequency lists in both small and large corpora. In fact, 82% of the top 50 words are function words. For Cosine and CSM, the top 50 extractions include some words that have particular technical uses in business such as market, price, cost, account, share, and firm. The LL(Chi2)/Yates lists seem to be well-suited to identifying ‘basic business words’ such as bank, asset, investor, shareholder, employee, credit, industry, capital, payment, stock, loan, exchange, and dividend. The MI and McNemar lists identify technical business words such as buyout, payout, arbitrage, subcontractor, shareholder, headhunter, issuer, drafter, liquidity, fiduciary, ledger, and volatility.

As we see from the data in the bottom three rows of Table 2, the average frequency score of each top 50 list, ranging from 61,909 to 134, and the rate of function words of each top 50 list, ranging from 82% to 0%, decreases from left to right or from Freq to McNemar. Function words are usually high-frequency words, also called structural words, that we cannot do without and are the kinds of words introduced very early in any type of language course. Thus, we can assume that high frequency words are familiar to learners and are therefore generally easier to learn. Next, correlating with the average frequency, the average word length of lists increases from left to right, ranging from 3.3 to 9.4. Takeda et al. (1994) showed that difficulty levels increase with increasing word length. Although we are aware that word difficulty seems to be influenced by many more factors than frequency and word length, this might support the possibility that specific statistical measures can be used to target specific grade-level vocabulary. This will be explored in the following sections.

5.3. Top 500 specialized word comparisons

The top 500 words extracted by each statistical measure were examined for their potential for pedagogical applications based on four criteria: average word length, BNC frequency, native speaker (US) grade level, and textbook coverage.

<table>
<thead>
<tr>
<th></th>
<th>LL</th>
<th>Yates</th>
<th>Chi2</th>
<th>CSM</th>
<th>MI</th>
<th>Cosine</th>
<th>Dice</th>
<th>Freq</th>
<th>McNemar</th>
</tr>
</thead>
<tbody>
<tr>
<td>LL</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.9</td>
<td>0.8</td>
<td>0.6</td>
<td>0.6</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Yates</td>
<td>1.0</td>
<td>0.9</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.5</td>
<td>0.5</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>Chi2</td>
<td>1.0</td>
<td>1.0</td>
<td>-</td>
<td>0.7</td>
<td>0.4</td>
<td>0.6</td>
<td>0.6</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>CSM</td>
<td>0.9</td>
<td>0.9</td>
<td>0.7</td>
<td>-</td>
<td>0.4</td>
<td>0.6</td>
<td>0.6</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>MI</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Cosine</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Dice</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.1</td>
<td>0.7</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Freq</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.1</td>
<td>0.7</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>McNemar</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.1</td>
<td>0.7</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>
### Table 2: Top 50 specialized words in commerce and finance as calculated using nine measures

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Freq.</th>
<th>CCASE</th>
<th>DSM</th>
<th>LL</th>
<th>Ch2</th>
<th>Yates</th>
<th>MI</th>
<th>McNemar</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The</td>
<td>The</td>
<td>Market</td>
<td>Market</td>
<td>Lending</td>
<td>Subcontractor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Be</td>
<td>Of</td>
<td>Company</td>
<td>Company</td>
<td>Buoyant</td>
<td>Acquirer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Of</td>
<td>Be</td>
<td>Berk</td>
<td>Berk</td>
<td>Long-run</td>
<td>Payout</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>To</td>
<td>To</td>
<td>The</td>
<td>Price</td>
<td>Arbitrage</td>
<td>Inter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>A</td>
<td>Business</td>
<td>Business</td>
<td>Subcontractor</td>
<td>Drafter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>And</td>
<td>And</td>
<td>In</td>
<td>Price</td>
<td>Investment</td>
<td>Stockmarket</td>
<td>No arbitrage</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>In</td>
<td>Will</td>
<td>Rate</td>
<td>Rate</td>
<td>Offer</td>
<td>Long-run</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>That</td>
<td>For</td>
<td>For</td>
<td>Cost</td>
<td>Firm</td>
<td>Drafter</td>
<td>Shareholding</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Have</td>
<td>That</td>
<td>Company</td>
<td>Firm</td>
<td>Cost</td>
<td>No-arbitrage</td>
<td>Headhunter</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>It</td>
<td>Have</td>
<td>Market</td>
<td>Tax</td>
<td>Shareholding</td>
<td>Tax-free</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>For</td>
<td>Will</td>
<td>By</td>
<td>Investment</td>
<td>Account</td>
<td>Headhunter</td>
<td>Buoyant</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>They</td>
<td>Company</td>
<td>On</td>
<td>Account</td>
<td>The</td>
<td>Payout</td>
<td>Cross-border</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>On</td>
<td>Market</td>
<td>Business</td>
<td>Share</td>
<td>Profit</td>
<td>Issuer</td>
<td>Headhunting</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Will</td>
<td>It</td>
<td>This</td>
<td>Profit</td>
<td>Contract</td>
<td>Liquidity</td>
<td>Arbitrage</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>This</td>
<td>By</td>
<td>May</td>
<td>Contract</td>
<td>Share</td>
<td>Salomon</td>
<td>Stockmarket</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>By</td>
<td>This</td>
<td>Bank</td>
<td>Of</td>
<td>Income</td>
<td>Settlement</td>
<td>Telegraph</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>With</td>
<td>On</td>
<td>Price</td>
<td>Income</td>
<td>Customer</td>
<td>Acquirer</td>
<td>Poultry</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>As</td>
<td>On</td>
<td>Cost</td>
<td>Financial</td>
<td>Asset</td>
<td>Volatility</td>
<td>Chargeable</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Not</td>
<td>They</td>
<td>Which</td>
<td>Customer</td>
<td>Financial</td>
<td>Accountancy</td>
<td>Ledger</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Or</td>
<td>Bank</td>
<td>Rate</td>
<td>Management</td>
<td>Inverter</td>
<td>Lender</td>
<td>Segmentation</td>
<td></td>
</tr>
</tbody>
</table>

