CHAPTER 3

THE BLC WORDLISTS

Having overviewed the major characteristics of the BLC lexicon in terms of POS distribution in the previous chapter, we now turn to the BLC wordlists which contain more detailed information as to the behavior of individual lexical items. First, in Section 3.1, I will define some of the major technical terms used either in the wordlists or in my discussion thereof. Then, in Section 3.2, the two main wordlists compiled for the current study, *i.e.* the COMPREHENSIVE BLC WORDLIST and the BLC KEYWORDS LIST, will be presented and their contents briefly explained, including the necessary comments as to how to "read" these wordlists. Technical procedure of wordlist compilation, including some of the AWK programs used in the process, will also be described in some detail where appropriate. Section 3.3 introduces four categorical wordlists; that is, the wordlists compiled for each of the four major POS categories, *i.e.* verbs, adverbs, adjectives, and nouns. Note that detailed discussions as to what these wordlists tell us about the lexico-grammatical characteristics of written Business English will be made in Chapter 4.

3.1 Definitions of technical terms

So far, I have used some of the technical terms related to lexical analysis without giving any particular definitions, assuming that they are self-explanatory and are interpreted the way they are generally understood in the academic circle. However, since these technical terms and their definitions actually imply the theoretical bases for the arguments made in this paper, and since I will be introducing several terms that are peculiar to the current study, I believe it appropriate at this point to briefly summarize the definitions of major technical terms as they are used in this paper before proceeding further. The following definitions are largely based on Francis and Kuc&era (1982, pp. 3-4), but are nevertheless not necessarily the same as those given in their work in matters of details.

Type and **Token**: A *type* refers to the generic form of a particular word and a *token* is any instance of that form including its inflected variations. For instance, the word, *go*, when cited as a lexical item, is a type. If this form occurs 10 times in a text, each of these forms is its token. There are, therefore, 10 tokens of the type *go* in this text. If there are 10 occurrences of *go* and two each of *goes*, *went* and *gone* in

the same text, we have a total of 16 tokens of the type *go*. These terms are also referred to as *"word type"* and *"word token"* respectively, with or without a hyphen in between where appropriate.

Type-Token ratio (*or* T-T ratio): The ratio of the total number of word types to that of word tokens in a given text or a corpus. If, for instance, the total number of word tokens of a particular corpus is 1,092,589 and that of word types 19,441, then the T-T ratio of this corpus is about 1.78% (= (19,441÷1,092,589)×100). This can also be expressed as 1:56.3 (= 1: 1,092,589÷19,441), meaning that in this corpus a particular word-type is used about 56.3 times on average. The problem with the T-T ratio is that it varies very widely in accordance with the size of the corpus. Generally, the larger the corpus, the smaller the T-T ratio. In order for the ratio to be meaningful for any comparative purpose, therefore, it is customary to calculate a *standardized* T-T ratio obtained by computing the ratio every *n* numbers of words (every 1,000 or 10,000 words, for instance) and averaging the results.[1]

Lemma (*or* Lemma group): A set of graphic words having the same stem and/or meaning and belonging to the same grammatical word class, differing only in inflection and/or spelling. Thus, *go, goes, going, went* and *gone* form a lemma or a lemma group. In the COMPREHENSIVE BLC WORDLIST, which we will see in detail shortly, all the inflected forms of verbs and plural forms of nouns are lemmatized to their base forms, but adjectives and adverbs like *interesting* and *interestingly* are not lemmatized to the noun *interest*, because they belong to different word classes. Similarly, words like *kindness* and *kindly* are not lemmatized to the adjective *kind* for the same reason. The present and past forms of modals are considered to constitute separate word types. Also, in the wordlists prepared for the current study, pronouns and hyphenated words are not lemmatized to their nominative variants in case of the former and to their main constituents in the latter. Some scholars call the entry word (headword) of a wordlist or a dictionary a lemma. This usage, however, is not adopted in the current study.

Lemmatization: This refers to the grouping of graphic words into appropriate lemma groups. This process is necessary to obtain accurate frequency information. A lemmatized wordlist, if properly compiled, can also provide information as to the comparative frequencies of individual graphic words belonging to the same lemma group more readily than does an unlemmatized wordlist. In the current study, lemmatization was done using the automatic lemmatization function of WordSmith.[2]

Base form: The generic member of a lemma consisting of stem alone without inflection. In the above instance of the lemma group GO (*go, goes, going, went, gone*), the simple present form, *go,* coincides with the base form. The base form of the BE verbs is *be* and *is, are, was, were, been,* and *being* are its inflectional variants.

Graphic word: The actual instance of a word that occurs in a running text. Francis and Kuc&era (*ibid.*) defines it as "a string of contiguous alphanumeric characters with space on either side" and it "may include hyphens and apostrophes but not other punctuations." Thus, *above-mentioned* constitutes one graphic word, whereas *above mentioned,* two. Also, *The* and *the* are considered two different graphic words.

Entry (*or* Entry word): The heading (*or* headword) of a wordlist, usually identical with the base form. For instance, the entry word for a lemma group ASK (*ask, asks, asking, asked*) is *ask.* As for nouns, all the plural forms are represented by their singular variants, except in cases where only the plural form appears in the corpus. If a word has no inflectional variant, it automatically becomes the entry word of that token. The present and past forms of modals, as mentioned earlier, are listed separately and, therefore, each graphic word (*i.e.*, will, would, can, could, etc.) becomes its own entry word.

Word level (WL): This refers to the difficulty level of each word in a running text—not necessarily for the native speakers of English, but for the average adult Japanese EFL learners. In the current study, all the lexical items are grouped into ten different levels of difficulty (*i.e.* 01, 02, 03, 04, 05, 06, 11, 17, 21, and 30). Levels 01 through 04 consist primarily of the words listed in the *JACET 4000 Basic Words* (JACET, 1993)[3] and the "*D4000*" wordlist[4] compiled by Professor Kaneda of Nagoya Gakuin University on the basis of 15 major lexical studies and relevant reference documents including the JACET list. Both wordlists are similar in nature and each contains about 4,000 basic English words that Japanese students should be able to recognize by the sophomore year of college. Combining these two wordlists and adjusting several discrepancies found between the two and also with reference to other "educational" wordlists such as the *Hokkaido University English Vocabulary List* (Sonoda, 1996), we obtained a fairly reliable preliminary wordlist containing the most basic 4000+ words with information as to the "difficulty level" of each entry. Levels 06, 11, 17, 21, and 30 have been defined based on a "general" purpose wordlist compiled by Shiro Akasegawa (Akasegawa, 1995). This wordlist, referred to as "*30000.LST*" by the original compiler, contains a total of 30,652 words grouped in five levels, *i.e.*, Levels 06, 11, 17, 21, and 30. In

Level 06 are contained the most basic 6,000 words, whereas each of the other four levels contains words above and up to the adjacent levels of difficulty defined primarily on the basis of frequency. By consolidating these three wordlists, *JACET 4000, D4000* and *30000.LST*, we have obtained the first version of the *WL-tag dictionary* file, `wrdlvl-2.dic`[5].

This tag dictionary file is meant to be used in conjunction with an AWK program package developed by this author[6], and currently contains a total of 34,386 word tokens (28,925 types) as shown in Table 3-1 below:

WL Tag	Corresponding Word Level	No. of Word Tokens 1)	No. of Word Types	Total
01	1 - 1,000	1,664	1,066	4,459
02	1,001 - 2,000	1,500	1,001	
03	2,001 - 3,000	1,404	1,005	
04	3,001 - 4,000	1,817	1,387	
05	NNP & Abbrev.2)	3,292	3,122	3,122
06	4,001 - 6,000	1,609	1,229	21,344
11	6,001 - 11,000	4,764	3,774	
17	11,001 - 17,000	6,000	5,264	
21	17,001 - 21,000	4,112	3,501	
30	21,001 - 30,000	8,224	7,576	
	Total	34,386	28,925	28,925

1) Including inflectional variants. Of the total word tokens, 62.2% are nouns, 19.09% adjectives, 13.6% verbs, 4.25% adverbs, and the remaining 0.86% are distributed among other POS categories.

2) Note that all the proper nouns (NNPs), abbreviated NNPs and acronyms, including the names of people, companies, organizations, products, towns, cities, countries, languages, etc. are defined as *Level 5* words.

Table 3-1 Number of words contained in the *WL-tag dictionary* (wrdlvl-2.dic)

Word-level tag (WL tag): A string of two-digit numerals that indicates the difficulty level of each graphic token in a running text. (*e.g.* the_01, company_02, enclose_03, etc.). As mentioned above, there are currently 10 WL tags being used. The tag information is stored in the *WL-tag dictionary*.

Keyness score (*or* K-score): A numeric index of the "keyness" of each word in a corpus. The "keyness" of word *x* is calculated by comparing the frequency of that word in the target corpus with that of the same word in a much larger reference corpus, taking also into consideration the total numbers of running words (tokens) in both corpora. In the current study, Ted Dunning's Log Likelihood test was used to calculate the statistical significance of a given K-score. This test "gives a better estimate of *keyness* (than does the classic χ^2 test), especially when contrasting long texts or a whole genre against your reference corpus." (Scott, 1998. p. 65). A detailed theoretical description of Dunning's Log Likelihood test can be found in Dunning (1993). In the current study, a word is considered to be "key" either positively or negatively if the *p* value obtained for that word is larger or equal to 0.000001. Further details about the K-score will be given in the next section when we discuss the BLC KEYWORDS LIST.

3.2 The BLC General Wordlists

The "general" wordlists consist of two types: the COMPREHENSIVE BLC WORDLIST and the BLC KEYWORDS LIST. The former is a lemmatized wordlist and contains a total of 6,408 entries (for a total of 1,070,644 tokens) whose frequencies are larger than or equal to five, constituting about 97.99% of the adjusted total word tokens of the BLC. (The remaining 2.01% are those whose frequencies are less than five. More details will be given later). The latter wordlist was compiled in order to obtain the information regarding the "Keyness" of the main wordlist entries.

1) THE COMPREHENSIVE BLC WORDLIST

The COMPREHENSIVE BLC WORDLIST has been compiled roughly in the following procedure. First, the plain text version of the original BLC was run through WordSmith to create a basic wordlist with frequency information. The wordlist was then lemmatized using the automatic lemmatization function of WordSmith.[7] Note that the total number of word tokens reported by WordSmith was 1,119,578. This, however, includes many non-text strings and symbols which were contained in the original corpus. Excluding these "bugs" manually, we obtained an adjusted total tokens of 1,092,589. The percentage figures shown in the fourth column of the above wordlist were also re-calculated accordingly. For instance, the percentage of *the* to the total word tokens has been changed to 4.1116 (*i.e.* (44,923÷1,092,589)×100) from the unadjusted figure of 4.0125 initially reported in the above wordlist. The total number of word types after adjustment is 19,411.

