論　文

日本大学生産工学部研究報告B
2007 年 6 月 第 40 巻

単語分布と専門語彙の関連付けに関する研究

内山将夫*・中條清美**

Linking Word Distribution to Technical Vocabulary

Masao UTIYAMA and Kiyomi CHUJO***

We describe a type of plot called *wordplot* that is designed to visualize the distribution of words in a specific domain in comparison with their distribution in general English. The wordplot is intended to give English teachers and developers of pedagogical materials a convenient way of selecting technical vocabulary. Words in certain regions of these wordplots occur frequently in specific domains and infrequently in general English. A detailed examination showed that the words in these regions were technical vocabulary, demonstrating the effectiveness of the wordplots in identifying technical vocabulary.

Keywords: Word Distribution, Technical Vocabulary, Scatter Plot, Visualization, Statistical Measures

1. Introduction

We describe a type of scatterplot that is designed to visualize the distribution of words in a specific domain in comparison with their distribution in general English. We use the British National Corpus (BNC) to represent general English and the Commerce and Natural Science domains, which are part of the BNC, to show the effectiveness of the plot in identifying technical vocabulary in these domains. The plot is intended to give English teachers and developers of pedagogical materials a convenient way of selecting technical vocabulary.

We call this particular type of scatterplot *wordplot*. A wordplot of a Commerce word list is shown in **Fig. 1**. The list consists of 2971 individual words with their frequencies in the Commerce domain. It was created from the Commerce component of the BNC as described in Section 3. In **Fig. 1**, this list is compared with the BNC High-Frequency Word List (BNC HFWL), a list of 13956 individual words with their frequency in the whole BNC (Chujo, 2004).[1)] Because each word in the Commerce word list is also included in the BNC HFWL, we can compare the frequency of each word in the Commerce word list with that of the same word in the BNC HFWL as described in **Table 1**. The sizes of these word lists are shown in **Table 2**.

In **Fig. 1**, the horizontal axes F_{global} and P_{global} represent the frequency and proportion of the frequency of each word in the BNC HFWL. The vertical axes F_{local} and P_{local} represent the frequency and

*情報通信研究機構主任研究員
**日本大学生産工学部教養・基礎科学系助教授

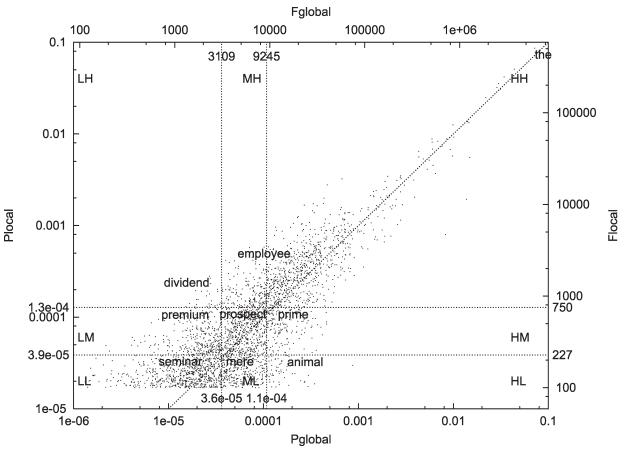


Fig. 1 Wordplot of Commerce word list

proportion of the frequency of each word in the Commerce word list and dots represent words in the Commerce word list.

Fig. 1 is divided into nine regions: HH, HM, HL, MH, MM, ML, LH, LM, and LL. X (and Y) of region XY means that the frequencies of words in that region are High (H), Medium (M), or Low (L) in the BNC HFWL (and Commerce word list). To set the boundaries for the X-regions (Y-regions), the words in the BNC HFWL (Commerce word list) were sorted in descending order of their frequency and divided into three sublists of the same length. The boundary between H and M was set at the highest frequency word in the medium-frequency word list and that between M and L was set at the highest frequency word in the low-frequency word list. The number of individual words in each region is shown in **Table 3**. The marginal frequencies are in the last column and last row.