### 5.3.1. Average word length

As the data in Table 2 suggests, word length increases as we move from left (Freq) to right (McNemar) for the top 50 words. To confirm this visual hypothesis holds true for the entire list, we computed the average word length of groups of 50 words from the lists, and the results up to the 500th word are illustrated in Fig. 1. Because the average word lengths for Freq and Dice were identical, and those of LL, Ch2, and Yates were similar to each other, only six lines are plotted in the graph.

The data shows that the measured extracted words of different lengths, with the longest words extracted by McNemar, then by MI, LL/Ch2/Yates, DSM, CCASE, and Freq/Dice, in that order. As Chujo and Takefuta (1989) and Takefuta et al. (1994) have shown, the average length of words can be a measure of difficulty level. This suggests that each measure identified different difficulty levels of words as its outstanding words. Interestingly, none of the measures use word length in their parameters, and yet the result implies a direct relationship to vocabulary level. Of course we are aware of the limitation of this type of statistical analysis, that, among the top 500 extractions there are 'long' words such as compounds like marketplace and headhunting and derivatives like reasonableness and segmentation and whose learning burden might be reduced if they contain base forms known by learners.

### 5.3.2. Vocabulary level defined by BNC frequency band distribution

The BNC represents present day general vocabulary usage. We examined the frequency distribution of the top 50 extracted words by using the BNC HFWL, which was divided into 14,000-word frequency bands of the most frequent words. BNC frequency bands '1000' indicates ranks 1–1000, 'frequency bands 2000' indicates ranks 1001–2000, etc. The percentages of the 500 words of each list that belong to each frequency band are shown in Table 3; a blank space indicates that no words belonged to that band. Because the scores for Freq and Dice were identical, and those of Ch2 and Yates were almost the same, only seven columns are shown.

![Fig. 1. Word length comparisons of 50-word groups of top 500 words.](image-url)
Table 3 shows clear graduations of frequency levels, and the top 500 extractions of each statistical measure are distributed as one might expect from the above observation. Looking across the table from Freq to McNemar, the top 500 words belong to increasingly lower frequency bands. The frequency bands to which more than 5% of extracted words belong can be used to clarify the frequency level comparisons. Most of the top 500 words from Freq and Dice belong to the 1000 and 2000 BNC frequency bands. In other words, most of the Freq and Dice words are included in the 1000 or 2000 most frequently appearing words in spoken and written English. More than 95% of the Cosine and CSM words belong to the top 1000–3000 BNC frequency bands. More than 65% of the LL, Chz2 and Yates words belong to the BNC most frequent 1000–4000 bands. Uniquely, MI extracts words evenly from all of the top 1000–8000 BNC frequency bands. And interestingly, McNemar extracts words from the BNC 4000–8000 frequency bands. The nine statistical measures clearly extract different outstanding levels of commerce and finance words.

5.3.4. Grade level rated by listing word vocabulary

To understand grade level definitions for these extracted words, we examined the levels of the top 500 extractions for word familiarity by native English speaking (NS) children (see Section 4.1.2). 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,” we determined at what grade level the majority of NS students would readily understand the central meaning of each word in the top 500 extractions produced by the nine statistical measures. (Note that grades 13 through 16 denote four years at the college or university level.) The results are shown in Table 4. The percentages of words not appearing in The Living Word Vocabulary are shown in the bottom row, denoted by ‘N/A’. Because the scores for Freq and Dice were identical, and those of Chz2 and Yates were almost the same, only seven columns are shown.