From this wordlist, a new text file containing all the entry words in "one-entry-per-line" format was created for the purpose of assigning a corresponding word-level tag to each of the entry words.

```
the
to
be
you
of
and
a
your
in
[...]
```

The above list was then sorted in alphabetical order, after adding a serial number (*i.e.* frequency rank order) to each entry so that the original list order can be restored later, as follows:

```
a  7
abandon  5150
ability  579
able  168
aboard  2619
about  55
above  437
above-mentioned  3823
[...]
```

This file was then run through an AWK program, `matchnew.awk`, which replaces each entry word in the input file with a corresponding WL tag as defined in the WL-tag dictionary file, `wrdlvl-2.dic` (mentioned in the program as the default dictionary file).[8] This produces an output like the following, in which the data in the first field have been replaced with corresponding WL tags:

```
01  7
04  5150
02  579
01  168
01  2619
01  55
01  437
04  3823
[...]
```

The above data were then re-sorted according to the frequency rank order in Field 2 ($2) to restore the original order of the wordlist. This output was merged into the original wordlist, which was then imported to MS Excel for further processing. At the same time, a separate list of "Keywords" (KBLC KEYWORDS LIST, to be explained in the next section) was compiled and the data obtained therein were incorporated into the wordlist, thus producing the final product, *e.g.* the COMPREHENSIVE BLC WORDLIST (See Figure 3-3).

This wordlist contains a total of 6,408 entries (word types), or a total of 1,070,644 tokens, whose frequencies are larger than or equal to five, covering about 97.99% of the adjusted total word tokens of the BLC. The remaining 2.01% are those whose frequencies are less than five, a breakdown of which is given in Table 3-2 below:

Frequency	No. of Word Types	% to Total W. Types	No. of Word Tokens	% to Total W. Tokens
≧5	6,408	32.96	1,070,644	97.99
= 4	818	4.21	3,272	0.30
= 3	1,340	6.89	4,020	0.37
= 2	2,605	13.40	5,210	0.48
= 1	8,270	42.54	8,270	0.76
errors [1)]	--	--	1,173	0.11
(Sub Total)	(13,033)	(67.04)	(21,945)	(2.01)
TOTAL	19,441	100.00	1,092,589	100.00

1) Include non-text strings and/or OCR reading errors.

Table 3-2 Comparison of the numbers and percentages of word types and tokens at different frequencies in the COMPREHENSIVE BLC WORDLIST

Note that the entries whose frequencies are less than five are omitted from the wordlist because they are considered more or less idiosyncratic items peculiar to the texts chosen for compiling the corpus and are of little value for the purpose of the current study.[9] The other more practical reason for not including these items is that they would have added another 168 pages to the current list, which

obviously is not desirable in order to keep this paper as reasonably concise as possible.[10]

Figure 3-3 on the next page shows the first page of the COMPREHENSIVE BLC WORDLIST (the full list is given in Appendix D1, Vol. 2). In Column A of the wordlist is the frequency rank order, followed by the entry word of each rank in Column B. The third column, or Column C, shows the Word Level of each entry ("1" corresponds to the 1000-word level (01), "2"= 2000-word level (02), "3"= 3000-word level (03), and so on).

In Columns D through G are the absolute frequency of each entry (= Freq.), the cumulative frequency (= Cum.Freq.), the percentage of each entry to the total word tokens (=%), and the cumulative percentages (=Cum.%), respectively. The "Cum. %" column shows, for instance, that the first 10 word types (Ranks 1 through 10) constitute about 26.35% of the total word tokens, and the first 50 word types (Ranks 1 through 50) comprise about 49.11% of the total word tokens, and so on.

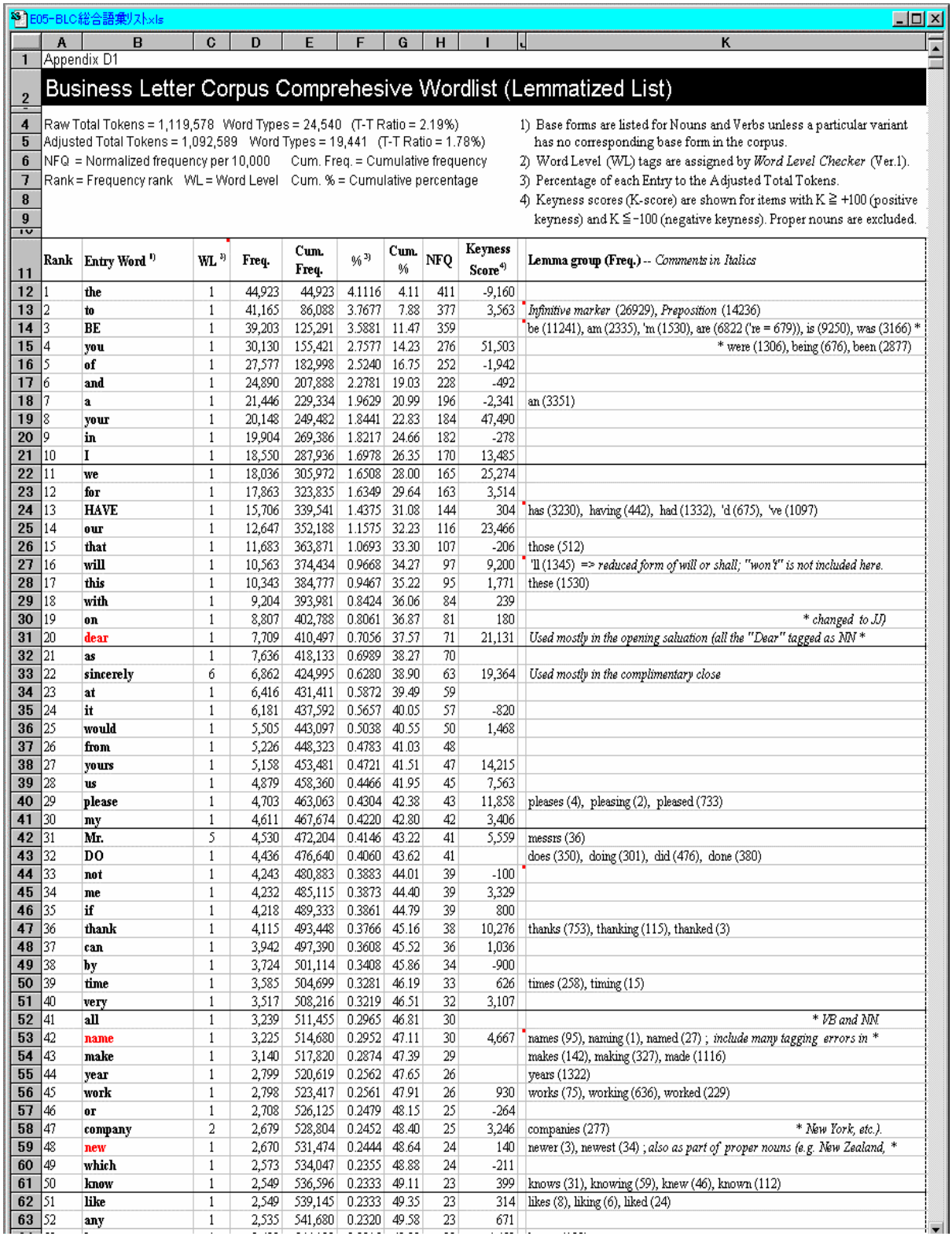
Column H (= NFQ) shows a "normalized" frequency of each entry. The NFQ of Item *n* has been obtained by the following formula:

$$\frac{nFRQ}{TWT} \times 10{,}000$$

where *nFRQ* is the absolute frequency of Item *n*, and *TWT* stands for the total number of word tokens. In other words, this means that the entry *the* (Rank 1), for instance, occurs about 411 times per every 10,000-word chunk, and *to* (Rank 2) occurs about 377 times in the same chunk of running text in the BLC.

The meaning of the "Keyness" score in the 9th column has already been explained briefly in the previous section. In this list, K-scores are given for those items whose K-scores are either larger or smaller than, or equal to, 100 on both the positive and negative sides (*i.e.* $+100 \leqq K \geqq -100$) for the sake of simplicity. A full information will be found in the BLC KEYWORDS LIST, which we discuss in the next section.

The final column shows the Lemma Group of each entry, as well as the frequencies of respective lexical items included in the lemma where applicable. My comments are added in *italics*. For instance, the comment added to *to* (Rank 2) reads "*Infinitive marker (26929), Preposition (14236),*" meaning that *to* is used 26,929 times as the infinitive marker *(i.e.* before a non-finite verb) and 14,236 times as a preposition (*i.e.* before a noun). Similarly, the comment added to *dear* (Rank 20) indicates that this item is used mostly in the opening salutation of the business messages contained in the BLC *(i.e.* "Dear Mr. Doe," or "Dear Sir," etc.), and that all the instances of *Dear_NN* have been manually changed to *Dear_JJ,* thus correcting the initial tagging errors.[11]



E05-BLC総合語彙リスト.xls

Appendix D1

Business Letter Corpus Comprehesive Wordlist (Lemmatized List)

Raw Total Tokens = 1,119,578　Word Types = 24,540　(T-T Ratio = 2.19%)
Adjusted Total Tokens = 1,092,589　Word Types = 19,441　(T-T Ratio = 1.78%)
NFQ = Normalized frequency per 10,000　　Cum. Freq. = Cumulative frequency
Rank = Frequency rank　WL = Word Level　Cum. % = Cumulative percentage

1) Base forms are listed for Nouns and Verbs unless a particular variant has no corresponding base form in the corpus.
2) Word Level (WL) tags are assigned by *Word Level Checker* (Ver.1).
3) Percentage of each Entry to the Adjusted Total Tokens.
4) Keyness scores (K-score) are shown for items with K ≧ +100 (positive keyness) and K ≦ -100 (negative keyness). Proper nouns are excluded.