A diagonal line, $P_{\text{local}}=P_{\text{global}}$, appears in **Fig. 1**. Words (dots) are distributed along this line, indicating that the frequency distribution of words in the whole BNC and in the Commerce domain are correlated. High-frequency words in the whole BNC tend to occur frequently in the Commerce domain. This tendency is also shown in **Table 3**. However, it can also be seen that some words occur more or less frequently in the Commerce domain than in the BNC.

A notable feature of **Fig. 1** is that words distributed in the upper region of the diagonal line ($P_{\text{local}} > P_{\text{global}}$) occur relatively more frequently in the Commerce domain. This observation is important because technical words tend to occur relatively frequently in a specific domain in comparison with in general English (Nation, 2001).[2] Thus, it is likely that technical vocabulary will be found in the upper region of this diagonal line.

Based on this observation, we conjectured that words in the LH and LM regions were likely to be technical words in the Commerce domain because words in these regions occur relatively frequently in that domain. To examine this idea, the highest F_{local} word in each region was displayed in **Fig. 1** to see which words were distributed in the region. The highest F_{local} 10 words in each region are listed in **Table 4**. In **Table 4**, an * denotes words that are not included in the JSH, a word list compiled from junior and senior high school English textbooks in Japan, as described in Section 2. The results generally supported our conjecture (the table is discussed in detail in Section 3). In the Appendix, we include **App. Fig. 1** and **App. Tables 1** and **2** to show that our observations

Table 1 Frequency pattern of word X

	SP	BNC
X	a	b
not X	c	d

This table indicates the pattern of the frequency of word X in the Commerce or Natural Science (SP) word lists in comparison with the frequency of the same word in the BNC HFWL: a and b stand for the frequencies of word X in the SP word list and the BNC HFWL, respectively, c stands for the number of running words in SP not involving word X, and d stands for the number of running words in BNC HFWL not involving word X. $a+c$ and $b+d$ are the numbers of total running words in these lists, whose actual values are listed in the column "No. of tokens" in **Table 2**. We use F_{local}, F_{global}, P_{local} and P_{global} to denote a, b, $\frac{a}{a+c}$, and $\frac{b}{b+d}$, respectively.

Table 2 Size of word lists

Word list	No. of types	No. of tokens
Commerce	2971	5877355
Natural Science	2323	805592
BNC HFWL	13956	86008037

"No. of types" is the number of individual words in each word list.

"No. of tokens" is the total number of running words in each word list, which is obtained by summing the frequencies of all words in each list. The procedure used to create these lists is described in Section 3.

Table 3 Number of individual words in each region for the Commerce word list

LH=29	MH=191	HH=771	991
LM=320	MM=506	HM=164	990
LL=641	ML=293	HL=56	990
990	990	991	2971

of the Commerce word list can also be applied to the Natural Science word list.

Below we describe the background to the development of wordplots and the procedure used to create the word lists we used in this work. We then examine the nature of the words in each region to evaluate the usefulness of wordplots as a means of identifying technical vocabulary.

Table 4 Highest ranked 10 words in each region for the Commerce word list

LH	MH	HH
dividend*	employee*	the
seller*	asset*	be
audit*	credit*	of
equity*	stock*	to
supplier*	cash*	a
discount*	loan	and
auditor*	consumer	in
accountant*	investor*	that
merger*	buyer*	have
monetary*	debt*	it

LM	MM	HM
premium*	prospect	prime
aggregate*	ordinary	image
lease*	external*	population
stake*	examination*	easily
organizational*	adviser*	context
innovation*	statutory*	pound
monopoly*	consultant*	near
incentive*	practical	version*
completion*	import	picture
publisher*	quation*	bear

LL	ML	HL
seminar*	mere	animal
auction*	tradition	surface
complexity*	skin	son
consensus*	pool	edge
respondent*	constitute*	husband
unchanged*	pure*	baby
redemption*	poverty	medium
goodwill*	healthy	father
reasonableness*	beginning	maybe
strictly*	distance	marry

2. Background

To become proficient in English, learners must expand their vocabulary (Nation, 2001).[3] Consequently, various word lists have been developed. These lists can be broadly categorized into three types, general, academic, and technical word lists, depending on their purposes.