The grade at which 80% of extracted words are understood can be used to clarify the grade level comparisons and is indicated by underlined scores in Table 4: 80% of the top 500 words from Freq, Dice, Cosine, and CSM are understood by 6th grade students, those of LL are known by 8th grade students, those of Chz2 and Yates are known by 10th grade students, those of MI by 12th, and those of McNemar by 13th grade students. Again, this confirms that each statistical measure identifies different grade levels of words.

Fig. 2. Percentage of top 500 words covered by JS1 textbooks.

---

Table 3
Frequency distribution of top 500 extractions

<table>
<thead>
<tr>
<th>BNC frequency bands</th>
<th>Freq</th>
<th>Dice</th>
<th>Cosine</th>
<th>CSM</th>
<th>LL</th>
<th>Chz2, Yates</th>
<th>MI</th>
<th>McNemar</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>93.6</td>
<td>75.0</td>
<td>66.6</td>
<td>47.4</td>
<td>43.2</td>
<td>11.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>5.8</td>
<td>13.6</td>
<td>19.2</td>
<td>19.0</td>
<td>18.4</td>
<td>13.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3000</td>
<td>0.6</td>
<td>7.0</td>
<td>10.4</td>
<td>15.2</td>
<td>15.4</td>
<td>17.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4000</td>
<td>2.4</td>
<td>2.6</td>
<td>8.0</td>
<td>5.2</td>
<td>14.2</td>
<td>26.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>1.2</td>
<td>0.8</td>
<td>3.4</td>
<td>4.0</td>
<td>10.6</td>
<td>34.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6000</td>
<td>0.4</td>
<td>0.4</td>
<td>3.0</td>
<td>1.2</td>
<td>10.2</td>
<td>15.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7000</td>
<td>0.2</td>
<td>1.4</td>
<td>1.0</td>
<td>2.0</td>
<td>8.6</td>
<td>9.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8000</td>
<td>1.4</td>
<td>1.6</td>
<td>6.4</td>
<td>6.4</td>
<td>7.4</td>
<td>4.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9000</td>
<td>2.4</td>
<td>2.4</td>
<td>4.2</td>
<td>2.4</td>
<td>4.2</td>
<td>4.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10,000</td>
<td>0.2</td>
<td>0.2</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11,000</td>
<td>0.2</td>
<td>0.2</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12,000</td>
<td>0.2</td>
<td>0.2</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13,000</td>
<td>0.2</td>
<td>0.2</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13,994</td>
<td>0.2</td>
<td>0.2</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: underlined scores show the grade at which 80% of the extracted words are understood.

Table 4
US grade level based on word familiarity

<table>
<thead>
<tr>
<th>Grade</th>
<th>Freq</th>
<th>Dice</th>
<th>Cosine</th>
<th>CSM</th>
<th>LL</th>
<th>Chz2, Yates</th>
<th>MI</th>
<th>McNemar</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>74.4</td>
<td>62.6</td>
<td>54.8</td>
<td>44.0</td>
<td>41.4</td>
<td>23.8</td>
<td>15.6</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>12.0</td>
<td>22.4</td>
<td>25.9</td>
<td>26.8</td>
<td>26.0</td>
<td>22.4</td>
<td>17.6</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>4.0</td>
<td>7.6</td>
<td>10.4</td>
<td>12.4</td>
<td>12.4</td>
<td>15.4</td>
<td>18.0</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.6</td>
<td>3.0</td>
<td>3.8</td>
<td>5.8</td>
<td>6.4</td>
<td>8.0</td>
<td>11.8</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1.0</td>
<td>3.4</td>
<td>3.8</td>
<td>6.2</td>
<td>6.4</td>
<td>12.2</td>
<td>14.4</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.2</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>2.0</td>
<td>2.6</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.6</td>
<td>1.0</td>
<td>2.8</td>
<td>4.2</td>
<td>9.4</td>
<td>10.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N/A</td>
<td>0.2</td>
<td>1.6</td>
<td>2.8</td>
<td>6.8</td>
<td>9.4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: underlined scores show the grade at which 80% of the extracted words are understood.
JSH textbook coverage is one way to obtain an accurate estimate of the vocabulary level of each extraction, which is crucial information to EFL learners. For ESP learners who want to acquire commerce and finance vocabulary, the percent of words that were covered by the JSH textbook vocabulary, represented by the lower section of each bar in Fig. 2, may be important information. Fig. 2 graphically illustrates that while 86% of the Freq/Dice top 500 extractions are covered in the JSH school textbooks, 70% of the Cosine, 62% of the CSM, 46% of the LL, and 42% of the Chi2/Yates extractions are covered, and only 14% of the MI and 7% of McNemar extractions are covered in the JSH school textbooks. Since some of the word forms, which appeared both in the JSH textbooks and the top 500 extractions in business fields, might not be used with the same meaning, the percentage might be lower than those in Fig. 2, if their meanings are considered. Overall, we can say that the data in Fig. 2 show that the nine different statistical measures extract words of quite different grade levels.