Rank	Entry Word [1]	WL [2]	Freq.	Cum. Freq.	% [3]	Cum. %	NFQ	Keyness Score [4]	Lemma group (Freq.) -- *Comments in Italics*
1	**the**	1	44,923	44,923	4.1116	4.11	411	-9,160	
2	**to**	1	41,165	86,088	3.7677	7.88	377	3,563	*Infinitive marker* (26929), *Preposition* (14236)
3	**BE**	1	39,203	125,291	3.5881	11.47	359		be (11241), am (2335), 'm (1530), are (6822 ('re = 679)), is (9250), was (3166) * * were (1306), being (676), been (2877)
4	**you**	1	30,130	155,421	2.7577	14.23	276	51,503	
5	**of**	1	27,577	182,998	2.5240	16.75	252	-1,942	
6	**and**	1	24,890	207,888	2.2781	19.03	228	-492	
7	**a**	1	21,446	229,334	1.9629	20.99	196	-2,341	an (3351)
8	**your**	1	20,148	249,482	1.8441	22.83	184	47,490	
9	**in**	1	19,904	269,386	1.8217	24.66	182	-278	
10	**I**	1	18,550	287,936	1.6978	26.35	170	13,485	
11	**we**	1	18,036	305,972	1.6508	28.00	165	25,274	
12	**for**	1	17,863	323,835	1.6349	29.64	163	3,514	
13	**HAVE**	1	15,706	339,541	1.4375	31.08	144	304	has (3230), having (442), had (1332), 'd (675), 've (1097)
14	**our**	1	12,647	352,188	1.1575	32.23	116	23,466	
15	**that**	1	11,683	363,871	1.0693	33.30	107	-206	those (512)
16	**will**	1	10,563	374,434	0.9668	34.27	97	9,200	'll (1345) => *reduced form of will or shall; "won't" is not included here.*
17	**this**	1	10,343	384,777	0.9467	35.22	95	1,771	these (1530)
18	**with**	1	9,204	393,981	0.8424	36.06	84	239	
19	**on**	1	8,807	402,788	0.8061	36.87	81	180	* *changed to JJ)*
20	**dear**	1	7,709	410,497	0.7056	37.57	71	21,131	*Used mostly in the opening salutation (all the "Dear" tagged as NN* *
21	**as**	1	7,636	418,133	0.6989	38.27	70		
22	**sincerely**	6	6,862	424,995	0.6280	38.90	63	19,364	*Used mostly in the complimentary close*
23	**at**	1	6,416	431,411	0.5872	39.49	59		
24	**it**	1	6,181	437,592	0.5657	40.05	57	-820	
25	**would**	1	5,505	443,097	0.5038	40.55	50	1,468	
26	**from**	1	5,226	448,323	0.4783	41.03	48		
27	**yours**	1	5,158	453,481	0.4721	41.51	47	14,215	
28	**us**	1	4,879	458,360	0.4466	41.95	45	7,563	
29	**please**	1	4,703	463,063	0.4304	42.38	43	11,858	pleases (4), pleasing (2), pleased (733)
30	**my**	1	4,611	467,674	0.4220	42.80	42	3,406	
31	**Mr.**	5	4,530	472,204	0.4146	43.22	41	5,559	messrs (36)
32	**DO**	1	4,436	476,640	0.4060	43.62	41		does (350), doing (301), did (476), done (380)
33	**not**	1	4,243	480,883	0.3883	44.01	39	-100	
34	**me**	1	4,232	485,115	0.3873	44.40	39	3,329	
35	**if**	1	4,218	489,333	0.3861	44.79	39	800	
36	**thank**	1	4,115	493,448	0.3766	45.16	38	10,276	thanks (753), thanking (115), thanked (3)
37	**can**	1	3,942	497,390	0.3608	45.52	36	1,036	
38	**by**	1	3,724	501,114	0.3408	45.86	34	-900	
39	**time**	1	3,585	504,699	0.3281	46.19	33	626	times (258), timing (15)
40	**very**	1	3,517	508,216	0.3219	46.51	32	3,107	
41	**all**	1	3,239	511,455	0.2965	46.81	30		* *VB and NN.*
42	**name**	1	3,225	514,680	0.2952	47.11	30	4,667	names (95), naming (1), named (27) ; *include many tagging errors in* *
43	**make**	1	3,140	517,820	0.2874	47.39	29		makes (142), making (327), made (1116)
44	**year**	1	2,799	520,619	0.2562	47.65	26		years (1322)
45	**work**	1	2,798	523,417	0.2561	47.91	26	930	works (75), working (636), worked (229)
46	**or**	1	2,708	526,125	0.2479	48.15	25	-264	
47	**company**	2	2,679	528,804	0.2452	48.40	25	3,246	companies (277) * *New York, etc.).*
48	**new**	1	2,670	531,474	0.2444	48.64	24	140	newer (3), newest (34) ; *also as part of proper nouns (e.g. New Zealand,* *
49	**which**	1	2,573	534,047	0.2355	48.88	24	-211	
50	**know**	1	2,549	536,596	0.2333	49.11	23	399	knows (31), knowing (59), knew (46), known (112)
51	**like**	1	2,549	539,145	0.2333	49.35	23	314	likes (8), liking (6), liked (24)
52	**any**	1	2,535	541,680	0.2320	49.58	23	671	

Figure 3-3　Business Letter Corpus Comprehensive Wordlist (MS Excel screen shot)

(See Appendix D1 (in Vol. 2) for the complete wordlist)

The wordlist, if one looks at it carefully and with a purpose, tells us a great many things as to the lexical characteristics of Business English. On a macro level, for instance, the data provided in Columns G and C, *i.e.* the "Cum.%" and "WL" columns, are instrumental in probing our 1st and 2nd hypotheses as to the nature of business lexicon (see Chapter 1). On a micro level, the wordlist also reveals quite a few interesting facts about individual lexical items. Of particular interests are such items as personal pronouns, modals, *if*, infinitival *to*, the definite and indefinite articles, prepositions (*of*, in particular), and the relative pronoun *which*—the items whose distribution in the BLC we found statistically significant in the POS analysis discussed in Chapter 2.

In addition, such lexical items as *sincerely, please, thank, look, order, enclose, hope* and *appreciate*—to mention only a few from the first two pages of the wordlist—also deserve special attention. But before we begin our discussion on these and other related matters in more detail, let us take a brief look at the BLC KEYWORDS LIST, from which the information regarding the "Keyness" of the main entries has been obtained.

2) THE BLC KEYWORDS LIST

The BLC KEYWORDS LIST has been compiled using the Keywords program of WordSmith. This program allows one to identify key words in a given text by comparing the words in that text with a reference set of words taken from a much larger corpus of text. To do so, we first need to create two sets of wordlist using WordSmith: one from the BLC and the other from the combined Reference Corpus which is about three times larger than the former. Since the former wordlist had already been created (Figure 3-1), all we had to do was to produce a corresponding wordlist for the combined Reference Corpus. The two wordlists were then fed into the Keywords program and the rest was done automatically, yielding an output like the following:

(WordSmith Keywords List) ==========

N	Word	BLC Freq.	%	Ref Freq.	%	Keyness	P
1	you	30,130	2.76	9,541	0.30	51503.0	0.000000
2	your	20,148	1.84	2,260	0.07	47489.5	0.000000
3	we	18,036	1.65	7,997	0.25	25274.0	0.000000
4	our	12,647	1.16	3,296	0.10	23465.8	0.000000
5	dear	7,716	0.70	153	0.00	21131.1	0.000000
6	sincerely	6,862	0.63	14	0.00	19364.1	0.000000
7	yours	5,158	0.47	81	0.00	14214.8	0.000000
8	i	18,550	1.70	16,200	0.51	13484.5	0.000000

```
9    please      4,703    0.43   323      0.01   11857.7   0.000000
10   thank       4,115    0.38   306      0.01   10275.9   0.000000
[...]
============================================================
```

Figure 3-4 Sample output of WordSmith Keywords List,
comparing the BLC and the combined Reference Corpus (BROWN+LOB+TIME)

This output was then imported to MS Excel and, after making the necessary adjustments and recalculation, we finally obtained the BLC KEYWORDS LIST as shown in Figure 3-5 on the next page. (Figure 3-5 shows the first and last pages of the KEYWORDS LIST. The full list is given in Appendix D3, Vol. 2).

On the first page is shown the first 51 items whose K-scores are extremely high on the positive (+) side, meaning that these lexical items are used more frequently in the BLC than in the combined Reference Corpus. The second page, on the other hand, shows the last 61 items on the list whose K-scores are extremely high on the negative (-) side. Negative K-scores mean that these lexical items occur much less often than would be expected by chance in comparison with the combined Reference Corpus.

Note that a word whose frequency in the BLC is not either unusually high or low in comparison with what one would normally expect on the basis of the much larger combined Reference Corpus does not get into the list. The verb *take*, for example, is not on the list because its normalized frequency in the BLC (1827 x 2.8976 = 5293) is not significantly different from its frequency in the combined Reference Corpus (= 5188).[12] Also not on the list are proper nouns, including names of companies, products, projects, buildings, towns, cities, countries, days and months. Non-standard acronyms and abbreviations, as well as non-text strings of alphanumerics have also been excluded from the list.

As mentioned earlier, a word is considered to be "key" either positively or negatively if $p \leqq 0.000001$. The full KEYWORDS LIST (See Appendix D3, Vol. 2) indicates that the items that are "*positively* key" are those between Keyness Ranks 1 and 1100, while the items that are "*negatively* key" are those below Rank 4245 down to the last item of the list.[13]

The most interesting thing about the KEYWORDS LIST is that it allows us to look at the wordlist data mentioned in the previous section from quite a different perspec-tive, providing additional insights into the nature of Business English. The most obvious example is the definite and indefinite articles *the* and *a/an*. These two entries are ranked 1st and 7th in the COMPREHENSIVE BLC WORDLIST (Figure 3-3) respectively, whereas they are ranked at or very close to the bottom of the KEYWORDS

E07-BLCキーワードリスト.xls

Appendix D2

Business Letter Corpus (BLC) Keywords List

☞ Proper nouns, including names of companies, products, projects, buildings, towns, cities, countries; names of days and months, etc., non-standard acronyms and abbreviations., and non-text strings of alphanumerics are excluded from the list.

1) Keyness scores were obtained by comparing the entries of the BLC Comprehensive Wordlist with those of the combined Reference Corpus Wordlist. Note that a word whose frequency in the BLC is not either unusually high or low in comparison with what one would normally expect on the basis of the Reference Corpus does not get into the list. The verb type, *take*, for example, is not on the list because its normalized frequency in the BLC (1827 x 2.83 = 5170) is not significantly different from its frequency in the Reference Corpus (= 5188). A word is considered to be "key" (either positively or negatively) if $p \leqq 0.000001$.

2) N. Freq. = (Combined Ref. Corpus Word Tokens ÷ BLC Word Tokens) × BLC Freq.