General words Lists of general words are intended to provide basic English words. An example is *A General Service List of English Words* (West, 1953).[4] It includes the 2000 most basic words from the *Interim Report on Vocabulary Selection* (Faucett et al., 1936)[5] that are considered necessary for learning English as a foreign language. The criteria for vocabulary selection include *word frequency, ease or difficulty of learning, necessity, cover, stylistic level*, and *intensive and emotional words*. We call this list the GSL in this paper.

We also created a general word list consisting of 3098 words, which we call the JSH. It was compiled from the top-selling series of junior and senior high school textbooks in Japan (the *New Horizon 1, 2, 3* series and the *Unicorn I, II* and *Reading* series). Japanese high school students generally use these or similar books to study English before entering university.

Academic words Academic word lists highlight words that university students are likely to meet in a wide range of academic texts. The 'Academic Word List (AWL)' (Coxhead, 2000)[6] for written academic English was compiled from a 3.5 million-word+ corpus of academic texts covering subject areas such as the arts, commerce, law, and science. The list contains 570 words selected according to *range* and *frequency* criteria and incorporates words not included in the GSL. More than 94% of the words in the list occur in 20 or more of the 28 subject areas of the corpus.

Technical words Unlike general and academic word lists that cover a wide range of general and academic English, technical vocabulary is intended to cover technical words that students meet in specific fields. Thus, English teachers must provide technical words that will satisfy their students' need. This is a challenging task when using the traditional vocabulary selection criteria of *frequency* and *range* (Sutarsyah et al., 1994).[7] Because the focus of these two 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. English teachers generally do not have this specialist knowledge so there is clearly a need for easy-to-use tools for identifying technical vocabulary.

A number of corpus-based studies have used certain statistical measures to identify technical vocabulary. For example, Nelson (2000)[8] used the log-likelihood ratio (LLR) statistic (Dunning, 1993)[9] to find words that are statistically more frequently used in business English than in general English by comparing the frequency with which each word occurred in a business English corpus with its frequency in the BNC. He was able to generate a list of business-related words such as, *business, market, customer, management, price,* and *bank*.

Recently, Chujo and Utiyama (2005)[10] compared nine statistical measures for identifying technical vocabulary in the Applied Science domain in the BNC. These measures included the frequency of each word in the domain (Freq), LLR, and pointwise mutual information (PMI) (Church and Hanks, 1989).[11] They found that each statistical measure extracted a word list suitable for different-level learners of scientific English, i.e., LLR and PMI extracted intermediate and advanced level technical words, respectively, and Freq extracted general words not specific to that domain.

We also applied the nine statistical measures and several other measures to the Commerce and Natural Science word lists and confirmed these results. Below we list the top 25 words in the Commerce word list identified by Freq, Fisher's exact test (Fisher) and PMI.[12]

Freq : the, be, of, to, a, and, in, that, have, it, for, they, on, will, this, by, with, as, not, or, you, which, he, from, at

Fisher : market, company, bank, the, business, price, rate, cost, firm, tax, investment, account, share, profit, contract, of, income, financial, customer, management, product, asset, will, trade, be

PMI : lading, buyout, long-run, arbitrage, subcontractor, stockmarket, offeror, drafter, no-arbitrage, shareholding, headhunter, payout, issuer, liquidity, salesperson, settlor, acquirer, volatility, accountancy, lender, depreciation, tax-free, investor, dividend, relocation

These lists indeed contain different-level words, which is demonstrated by comparing them with other reference word lists such as the JSH.

However, we noticed a gap in the difficulty levels between the word lists extracted by Fisher and PMI. Because the words identified by PMI seem to be much more difficult than those identified by Fisher, we believe that some words are not identified by these measures.

This concern is illustrated in **Fig. 2**. We classified the 2971 words in **Fig. 1** into four classes (*Top, Middle, Bottom*, and *Others*) according to the ranking identified by Freq, Fisher and PMI. We first identified 1741 words that occurred statistically more significantly (at the 5% level) in the Commerce word list, based on Fisher's exact test. We then extracted top 1740 (=580×3) words in each measure and divided them into Top-, Middle-, and Bottom- 580 words. We classified the remaining 1231 words as Others. Those classified words are shown in **Figs. 2a**, **2b**, and **2c** for Freq, Fisher and PMI, respectively.