6. Conclusion

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, 2004; Utiyama et al., 2004). Our study in combination with these previous studies shows that the results of the statistical measures on corpora are robust; i.e., the results are similar even though the examined corpora sizes are different. The obvious pedagogical implication is that these statistical tools can very effectively be used to automatically extract various types of specialized lists that can be quickly and accurately targeted to learners' vocabulary or proficiency levels. For example, we can infer that the basic business words extracted by Cosine and CSM would be good for beginning-level business English learners, the LL/Chi2/Yates lists would be suitable for intermediate-level business English learners, MI and McNemar would be appropriate for advanced-level business English learners, and Freq and Dice would be useful for business students who need to consolidate JSH vocabulary while learning basic business words.

The good news for teachers and material writers is that these statistical tools can help them to select technical vocabulary automatically without much specialist knowledge. Using extracted lists, teachers and material writers can easily manually delete less relevant candidates. One of the challenges in interpreting the results is that the meaning of each word was not considered in generating any of the lists. Also only single-word units were considered, and multi-word units and collocations were not considered in this study. Users of the statistical methodology described in this study would need to be aware of these limitations. It might also be useful to explore to what extent the business vocabulary identified using the approach employed in this study appears in references such as the Longman Business English Dictionary (2000). Our current direction is in selecting three-level business word lists based on the results of this study, and in determining the most practical way to use these nine resulting lists. For example, in order to create a beginning-level list, we are exploring whether it is more efficient to choose words targeted by using only one statistic or combining the results from two statistics such as Cosine and CSM.

Further work also needs to be done to expand this research to the other fourteen domain-specific components of the BNC, such as social science and arts, to define specialized vocabularies in each domain, and to apply other statistical measures to find more useful formulas for identifying specialized vocabularies. Finally, we would like to develop these specialized vocabularies into e-learning materials for vocabulary building.

Acknowledgements

This study was funded by a Grant-in-aid for Scientific Research (No. 17520401) from the Japan Society for the Promotion of Science and Ministry of Education, Science, Sports and Culture. It was also supported in part by Shogakukan Inc. We are grateful to the anonymous reviewers for detailed comments on an initial draft of this article.

Appendix

Statistics for determining whether a word appears more frequently in a specific corpus than in a general corpus are described here. The statistical measures used were Freq, Dice, Cosine, CSM, LL, Chi2, Yates, MI, and McNemar. These statistical scores can be calculated by using spreadsheet applications such as Excel. Their formulas are given below.

The statistic 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 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 word list
- $b$ stands for the frequency of word $X$ in the BNC HFWL
- $c$ stands for the number of running words in Commerce not involving word $X$
- $d$ stands for the number of running words in BNC HFWL not involving word $X$
- $n$ denotes $a + b + c + d$

\[
\begin{array}{|c|c|}
\hline
\text{Commerce} & \text{BNC HFWL} \\
\hline
X & a & b \\
not X & c & d \\
\hline
\end{array}
\]

\[
L_{L_1} = a \log(an/(a + b)(a + c)) + b \log(bn/(a + b)(b + d)) + c \log(cn/(c + d)(a + c)) + d \log(dn/(c + d)(b + d))
\]

\[
\text{Chi}_2 = (n(ad - bc)^2)/((a + b)(c + d)(a + c)(b + d))
\]

\[
\text{Yates}_4 = n((ad - bc) - n/2)^2/((a + b)(c + d)(a + c)(b + d))
\]

Correction of the above three measures:

$$LL = \text{sign}(ad - bc) \times LL_0$$

$$Ch2 = \text{sign}(ad - bc) \times Ch2_0$$

$$Yates = \text{sign}(ad - bc) \times \text{Yates}_0$$

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

$$\text{Diee} = \frac{2\alpha^2}{(2\alpha + b + c)}$$

$$\text{Cosine} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

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

$$\text{MD} = \log_{10}(\text{BIC}(a,b,c))$$

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

$$\text{Freq} = a$$

References