Keyness Rank	Entry Word	Business Letter Corpus (BLC)				Combined Ref. Corpus (Brown+LOB+TIME)		Keyness[1] Score	*p*	
		BLC Freq.	%	N. Freq.[2]	%	Ref.C Freq.	%			
1	you	30,130	2.7577	87,306	2.7577	9,541	0.3014	51,503	0.000000	*
2	your	20,148	1.8441	58,382	1.8441	2,260	0.0714	47,490	0.000000	*
3	we	18,036	1.6508	52,262	1.6508	7,997	0.2526	25,274	0.000000	*
4	our	12,647	1.1575	36,646	1.1575	3,296	0.1041	23,466	0.000000	*
5	dear	7,716	0.7062	22,358	0.7062	153	0.0048	21,131	0.000000	*
6	sincerely	6,862	0.6280	19,884	0.6280	14	0.0004	19,364	0.000000	*
7	yours	5,158	0.4721	14,946	0.4721	81	0.0026	14,215	0.000000	*
8	I	18,550	1.6978	53,751	1.6978	16,200	0.5117	13,485	0.000000	*
9	please	4,703	0.4304	13,628	0.4304	323	0.0102	11,858	0.000000	*
10	thank	4,115	0.3766	11,924	0.3766	306	0.0097	10,276	0.000000	*
11	will	10,563	0.9668	30,608	0.9668	7,962	0.2515	9,200	0.000000	*
12	us	4,879	0.4466	14,138	0.4466	1,828	0.0577	7,563	0.000000	*
13	enclose	2,037	0.1864	5,902	0.1864	38	0.0012	5,579	0.000000	*
14	appreciate	1,848	0.1691	5,355	0.1691	119	0.0038	4,687	0.000000	*
15	letter	2,420	0.2215	7,012	0.2215	639	0.0202	4,450	0.000000	*
16	forward	1,863	0.1705	5,398	0.1705	417	0.0132	3,649	0.000000	*
17	to	41,165	3.7677	119,281	3.7677	81,724	2.5814	3,563	0.000000	*
18	for	17,863	1.6349	51,761	1.6349	29,040	0.9173	3,514	0.000000	*
19	my	4,611	0.4220	13,361	0.4220	3,960	0.1251	3,406	0.000000	*
20	me	4,232	0.3873	12,263	0.3873	3,462	0.1094	3,329	0.000000	*
21	request	1,439	0.1317	4,170	0.1317	179	0.0057	3,306	0.000000	*
22	company	2,679	0.2452	7,763	0.2452	1,433	0.0453	3,246	0.000000	*
23	product	1,748	0.1600	5,065	0.1600	493	0.0156	3,124	0.000000	*
24	very	3,517	0.3219	10,191	0.3219	2,611	0.0825	3,107	0.000000	*
25	customer	1,322	0.1210	3,831	0.1210	210	0.0066	2,872	0.000000	*
26	hope	2,028	0.1856	5,876	0.1856	1,095	0.0346	2,437	0.000000	*
27	receive	1,715	0.1570	4,969	0.1570	764	0.0241	2,380	0.000000	*
28	copy	1,068	0.0977	3,095	0.0977	180	0.0057	2,283	0.000000	*
29	sale	1,390	0.1272	4,028	0.1272	513	0.0162	2,169	0.000000	*
30	information	1,444	0.1322	4,184	0.1322	576	0.0182	2,152	0.000000	*
31	order	2,081	0.1905	6,030	0.1905	1,346	0.0425	2,126	0.000000	*
32	memo	764	0.0699	2,214	0.0699	26	0.0008	2,036	0.000000	*
33	send	1,575	0.1442	4,564	0.1442	791	0.0250	2,004	0.000000	*
34	wish	1,299	0.1189	3,764	0.1189	517	0.0163	2,004	0.000000	*
35	meeting	1,472	0.1347	4,265	0.1347	718	0.0227	1,915	0.000000	*
36	service	1,904	0.1743	5,517	0.1743	1,263	0.0399	1,896	0.000000	*
37	re	1,066	0.0976	3,089	0.0976	316	0.0100	1,861	0.000000	*
38	schedule	949	0.0869	2,750	0.0869	236	0.0075	1,785	0.000000	*
39	this	10,343	0.9467	29,970	0.9467	17,481	0.5522	1,771	0.000000	*
40	payment	973	0.0891	2,819	0.0891	275	0.0087	1,736	0.000000	*
41	account	1,286	0.1177	3,726	0.1177	610	0.0193	1,707	0.000000	*
42	invoice	603	0.0552	1,747	0.0552	0	0.0000	1,702	0.000000	*
43	business	1,979	0.1811	5,734	0.1811	1,508	0.0476	1,695	0.000000	*
44	regard	1,208	0.1106	3,500	0.1106	550	0.0174	1,651	0.000000	*
45	contact	856	0.0783	2,480	0.0783	260	0.0082	1,477	0.000000	*
46	would	5,505	0.5038	15,952	0.5038	8,071	0.2549	1,468	0.000000	*
47	price	1,342	0.1228	3,889	0.1228	809	0.0256	1,465	0.000000	*
48	gentleman	662	0.0606	1,918	0.0606	118	0.0037	1,392	0.000000	*
49	opportunity	968	0.0886	2,805	0.0886	417	0.0132	1,373	0.000000	*
50	credit	863	0.0790	2,501	0.0790	319	0.0101	1,344	0.000000	*
51	office	1,437	0.1315	4,164	0.1315	1,030	0.0325	1,318	0.000000	*

(1st Page: Keyness Ranks 1-51)

E07-BLCキーワードリスト.xls

	A	B	C	D	E	F	G	H	I	J	K
5447	K.Rank	Entry word	BLC Freq.	%	N. Freq.	%	Ref.C Freq.	%	Keyness	*p*	
5448	5352	state	540	0.0494	1,565	0.0494	3,272	0.1034	-282	0.000000	*
5449	5353	church	38	0.0035	110	0.0035	1,044	0.0330	-282	0.000000	*
5450	5354	power	112	0.0103	325	0.0103	1,438	0.0454	-286	0.000000	*
5451	5355	young	121	0.0111	351	0.0111	1,521	0.0480	-298	0.000000	*
5452	5356	face	127	0.0116	368	0.0116	1,568	0.0495	-304	0.000000	*
5453	5357	play	110	0.0101	319	0.0101	1,497	0.0473	-309	0.000000	*
5454	5358	hand	188	0.0172	545	0.0172	1,894	0.0598	-315	0.000000	*
5455	5359	against	195	0.0178	565	0.0178	1,935	0.0611	-317	0.000000	*
5456	5360	girl	14	0.0013	41	0.0013	996	0.0315	-319	0.000000	*
5457	5361	nation	37	0.0034	107	0.0034	1,144	0.0361	-322	0.000000	*
5458	5362	government	202	0.0185	585	0.0185	1,991	0.0629	-324	0.000000	*
5459	5363	million	109	0.0100	316	0.0100	1,542	0.0487	-325	0.000000	*
5460	5364	too	348	0.0319	1,008	0.0319	2,649	0.0837	-327	0.000000	*
5461	5365	old	214	0.0196	620	0.0196	2,109	0.0666	-344	0.000000	*
5462	5366	eye	38	0.0035	110	0.0035	1,244	0.0393	-355	0.000000	*
5463	5367	black	43	0.0039	125	0.0039	1,326	0.0419	-374	0.000000	*
5464	5368	even	659	0.0603	1,910	0.0603	4,134	0.1306	-380	0.000000	*
5465	5369	white	77	0.0070	223	0.0070	1,537	0.0485	-380	0.000000	*
5466	5370	seem	294	0.0269	852	0.0269	2,600	0.0821	-382	0.000000	*
5467	5371	never	186	0.0170	539	0.0170	2,107	0.0666	-385	0.000000	*
5468	5372	turn	183	0.0167	530	0.0167	2,094	0.0661	-385	0.000000	*
5469	5373	political	46	0.0042	133	0.0042	1,389	0.0439	-390	0.000000	*
5470	5374	up	1,131	0.1035	3,277	0.1035	6,063	0.1915	-404	0.000000	*
5471	5375	little	246	0.0225	713	0.0225	2,471	0.0780	-410	0.000000	*
5472	5376	page	191	0.0175	553	0.0175	2,210	0.0698	-410	0.000000	*
5473	5377	than	1,031	0.0944	2,987	0.0944	5,752	0.1817	-420	0.000000	*
5474	5378	child	116	0.0106	336	0.0106	1,888	0.0596	-429	0.000000	*
5475	5379	bush	4	0.0004	12	0.0004	1,308	0.0413	-452	0.000000	*
5476	5380	where	291	0.0266	843	0.0266	2,887	0.0912	-474	0.000000	*
5477	5381	down	230	0.0211	666	0.0211	2,614	0.0826	-479	0.000000	*
5478	5382	himself	26	0.0024	75	0.0024	1,545	0.0488	-486	0.000000	*
5479	5383	life	217	0.0199	629	0.0199	2,584	0.0816	-490	0.000000	*
5480	5384	and	24,890	2.2781	72,122	2.2781	82,449	2.6043	-492	0.000000	*
5481	5385	one	1,987	0.1819	5,758	0.1819	9,639	0.3045	-493	0.000000	*
5482	5386	house	183	0.0167	530	0.0167	2,481	0.0784	-511	0.000000	*
5483	5387	when	1,346	0.1232	3,900	0.1232	7,367	0.2327	-515	0.000000	*
5484	5388	there	1,473	0.1348	4,268	0.1348	7,870	0.2486	-520	0.000000	*
5485	5389	then	435	0.0398	1,260	0.0398	3,824	0.1208	-558	0.000000	*
5486	5390	out	1,051	0.0962	3,045	0.0962	6,421	0.2028	-563	0.000000	*
5487	5391	war	19	0.0017	55	0.0017	1,763	0.0557	-579	0.000000	*
5488	5392	world	236	0.0216	684	0.0216	2,968	0.0937	-584	0.000000	*
5489	5393	woman	62	0.0057	180	0.0057	2,058	0.0650	-590	0.000000	*
5490	5394	go	801	0.0733	2,321	0.0733	5,946	0.1878	-711	0.000000	*
5491	5395	into	661	0.0605	1,915	0.0605	5,350	0.1690	-713	0.000000	*
5492	5396	it	6,181	0.5657	17,910	0.5657	26,102	0.8245	-820	0.000000	*
5493	5397	by	3,724	0.3408	10,791	0.3408	17,941	0.5667	-900	0.000000	*
5494	5398	him	689	0.0631	1,996	0.0631	6,273	0.1981	-951	0.000000	*
5495	5399	its	584	0.0535	1,692	0.0535	6,090	0.1924	-1045	0.000000	*
5496	5400	their	1,221	0.1118	3,538	0.1118	9,073	0.2866	-1087	0.000000	*
5497	5401	who	1,032	0.0945	2,990	0.0945	8,466	0.2674	-1148	0.000000	*
5498	5402	man	202	0.0185	585	0.0185	5,001	0.1580	-1334	0.000000	*
5499	5403	they	1,334	0.1221	3,865	0.1221	11,400	0.3601	-1620	0.000000	*
5500	5404	of	27,577	2.5240	79,908	2.5240	104,495	3.3006	-1942	0.000000	*
5501	5405	but	1,937	0.1773	5,613	0.1773	15,224	0.4809	-1965	0.000000	*
5502	5406	she	509	0.0466	1,475	0.0466	8,772	0.2771	-2052	0.000000	*
5503	5407	her	596	0.0545	1,727	0.0545	9,259	0.2925	-2062	0.000000	*
5504	5408	say	686	0.0628	1,988	0.0628	10,436	0.3296	-2300	0.000000	*
5505	5409	a	21,446	1.9629	62,143	1.9629	87,110	2.7515	-2341	0.000000	*
5506	5410	his	1,066	0.0976	3,089	0.0976	19,766	0.6243	-4786	0.000000	*
5507	5411	he	1,484	0.1358	4,300	0.1358	25,928	0.8190	-6131	0.000000	*
5508	5412	the	44,923	4.1116	130,171	4.1116	205,132	6.4794	-9160	0.000000	*
5509											
5510	[end of list]										

Figure 3-5 BLC Keywords List screen shot (Last page: Keyness Ranks 5352-5412)

(See Appendix D3 for the complete Keywords List)

LIST with extremely high negative K-scores (-9160 for *the* and -2341 for *a/an*). This means that, although both the definite and indefinite articles are among the most frequent items in the BLC, they are nevertheless used much less frequently in written business discourse than in otherwise. What this and other similar finding entail we will discuss in detail in Chapter 4. For the moment, it suffices to confirm that simple frequency information alone may be quite insufficient and sometimes can be misleading in understanding the "true" picture of a given lexicon—which has been the main rationale for compiling the BLC KEYWORDS LIST.