Fig. 2 shows that each measure extracted words from different regions of the wordplot. **Fig. 2d** shows the words that were included in the Top-580 words of some measures.[13] It clearly shows that some regions of the wordplot were not covered by these measures. The same phenomenon was also observed in the case of the Natural Science word list.

It might be possible to use a statistical measure to extract words from the regions that were not covered. However, we opted to examine the nature of the words in each region of the wordplots before devising any statistical measures.

Our examination showed that we could characterize each region in the wordplots in terms of general, academic, and technical vocabulary, as described in the next section. This finding suggested the idea of using wordplots as a tool for identifying technical vocabulary. In addition, wordplots could be combined with existing statistical measures to guide users to interesting top-ranked regions or words identified by these measures, as shown in **Fig. 2**. (We are now developing a program that allows users to examine wordplots interactively.)

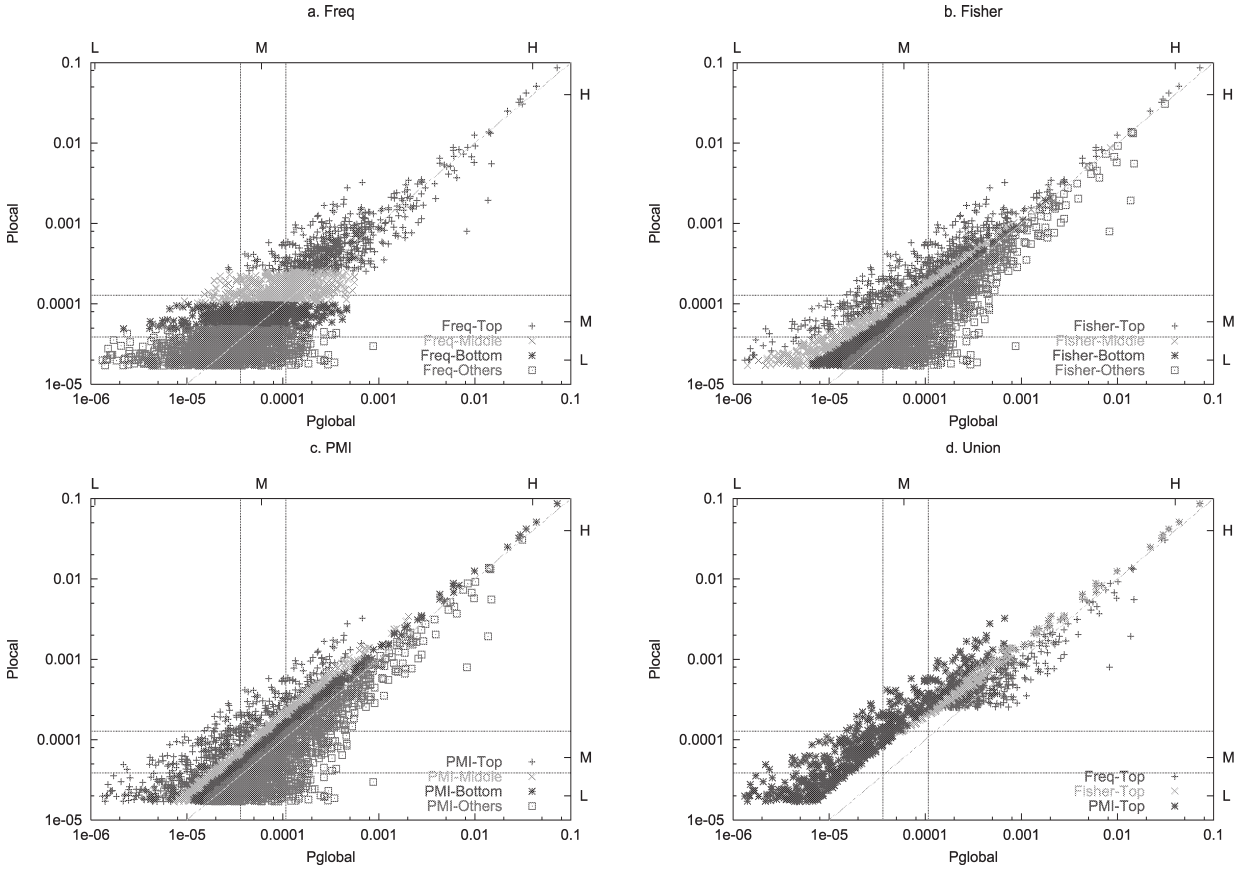


Fig. 2 Distributions of measures for Commerce word list

3. Observations

3.1 Data

We used the BNC HFWL to represent general English and the Commerce and Natural Science word lists to demonstrate how wordplots can identify technical vocabulary in these domains. These lists were created as follows.

We first lemmatized the Commerce component of the BNC by using the CLAWS7 program,[14] which provides possible base forms and part-of-speech information for each word. This created a list of 154669 individual lemmatized words. (For example, *finances, financing*, and *financed* would be listed as *finance*.) Then, words that occurred fewer than 100 times in the Commerce domain were deleted. Next, all proper nouns and numerals were identified by their part-of-speech tags and deleted manually because "they are of high frequency in particular texts but not in others ... and they could not sensibly pre-taught because their use in the text reveals their meaning" (Nation, 2001, pp.19-20)[15]. Finally, this process yielded a 2971 word list, which we call the Commerce word list.