3.3 The BLC Categorical Wordlists

In addition to the two main wordlists discussed above, I have also compiled several categorical wordlists for each of the four major word classes, *i.e.* verbs, adverbs, adjectives, and nouns. These wordlists are meant, first of all, to supplement the COMPREHENSIVE BLC WORDLIST which is useful in grasping an overall picture of the BLC lexicon but is nevertheless too inclusive to identify more fine-tuned categorical characteristics thereof. Suppose we want to find out the first most frequently used 50 verbs in the BLC lexicon, for instance, we have to check every entry in the COMPREHENSIVE BLC WORDLIST one by one for the first few pages until we get to the 50th verb. Similarly, this all-inclusive wordlist, however useful it may be for other purposes, is of little help in finding out, say, the number of adverbs needed to cover 90% of all the adverb occurrences in the BLC lexicon. If we have categorically compiled wordlists, such otherwise cumbersome work as mentioned above can be readily accomplished. The categorical wordlists are also indispensable for extracting what I would call the "core" vocabulary of Business English for each of the four major POS categories. We will get back to this latter topic in Chapter 4, but before proceeding further let us now briefly review the technical procedures through which these wordlists have been compiled.

1) THE BLC VERB LIST

The main BLC VERB LIST (LEMMATIZED USAGE RANK LIST) has been produced in roughly the following process (See Figure 3-7 for a flow chart).

First, the POS-tagged version of the BLC was run through a multi-function wordlist compiler, `mk_list.awk`, which was written by this author for the current study[14], to produce a simple wordlist containing all the words used in the BLC with corresponding POS tags. The following command line was executed from MS-DOS prompt, where "`blc.tag`" is the input file (*i.e.* the tagged version of the BLC), and "`blc_list.txt`" is the name given to the output file:

```
> jgawk -f mk_list.awk blc.tag > blc_list.txt
```

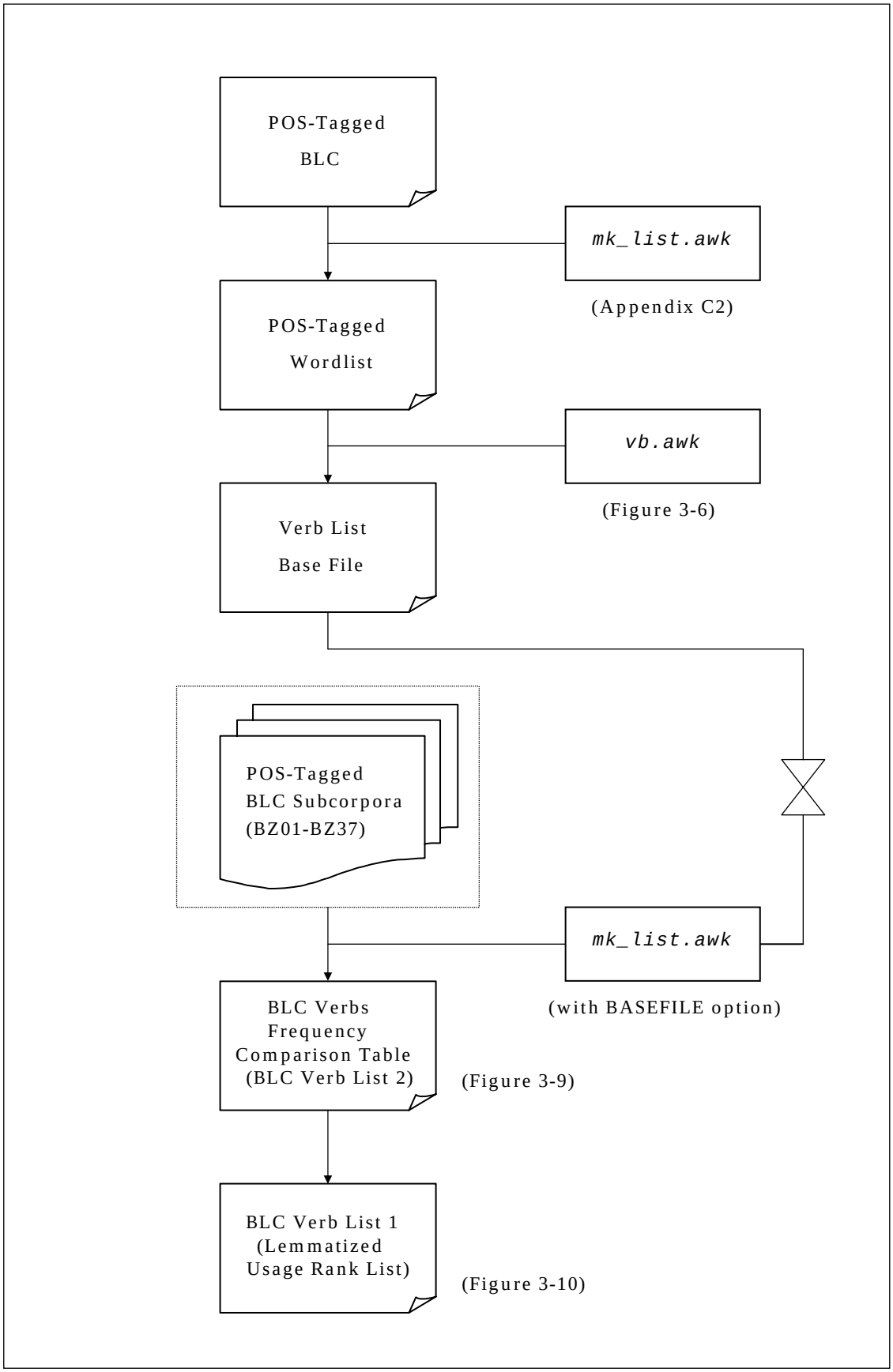


Figure 3-7 A flow chart showing steps to produce the Lemmatized BLC Verb List 1

The POS-tagged wordlist thus compiled was then run through an AWK program, `vb.awk`, to extract all the words with verb tags *(i.e.* VB, VBD, VBG, VBN, VBZ, and VBP).[15] This produced a preliminary list of verb-tag combinations containing all the verbs used in the BLC in their graphic forms. In order to change this list into a verb "base file" to be used in the next process, all the VB+ tags were deleted, leaving only the verb entries in Filed 1 ($1).

Next, POS-tagged versions of the 34 BLC subcorpora (from BZ01 to BZ37, excluding BZ7, BZ11 and BZ14) were run through the above-mentioned wordlist compiler, `mk_list.awk`, but this with the BASEFILE option (For details, see [Usage-2] of the source program, Appendix C2). This option allows one to create a consolidated word frequency comparison table from multiple input files for those words listed in an existing wordlist, *i.e.* the Base File. The following command line was executed from MS-DOS prompt, where "`base=`" is the statement to use the BASEFILE option and "`vb_base.txt`" is the name of the basefile to be used. The next command block, `c:¥temp¥*.* > blcvbtbl.txt`, declares that all the files stored in the "temp" directory (in this case, the 34 BLC subcorpora) are used as input files and that the results are to be written onto a new file, `blcvbtbl.txt`.

```
> jgawk -f mk_list.awk base=vb_base.txt c:¥temp¥*.* > blcvbtbl.txt
```

The following is an excerpt from the output of the above operation. In Field 1 ($1) are shown graphic verb entries in alphabetical order and Fields 2-35 ($1-$35) show frequencies of each entry in the 34 subcorpora respectively. The total number of occurrences of each entry is given in Field 36 ($36).

```
(Consolidated word freq. comparison table for BLC verbs)=============
WORDLIST=  BZ01 BZ02 BZ03 BZ04 BZ05 BZ06 BZ08 BZ09 BZ10 BZ12 BZ13 BZ15 BZ16 BZ17 BZ18
BZ19 BZ20 BZ21 BZ22 BZ23 BZ24 BZ25 BZ26 BZ27 BZ28 BZ29 BZ30 BZ31 BZ32 BZ33 BZ34 BZ35
BZ36 BZ37 TOTAL
[...]
abandon  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1
abandoned  0 0 0 0 0 0 0 2 0 2 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5
abated  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1
abide  0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2
abolish  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 2
abolished  0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
abound  0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
absorb  3 0 0 0 1 0 0 1 0 1 0 0 0 0 0 14 0 0 0 2 0 0 0 0 0 0 0 0 0 0 1 0 1 0 24
absorbed  0 0 0 2 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 5
```

```
absorbing 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 2
absorbs 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1
abuse 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
abused 0 0 0 0 0 0 0 2 0 0 0 2 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5
abusing 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2
[...]
======================================================================
```

Figure 3-8 A sample output of the consolidated word frequency comparison table for the BLC subcorpora, produced via *mk_list.awk* (excerpt)

This output was then imported to MS Excel and, after making the necessary adjustments and calculation, the GRAPHIC-WORD-BASED FREQUENCY COMPARISON TABLE (BLC VERB LIST 2) has been created. The first page of this table is given in Figure 3-9 on the next page for reference.[16] This table is not only instrumental in seeing how each verb entry is distributed among the 34 subcorpora, but also in calculating the "Usage" score of each entry. Before proceeding further, I will briefly define this notion of "Usage" since it is one of the primary statistical measures used in the current study in defining the relative importance of individual lexical items.

Definition of "Usage"

The notion of "Usage" (or *U* score) was first proposed in Juilland and Chang-Rodriguez (1964), and was subsequently adopted in Kuc&era and Francis (1967) in their study of the Brown Corpus with some modification. It is a product of the actual frequency of an item (in our case, a lexical item) in a corpus multiplied by the coefficient of dispersion of that item. The dispersion is a measure of how each item is evenly (or unevenly) distributed among subcorpora, and by multiplying the coefficient of dispersion of the lexical item in question with its actual frequency, we are able to get a better estimate of the relative importance of that item in the entire corpus. The *U* score of an item *x* (*Ux*) is calculated by the following formula:[17]

$$Ux = Fx \times (1 - SDx) \div Mx\sqrt{n-1}$$

where

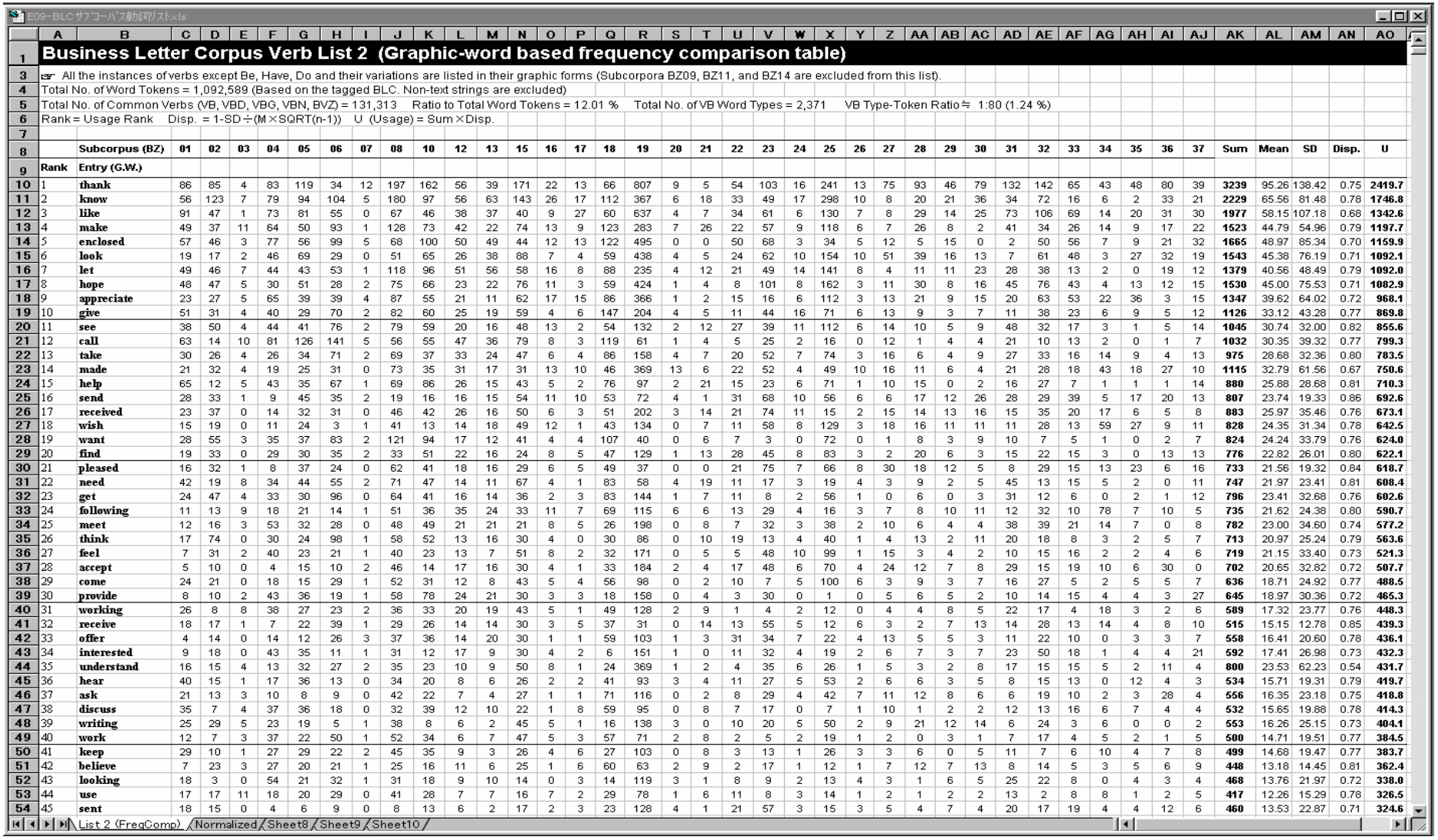
F*x* = Total frequency of Item *x*

SD*x* = Standard deviation of Item *x*

M*x* = Mean frequency of Item *x*

n = Total number of subcorpora

Figure 3-9　Business Letter Corpus Verb List 2: Graphic-word based Frequency Comparison Table (MS Excel Screen Shot)



E09-BLCサブコーパス動詞リスト.xls

Business Letter Corpus Verb List 2 (Graphic-word based frequency comparison table)

☞ All the instances of verbs except Be, Have, Do and their variations are listed in their graphic forms (Subcorpora BZ09, BZ11, and BZ14 are excluded from this list).

Total No. of Word Tokens = 1,092,589 (Based on the tagged BLC. Non-text strings are excluded)

Total No. of Common Verbs (VB, VBD, VBG, VBN, BVZ) = 131,313　Ratio to Total Word Tokens = 12.01 %　Total No. of VB Word Types = 2,371　VB Type-Token Ratio≒ 1:80 (1.24 %)

Rank = Usage Rank　Disp. = 1-SD÷(M×SQRT(n-1))　U (Usage) = Sum×Disp.

	Subcorpus (BZ)	01	02	03	04	05	06	07	08	10	12	13	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	Sum	Mean	SD	Disp.	U
Rank	Entry (G.W.)																																								
1	thank	86	85	4	83	119	34	12	197	162	56	39	171	22	13	66	807	9	5	54	103	16	241	13	75	93	46	79	132	142	65	43	48	80	39	3239	95.26	138.42	0.75	2419.7
2	know	56	123	7	79	94	104	5	180	97	56	63	143	26	17	112	367	6	18	33	49	17	298	10	8	20	21	36	34	72	16	6	2	33	21	2229	65.56	81.48	0.78	1746.8
3	like	91	47	1	73	81	55	0	67	46	38	37	40	9	27	60	637	4	7	34	61	6	130	7	8	29	14	25	73	106	69	14	20	31	30	1977	58.15	107.18	0.68	1342.6
4	make	49	37	11	64	50	93	1	128	73	42	22	74	13	9	123	283	7	26	22	57	9	118	6	7	26	8	2	41	34	26	14	9	17	22	1523	44.79	54.96	0.79	1197.7
5	enclosed	57	46	3	77	56	99	5	68	100	50	49	44	12	13	122	495	0	0	50	68	3	34	5	12	5	15	0	2	50	56	7	9	21	32	1665	48.97	85.34	0.70	1159.9
6	look	19	17	2	46	69	29	0	51	65	26	38	88	7	4	59	438	4	5	24	62	10	154	10	51	39	16	13	7	61	48	3	27	32	19	1543	45.38	76.19	0.71	1092.1
7	let	49	46	7	44	43	53	1	118	96	51	56	58	16	8	88	235	4	12	21	49	14	141	8	4	11	11	23	28	38	13	2	0	19	12	1379	40.56	48.49	0.79	1092.0
8	hope	48	47	5	30	51	28	2	75	66	23	22	76	11	3	59	424	1	4	8	101	8	162	3	11	30	8	16	45	76	43	4	13	12	15	1530	45.00	75.53	0.71	1082.9
9	appreciate	23	27	5	65	39	39	4	87	55	21	11	62	17	15	86	366	1	2	15	16	6	112	3	13	21	9	15	20	63	53	22	36	3	15	1347	39.62	64.02	0.72	968.1
10	give	51	31	4	40	29	70	2	82	60	25	19	59	4	6	147	204	4	5	11	44	16	71	6	13	9	3	7	11	38	23	6	9	5	12	1126	33.12	43.28	0.77	869.8
11	see	38	50	4	44	41	76	2	79	59	20	16	48	13	2	54	132	2	12	27	39	11	112	6	14	10	5	9	48	32	17	3	1	5	14	1045	30.74	32.00	0.82	855.6
12	call	63	14	10	81	126	141	5	56	55	47	36	79	8	3	119	61	1	4	5	25	2	16	0	12	1	4	4	21	10	13	2	0	1	7	1032	30.35	39.32	0.77	799.3
13	take	30	26	4	26	34	71	2	69	37	33	24	47	6	4	86	158	4	7	20	52	7	74	3	16	6	4	9	27	33	16	14	9	4	13	975	28.68	32.36	0.80	783.5
14	made	21	32	4	19	25	31	0	73	35	31	17	31	13	10	46	369	13	6	22	52	4	49	10	16	11	6	4	21	28	18	43	18	27	10	1115	32.79	61.56	0.67	750.6
15	help	65	12	5	43	35	67	1	69	86	26	15	43	5	2	76	97	2	21	15	23	6	71	1	10	15	0	2	16	27	7	1	1	1	14	880	25.88	28.68	0.81	710.3
16	send	28	33	1	9	45	35	2	19	16	16	15	54	11	10	53	72	4	1	31	68	10	56	6	6	17	12	26	28	29	39	5	17	20	13	807	23.74	19.33	0.86	692.6
17	received	23	37	0	14	32	31	0	46	42	26	16	50	6	3	51	202	3	14	21	74	11	15	2	15	14	13	16	15	35	20	17	6	5	8	883	25.97	35.46	0.76	673.1
18	wish	15	19	0	11	24	3	1	41	13	14	18	49	12	1	43	134	0	7	11	58	8	129	3	18	16	11	11	11	28	13	59	27	9	11	828	24.35	31.34	0.78	642.5
19	want	28	55	3	35	37	83	2	121	94	17	12	41	4	4	107	40	0	6	7	3	0	72	0	1	8	3	9	10	7	5	1	0	2	7	824	24.24	33.79	0.76	624.0
20	find	19	33	0	29	30	35	2	33	51	22	16	24	8	5	47	129	1	13	28	45	8	83	3	2	20	6	3	15	22	15	3	0	13	13	776	22.82	26.01	0.80	622.1
21	pleased	16	32	1	8	37	24	0	62	41	18	16	29	6	5	49	37	0	0	21	75	7	66	8	30	18	12	5	8	29	15	13	23	6	16	733	21.56	19.32	0.84	618.7
22	need	42	19	8	34	44	55	2	71	47	14	11	67	4	1	83	58	4	19	11	17	3	19	4	3	9	2	5	45	13	15	5	2	0	11	747	21.97	23.41	0.81	608.4
23	get	24	47	4	33	30	96	0	64	41	16	14	36	2	3	83	144	1	7	11	8	2	56	1	0	6	0	3	31	12	6	0	2	1	12	796	23.41	32.68	0.76	602.6
24	following	11	13	9	18	21	14	1	51	36	35	24	33	11	7	69	115	6	6	13	29	4	16	3	7	8	10	11	12	32	10	78	7	10	5	735	21.62	24.38	0.80	590.7
25	meet	12	16	3	53	32	28	0	48	49	21	21	21	8	5	26	198	0	8	7	32	3	38	2	10	6	4	4	38	39	21	14	7	0	8	782	23.00	34.60	0.74	577.2
26	think	17	74	0	30	24	98	1	58	52	13	16	30	4	0	30	86	0	10	19	13	4	40	1	4	13	2	11	20	18	8	3	2	5	7	713	20.97	25.24	0.79	563.6
27	feel	7	31	2	40	23	21	1	40	23	13	7	51	8	2	32	171	0	5	5	48	10	99	1	15	3	4	2	10	15	16	2	2	4	6	719	21.15	33.40	0.73	521.3
28	accept	5	10	0	4	15	10	2	46	14	17	16	30	4	1	33	184	2	4	17	48	6	70	4	24	12	7	8	29	15	19	10	6	30	0	702	20.65	32.82	0.72	507.7
29	come	24	21	0	18	15	29	1	52	31	12	8	43	5	4	56	98	0	2	10	7	5	100	6	3	9	3	7	16	27	5	2	5	5	7	636	18.71	24.92	0.77	488.5
30	provide	8	10	2	43	36	19	1	58	78	24	21	30	3	3	18	158	0	4	3	30	0	1	0	5	6	5	2	10	14	15	4	4	3	27	645	18.97	30.36	0.72	465.3
31	working	26	8	8	38	27	23	2	36	33	20	19	43	5	1	49	128	2	9	1	4	2	12	0	4	4	8	5	22	17	4	18	3	2	6	589	17.32	23.77	0.76	448.3
32	receive	18	17	1	7	22	39	1	29	26	14	14	30	3	5	37	31	0	14	13	55	5	12	6	3	2	7	13	14	28	13	14	4	8	10	515	15.15	12.78	0.85	439.3
33	offer	4	14	0	14	12	26	3	37	36	14	20	30	1	1	59	103	1	3	31	34	7	22	4	13	5	5	3	11	22	10	0	3	3	7	558	16.41	20.60	0.78	436.1
34	interested	9	18	0	43	35	11	1	31	12	17	9	30	4	2	6	151	1	0	11	32	4	19	2	6	7	3	7	23	50	18	1	4	4	21	592	17.41	26.98	0.73	432.3
35	understand	16	15	4	13	32	27	2	35	23	10	9	50	8	1	24	369	1	2	4	35	6	26	1	5	3	2	8	17	15	15	5	2	11	4	800	23.53	62.23	0.54	431.7
36	hear	40	15	1	17	36	13	0	34	20	8	6	26	2	2	41	93	3	4	11	27	5	53	2	6	6	3	5	8	15	13	0	12	4	3	534	15.71	19.31	0.79	419.7
37	ask	21	13	3	10	8	9	0	42	22	7	4	27	1	1	71	116	0	2	8	29	4	42	7	11	12	8	6	6	19	10	2	3	28	4	556	16.35	23.18	0.75	418.8
38	discuss	35	7	4	37	36	18	0	32	39	12	10	22	1	8	59	95	0	8	7	17	0	7	1	10	1	2	2	12	13	16	6	7	4	4	532	15.65	19.88	0.78	414.3
39	writing	25	29	5	23	19	5	1	38	8	6	2	45	5	1	16	138	3	0	10	20	5	50	2	9	21	12	14	6	24	3	6	0	0	2	553	16.26	25.15	0.73	404.1
40	work	12	7	3	37	22	50	1	52	34	6	7	47	5	3	57	71	2	8	2	5	2	19	1	2	0	3	1	7	17	4	5	2	1	5	500	14.71	19.51	0.77	384.5
41	keep	29	10	1	27	29	22	2	45	35	9	3	26	4	6	27	103	0	8	3	13	1	26	3	3	6	0	5	11	7	6	10	4	7	8	499	14.68	19.47	0.77	383.7
42	believe	7	23	3	27	20	21	1	25	16	11	6	25	1	6	60	63	2	9	2	17	1	12	1	7	12	7	13	8	14	5	3	5	6	9	448	13.18	14.45	0.81	362.4
43	looking	18	3	0	54	21	32	1	31	18	9	10	14	0	3	14	119	3	1	8	9	2	13	4	3	1	6	5	25	22	8	0	4	3	4	468	13.76	21.97	0.72	338.0
44	use	17	17	11	18	20	29	0	41	28	7	7	16	7	2	29	78	1	6	11	8	3	14	1	2	1	2	2	13	2	8	8	1	2	5	417	12.26	15.29	0.78	326.5
45	sent	18	15	0	4	6	9	0	8	13	6	2	17	2	3	23	128	4	1	21	57	3	15	3	5	4	7	4	20	17	19	4	4	12	6	460	13.53	22.87	0.71	324.6