The Natural Science word list and BNC HFWL were created using the same procedure as for the Commerce word list. The procedure used to compile the BNC HFWL is detailed in (Chujo, 2004).[16]

We used the GSL, JSH and AWL as reference lists to see which words were distributed in each region of the wordplots of the Commerce and Natural Science word lists. As described in the previous section, these lists contain general and academic words. We also used the function word list (FWL) from (Nation, 2001, pp. 430-431)[17] as a reference word list.[18] The numbers of individual words in these reference lists that are contained in the BNC HFWL, Commerce and Natural Science word lists are shown in **Table 5**.

Table 5 Number of individual words contained in word lists

Word list	FWL	GSL	JSH	AWL
Commerce	152	1110	1548	409
Natural Science	142	934	1261	309
BNC HFWL	164	1855	2889	560

3.2 Comparison with the reference word lists

In this section, we compare the nine word lists extracted from the nine regions (HH, HM ... LL) of

the Commerce wordplot in **Fig. 1** with the reference word lists. The statistics pertaining to the Natural Science word list are in the **Appendix**. These statistics show similar patterns to those in the Commerce word list.

We defined three indicators, percent inclusion at rank 50 (PI@50), percent inclusion (PI) and relative recall (RR) to compare each word list with the reference word lists.

Let TOP and ALL be the top 50 words and all words in the word list extracted from a region. These words are sorted in descending order of their frequency in the Commerce word list. Let REF be the all words in a reference list. Then, $\mathrm{PI@50}=\frac{|\mathrm{TOP}\cap\mathrm{REF}|}{|\mathrm{TOP}|}\times 100$, $\mathrm{PI}=\frac{|\mathrm{ALL}\cap\mathrm{REF}|}{|\mathrm{ALL}|}\times 100$. PI@50 and PI are used to show what percentages of the words in the top-50 and all words in each region are contained in the reference list. To define RR, we first define recall $R=\frac{|\mathrm{ALL}\cap\mathrm{REF}|}{|\mathrm{REF}|}$. Because the size of ALL varies according to each region (see **Table 3**), we divide R by the expected recall $E=\frac{|\mathrm{ALL}|}{2971}$ to calculate $\mathrm{RR}=\frac{R}{E}\times 100$. When RR>=100, then the region contains many more words in the reference list compared with the expected recall.