List 2 (FreqComp) / Normalized / Sheet8 / Sheet9 / Sheet10

The *U* score takes any number larger than 0 (U = 0 when Freq.= 0) and the larger the figure the more important a particular lexical item in a given corpus of running texts in that it is not only used very frequently but is also distributed evenly among the subcorpora of that corpus.

In the GRAPHIC-WORD-BASED FREQUENCY COMPARISON TABLE (Figure 3-9), the entries are sorted in descending order of the *U* score that is given in the last column of the table for each of the graphic entries. By lemmatizing all the entries to their headwords, we finally get a revised, more easy-to-read and information-rich wordlist for the BLC verbs (*i.e.* the BLC VERB LIST 1: LEMMATIZED USAGE RANK LIST) as shown in Figure 3-10 on the next page, a full list of which can be found in Appendix E1 of Vol. 2 of this paper.

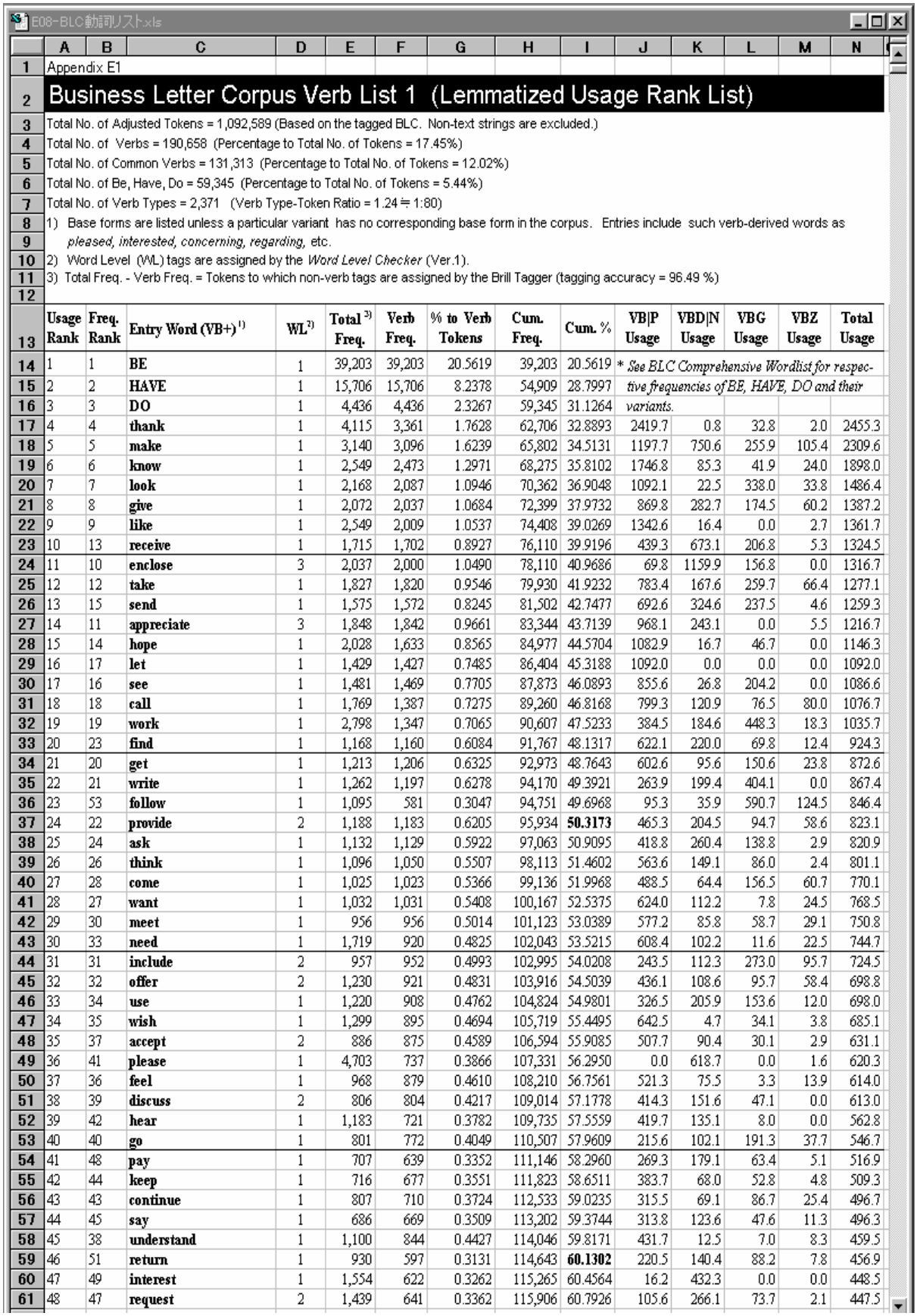
In Column A of this verb list is the usage rank order, followed by the frequency rank order and the entry words in their lemmatized forms in Columns B and C respectively. Note that the verbs that appear only in a particular inflectional variant get into the list in their inflected forms. Also, the entries include such verb-derived words as *pleased, interested, concerning, regarding* and so on for technical reasons.

The fourth column, or Column D, shows the Word Level of each entry. As already explained in the previous subsection, the figures given in this column correspond to respective word levels, *i.e.* "1" corresponds to the 1000-word level (01), "2" to the 2000-word level (02), "3" to the 3000-word level (03), and so on.

In Columns E and F are the total frequency of each entry (= Total Freq.) and the total frequency of "verb" instances of that entry (= Verb Freq.). The entry *thank*, for instance, occurred in the BLC a total of 4,115 times, of which 3,361 instances are the cases of verbs.[18]

Between Columns G and I are given the percentage of each entry to the total verb tokens (= % to Verb Tokens), the cumulative frequencies (Cum. Freq.) and the corresponding cumulative percentages to the total tokens (Cum.%) respectively. The "Cum. %" column shows, for instance, that the first 24 verb types constitute about 50.32% of the total verb tokens, and the first 46 verb types comprise about 60.13% of the total verb tokens, and so on.