Table 6 Percent inclusion at rank 50 for Commerce word list

Region	FWL	GSL	JSH	AWL
HH	82.0	100.0	100.0	0.0
HM	8.0	70.0	90.0	14.0
HL	2.0	92.0	98.0	2.0
MH	0.0	14.0	32.0	38.0
MM	0.0	28.0	46.0	22.0
ML	4.0	46.0	56.0	18.0
LH	0.0	6.9	3.4	0.0
LM	0.0	2.0	8.0	8.0
LL	0.0	8.0	14.0	12.0

Table 7 Percent inclusion for Commerce word list

Region	FWL	GSL	JSH	AWL
HH	16.5	71.3	88.8	12.3
HM	4.3	73.8	89.6	12.2
HL	3.6	91.1	98.2	1.8
MH	0.5	23.0	37.7	26.7
MM	1.6	29.4	53.0	22.3
ML	2.4	47.4	60.8	12.6
LH	0.0	6.9	3.4	0.0
LM	0.0	3.8	14.7	11.6
LL	0.0	6.6	14.8	8.6

Table 8 Relative recall for Commerce word list

Region	FWL	GSL	JSH	AWL
HH	322.0	190.9	170.5	89.5
HM	83.4	197.5	172.0	88.6
HL	69.8	243.8	188.5	13.0
MH	10.2	61.7	72.3	194.0
MM	30.9	78.8	101.7	162.2
ML	46.7	127.0	116.6	91.7
LH	0.0	18.5	6.6	0.0
LM	0.0	10.0	28.2	84.0
LL	0.0	17.5	28.4	62.3

PI@50, PI, and RR for the Commerce word list are shown in **Tables 6, 7** and 8, respectively. From these tables, it is clear that words in the FWL occur mainly in the HH region. This is expected because function words occur frequently regardless of domains. Next, general words represented by the GSL and JSH occur mainly in HH, HM, and HL. That is, words in these regions are regarded as general words. Finally, words in the AWL often occur in MH and MM, which suggests that words in these regions are academic words. Words in ML seem to be a mixture of general and academic words.

The tables clearly show that words in LH, LM, and LL are not covered by the reference word lists. That is, words in these regions are neither general nor academic words. We will now examine the words contained in these and other regions by looking at the top-50 words in each region. (In future work, we will compare the words in these regions with references such as dictionaries of business English.)

3.3 Qualitative evaluation

The top-10 words from each of the nine regions are shown in **Table 4** for the Commerce word list and in **App. Table 1** for the Natural Science word list. Since **Table 4** and **App. Table 1** show similar tendencies in the extraction of each word lists, we will focus on the word lists in the Commerce domain. Note that the Commerce domain includes texts from related fields such as accounting, advertising, banking, public relations, trading, and sales.

When we look at the word lists in **Table 4** and **App. Table 1**, the vocabulary seems to fall into three categories: (1) general vocabulary, (2) academic vocabu-

lary, and (3) technical vocabulary. These categories are denoted by headings G (general), A (academic), and T (technical). We can further classify these three groupings into subgroups 1, 2, and 3, based on their frequency in the Commerce domain.

(1) General vocabulary (G1, G2, G3＝HH, HM, HL) In general, these three lists contain the most high-frequency, general-purpose vocabulary in English, and are typically found in the GSL as well as in the JSH.

The G1 list contains a high number of function words such as ***the***, ***a***, ***be***, ***of***, and ***to***. In the G2 list, there are some function words such as ***near*** and ***else***, and many content words, e. g., nouns such as ***image*** and ***population***, verbs such as ***replace***, adjectives such as ***prime*** and ***popular***, and adverbs such as ***easily***. The G3 list includes mainly broader content words such as ***animal***, ***surface***, ***son***, ***edge***, and ***husband***.

(2) Academic vocabulary (A1, A2, A3＝MH, MM, ML) Generally, these three lists are a mixture of the three types of vocabulary (general, academic and technical). However, they seem to be characterized by the presence of the words in the AWL (the AWL consists of words that are not included in the GSL but that are used in a wide range of disciplines including commerce).

The majority of the words in the A1 list belong to the business world and the homogeneity of this group is remarkable: "people" such as ***employee***, ***consumer***, ***investor***, ***buyer***, ***shareholder***, ***employer***, and ***chairman***; "business events" such as ***transaction*** and ***purchase***; "money/finance" such as ***asset***, ***credit***, ***stock***, ***cash***, ***loan***, ***debt*** and ***liability***. Many words (particularly after top 10) are words also included in the AWL, probably in the commerce discipline.

The A2 and A3 lists include a great deal of general-purpose and academic vocabulary. Words in these lists are not unique to a particular field and seem of little interest from the viewpoint of selecting technical vocabulary.

(3) Technical vocabulary (T1, T2, T3＝LH, LM, LL) Examination of this extracted vocabulary shows that the "T" category are found at all three levels (1, 2 and 3) and that the quantity of this type of vocabulary seems to decrease in order from 1 to 3. Almost all the words in the T1 list are purely business-related "specialist vocabulary" and they display a specific business focus. They include "business people" such as ***seller***, ***supplier***, ***auditor***, and ***accountant***; "business activities and events" such as ***takeover***, ***audit***, ***merger***, and ***valuation***; and "money/finance" such as ***dividend***, ***equity***, ***discount***, ***taxation*** and ***monetary***.

These words in the T1 list look more like a "lexis for talking about business" rather than a "lexis for doing business" (Pickett, 1998)[19] typically seen in the A1 list. In other words, these words might be more suitable for students of economics rather than of business.

We might expect to find the majority of the T2 words in a business environment. They are closely related to this discipline but may not be as strictly technical as T1. In the T3 list, some words might be used in the business world, e.g. s***eminar***, ***auction***, ***consensus***, ***redemption***, and ***decision-making***. Several are less directly related to business and a few have no direct connection with business.

4. Conclusion

Technical vocabulary is difficult to obtain using the traditional vocabulary selection criteria of *frequency* and *range*. 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 technical domain. English teachers generally do not have this specialist knowledge so there is clearly a need for easy-to-use tools for identifying technical vocabulary.

To provide such a tool, we introduced a particular type of plot called *wordplot* that can be used to visualize the distribution of words in a specific domain in comparison with their distribution in general English. We showed that words in the LH and LM regions in the wordplots of the Commerce and Natural Science domains, which occur frequently (high and medium) in these domains but infrequently (low) in the whole BNC, can be categorized as technical vocabulary. In future work, we want to apply wordplots to other domains to further examine their usefulness in identifying technical vocabulary.

References

1) Kiyomi Chujo. (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, editors, *English Corpora under Japanese Eyes*, pages 231-249. Rodopi.

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

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

4) M. West. (1953) *A General Service List of English Words*. Longman.

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

6) Averil Coxhead. (2000) A new academic word list. *TESOL Quarterly*, 34(2): 213-238.

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

8) M. Nelson. (2000) *A corpus-based study of business English and business English teaching materials*. PhD thesis, University of Manchester.

9) Ted Dunning. (1993) Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1): 61-74.

10) Kiyomi Chujo and Masao Utiyama. (2005) Selecting level-specific BNC applied science vocabulary using statistical measures. *Selected Papers from the 14th International Symposium on English Teaching*. English Teacher's Association, 195-202.

11) Kenneth W. Church and Patrick Hanks. (1989) Word association norms, mutual information, and lexicography. In *ACL-89*, pages 76-83.

12) These measures are defined as follows by using a, b, c, and d from **Table 1**: Freq$=a$, PMI$=\log \frac{an}{(a+b)(a+c)}$, and Fisher$=\sum_{x \geq a} \frac{{}_{a+c}C_{x\ b+d}C_{a+b-x}}{{}_{n}C_{a+b}}$, where $n=a+b+c+d$. The words extracted by Fisher and LLR are very similar. For the Commerce and Natural Science word lists, these two measures extracted the same word sets of up to 450 top-ranked words. We use Fisher's exact test in this paper because a p-value calculated by Fisher's exact test is not an approximation regardless of the sample size.

13) In terms of pedagogical materials, the top-ranked 580 words are more important than other words because class time and independent study time are too limited to learn all words.

14) http://www.comp.lancs.ac.uk/ucrel/claws/

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

16) Kiyomi Chujo. (2004) Measuring vocabulary levels of English textbooks and tests using a BNC lem matised high frequency word list. In J. Nakamura, N. Inoue, and T. Tabata, editors, *English Corpora under Japanese Eyes*, pages 231-249. Rodopi.

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

18) Function words express grammatical relationships with other words within a sentence. They may be prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles or particles.

19) Pickett. (1988) cited from Nelson (2000).

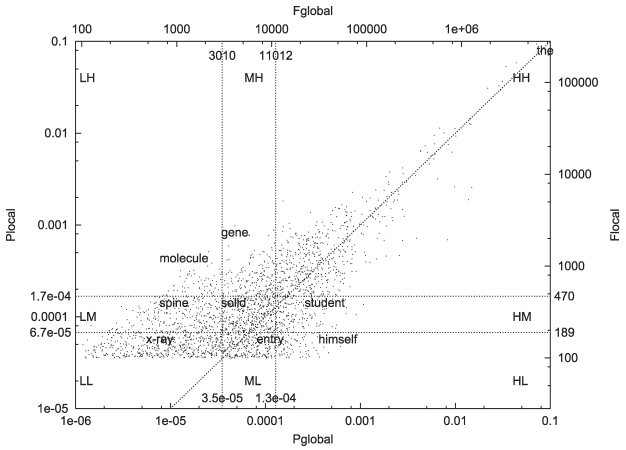
Appendix. Statistics for the Natural Science word list (NSWL)

App. Table 1 Top 10 words for the NSWL

LH	MH	HH
molecule*	gene*	the
particle*	surface	be
insect*	protein*	of
mouse	sequence*	and
electron*	temperature	a
polymer*	solution*	in
dolphin*	disease	to
clinical*	equation*	that
measurement*	complex	have
matrix*	female	it

LM	MM	HM
spine*	solid	student
node*	tail	association*
gravitational*	suitable*	performance
linear	calculate*	image
ion*	fully*	pass
mutant*	typical	purpose
segment*	largely*	technology
strand*	metal*	transfer
shield	concept*	experience
antibody*	prevent	education

LL	ML	HL
X-ray	entry*	himself
proof	proper*	trade
thickness*	yellow	labor
abundant*	electric*	commission*
temperate*	library	lot
vapor*	locate	everything
chronic*	waste	investment
liver*	gradually	aim*
cattle	summary*	anyone
accuracy*	totally*	charge



App. Fig. 1 Wordplot of the NSWL

App. Table 2 Number of individual words in each region for the NSWL

LH＝65	MH＝176	HH＝534	775
LM＝253	MM＝356	HM＝165	774
LL＝456	ML＝242	HL＝76	774
774	774	775	2323

App. Table 3 Percent inclusion at rank 50 for the NSWL

Region	FWL	GSL	JSH	AWL
HH	90.0	100.0	100.0	0.0
HM	4.0	68.0	98.0	12.0
HL	10.0	64.0	90.0	10.0
MH	0.0	48.0	60.0	20.0
MM	0.0	34.0	48.0	22.0
ML	0.0	48.0	64.0	22.0
LH	0.0	10.0	14.0	4.0
LM	0.0	14.0	16.0	8.0
LL	0.0	8.0	14.0	2.0

App. Table 4 Percent inclusion for the NSWL

Region	FWL	GSL	JSH	AWL
HH	22.3	78.5	93.6	11.0
HM	4.8	73.3	89.7	9.7
HL	6.6	69.7	88.2	13.2
MH	1.1	38.1	56.2	23.9
MM	1.4	36.2	58.1	21.1
ML	1.2	38.0	56.2	22.3
LH	0.0	10.8	18.5	4.6
LM	0.0	9.9	14.2	7.9
LL	0.0	4.6	12.3	6.6

App. Table 5 Relative recall for the NSWL

Region	FWL	GSL	JSH	AWL
HH	364.6	195.2	172.5	83.1
HM	79.3	182.4	165.2	72.9
HL	107.6	173.4	162.4	98.9
MH	18.6	94.7	103.6	179.4
MM	23.0	90.1	107.1	158.4
ML	20.3	94.6	103.5	167.8
LH	0.0	26.8	34.0	34.7
LM	0.0	24.6	26.2	59.4
LL	0.0	11.5	22.6	49.5

（H 18.12.20 受理）