Also important are the information provided in Columns J through M, in which the usage scores of respective inflectional variants of a given verb are given. Note that VB (= present tense) and VBP (= non-third-person singular present) are lumped together in the 10th column as "VB|P." Also lumped together in this list are the instances of VBD (= past tense) and VBN (= past participle) since the distinction between the two variants are not clear-cut from their surface forms alone. By scanning this part of the list, we are able to find, for instance, that the verb *thank* (Rank 4) is overwhelmingly used in the present tense, whereas the verb



E08-BLC動詞リスト.xls

Appendix E1

Business Letter Corpus Verb List 1 (Lemmatized Usage Rank List)

Total No. of Adjusted Tokens = 1,092,589 (Based on the tagged BLC. Non-text strings are excluded.)
Total No. of Verbs = 190,658 (Percentage to Total No. of Tokens = 17.45%)
Total No. of Common Verbs = 131,313 (Percentage to Total No. of Tokens = 12.02%)
Total No. of Be, Have, Do = 59,345 (Percentage to Total No. of Tokens = 5.44%)
Total No. of Verb Types = 2,371 (Verb Type-Token Ratio = 1.24 ≒ 1:80)
1) Base forms are listed unless a particular variant has no corresponding base form in the corpus. Entries include such verb-derived words as *pleased, interested, concerning, regarding,* etc.
2) Word Level (WL) tags are assigned by the *Word Level Checker* (Ver.1).
3) Total Freq. - Verb Freq. = Tokens to which non-verb tags are assigned by the Brill Tagger (tagging accuracy = 96.49 %)

Usage Rank	Freq. Rank	Entry Word (VB+) [1]	WL [2]	Total Freq. [3]	Verb Freq.	% to Verb Tokens	Cum. Freq.	Cum. %	VB|P Usage	VBD|N Usage	VBG Usage	VBZ Usage	Total Usage
1	1	**BE**	1	39,203	39,203	20.5619	39,203	20.5619	* *See BLC Comprehensive Wordlist for respec-*				
2	2	**HAVE**	1	15,706	15,706	8.2378	54,909	28.7997	*tive frequencies of BE, HAVE, DO and their*				
3	3	**DO**	1	4,436	4,436	2.3267	59,345	31.1264	*variants.*				
4	4	**thank**	1	4,115	3,361	1.7628	62,706	32.8893	2419.7	0.8	32.8	2.0	2455.3
5	5	**make**	1	3,140	3,096	1.6239	65,802	34.5131	1197.7	750.6	255.9	105.4	2309.6
6	6	**know**	1	2,549	2,473	1.2971	68,275	35.8102	1746.8	85.3	41.9	24.0	1898.0
7	7	**look**	1	2,168	2,087	1.0946	70,362	36.9048	1092.1	22.5	338.0	33.8	1486.4
8	8	**give**	1	2,072	2,037	1.0684	72,399	37.9732	869.8	282.7	174.5	60.2	1387.2
9	9	**like**	1	2,549	2,009	1.0537	74,408	39.0269	1342.6	16.4	0.0	2.7	1361.7
10	13	**receive**	1	1,715	1,702	0.8927	76,110	39.9196	439.3	673.1	206.8	5.3	1324.5
11	10	**enclose**	3	2,037	2,000	1.0490	78,110	40.9686	69.8	1159.9	156.8	0.0	1316.7
12	12	**take**	1	1,827	1,820	0.9546	79,930	41.9232	783.4	167.6	259.7	66.4	1277.1
13	15	**send**	1	1,575	1,572	0.8245	81,502	42.7477	692.6	324.6	237.5	4.6	1259.3
14	11	**appreciate**	3	1,848	1,842	0.9661	83,344	43.7139	968.1	243.1	0.0	5.5	1216.7
15	14	**hope**	1	2,028	1,633	0.8565	84,977	44.5704	1082.9	16.7	46.7	0.0	1146.3
16	17	**let**	1	1,429	1,427	0.7485	86,404	45.3188	1092.0	0.0	0.0	0.0	1092.0
17	16	**see**	1	1,481	1,469	0.7705	87,873	46.0893	855.6	26.8	204.2	0.0	1086.6
18	18	**call**	1	1,769	1,387	0.7275	89,260	46.8168	799.3	120.9	76.5	80.0	1076.7
19	19	**work**	1	2,798	1,347	0.7065	90,607	47.5233	384.5	184.6	448.3	18.3	1035.7
20	23	**find**	1	1,168	1,160	0.6084	91,767	48.1317	622.1	220.0	69.8	12.4	924.3
21	20	**get**	1	1,213	1,206	0.6325	92,973	48.7643	602.6	95.6	150.6	23.8	872.6
22	21	**write**	1	1,262	1,197	0.6278	94,170	49.3921	263.9	199.4	404.1	0.0	867.4
23	53	**follow**	1	1,095	581	0.3047	94,751	49.6968	95.3	35.9	590.7	124.5	846.4
24	22	**provide**	2	1,188	1,183	0.6205	95,934	**50.3173**	465.3	204.5	94.7	58.6	823.1
25	24	**ask**	1	1,132	1,129	0.5922	97,063	50.9095	418.8	260.4	138.8	2.9	820.9
26	26	**think**	1	1,096	1,050	0.5507	98,113	51.4602	563.6	149.1	86.0	2.4	801.1
27	28	**come**	1	1,025	1,023	0.5366	99,136	51.9968	488.5	64.4	156.5	60.7	770.1
28	27	**want**	1	1,032	1,031	0.5408	100,167	52.5375	624.0	112.2	7.8	24.5	768.5
29	30	**meet**	1	956	956	0.5014	101,123	53.0389	577.2	85.8	58.7	29.1	750.8
30	33	**need**	1	1,719	920	0.4825	102,043	53.5215	608.4	102.2	11.6	22.5	744.7
31	31	**include**	2	957	952	0.4993	102,995	54.0208	243.5	112.3	273.0	95.7	724.5
32	32	**offer**	2	1,230	921	0.4831	103,916	54.5039	436.1	108.6	95.7	58.4	698.8
33	34	**use**	1	1,220	908	0.4762	104,824	54.9801	326.5	205.9	153.6	12.0	698.0
34	35	**wish**	1	1,299	895	0.4694	105,719	55.4495	642.5	4.7	34.1	3.8	685.1
35	37	**accept**	2	886	875	0.4589	106,594	55.9085	507.7	90.4	30.1	2.9	631.1
36	41	**please**	1	4,703	737	0.3866	107,331	56.2950	0.0	618.7	0.0	1.6	620.3
37	36	**feel**	1	968	879	0.4610	108,210	56.7561	521.3	75.5	3.3	13.9	614.0
38	39	**discuss**	2	806	804	0.4217	109,014	57.1778	414.3	151.6	47.1	0.0	613.0
39	42	**hear**	1	1,183	721	0.3782	109,735	57.5559	419.7	135.1	8.0	0.0	562.8
40	40	**go**	1	801	772	0.4049	110,507	57.9609	215.6	102.1	191.3	37.7	546.7
41	48	**pay**	1	707	639	0.3352	111,146	58.2960	269.3	179.1	63.4	5.1	516.9
42	44	**keep**	1	716	677	0.3551	111,823	58.6511	383.7	68.0	52.8	4.8	509.3
43	43	**continue**	1	807	710	0.3724	112,533	59.0235	315.5	69.1	86.7	25.4	496.7
44	45	**say**	1	686	669	0.3509	113,202	59.3744	313.8	123.6	47.6	11.3	496.3
45	38	**understand**	1	1,100	844	0.4427	114,046	59.8171	431.7	12.5	7.0	8.3	459.5
46	51	**return**	1	930	597	0.3131	114,643	**60.1302**	220.5	140.4	88.2	7.8	456.9
47	49	**interest**	1	1,554	622	0.3262	115,265	60.4564	16.2	432.3	0.0	0.0	448.5
48	47	**request**	2	1,439	641	0.3362	115,906	60.7926	105.6	266.1	73.7	2.1	447.5

Figure 3-10　Business Letter Corpus Verb List 1 (MS Excel Screen Shot)
(See Appendix E1 for the complete wordlist)

enclose (Rank 11) is most likely to be used in either the past or past participle forms and only secondary in the progressive form. The last column shows the aggregated usage scores, based on which the entries are being ranked in the current list.

2) BLC Wordlists for Adverbs, Adjectives and Nouns

The technical procedure for compiling the wordlists for adverbs, adjectives and nouns is basically the same as that of the verb list mentioned above, except that slightly modified versions of `vb.awk` (Figure 3-6) were used to extract relevant word-tag combinations.[19]

Each of the preliminary wordlists compiled via these revised AWK programs was then used as the base files in the subsequent step in which the 34 BLC subcorpora were run through `mk_list.awk` with the BASEFILE option. This yielded three different sets of the CONSOLIDATED WORD FREQUENCY COMPARISON TABLE in the same format as the one shown in Figure 3-8. By importing these frequency comparison tables to MS Excel one by one and making the necessary adjustments and calculation, we have obtained the final versions of the categorical wordlist for each the three POS categories (Figures 3-11 through 3-13).[20]

Figure 3-11 shows the first page of the BLC ADVERB LIST 1 (USAGE RANK LIST). From the 1st to 8th columns (Columns A through H) of this list are given the same types of information as those in the first nine columns of the verb list. The 9th column, or Column I, contains the average frequency of each adverb entry in the 34 subcorpora. It shows that the adverb *sincerely* (Rank 1), for instance, occurs an average of 201.21 times per subcorpus.

In Column K is given the coefficient of dispersion (= Disp.) for each entry. As noted earlier, the dispersion value takes any number between 0 and 1, where "0" means that the distribution of the item in question is maximally skewed and "1" highly even. The dispersion of 0.7 obtained for *sincerely*, for instance, indicates that this adverb is not only very frequent in the BLC, but is also fairly evenly distributed among the 34 subcorpora. One possible interpretation of this is that, since this adverb is mostly used in the complimentary close of the business messages contained in the BLC, we may just as well hypothesize that "*sincerely*" is the most common form of complimentary close in today's business messages.[21] On the other hand, the dispersion of 0.14 for the adverb *consequently* (Usage Rank 122; Frequency Rank 54) means that the distribution of this adverb is highly skewed and that its use is most likely to be concentrated in a particular subcorpus. A closer look at the corpus data reveals that about 85% of all the occurrences of *consequently* (*i.e.* 181 times out of 212) are found in Subcorpus BZ16. The difference between the usage and frequency rankings properly reflects the idiosyncrasy of this particular lexical item in the BLC and, in effect, confirms the viability of the

notion of "Usage." The last column shows the usage scores and, as was also the case with the verb list, all the adverbs in the current list have been ranked as per their respective usage scores.

Figure 3-12 shows the first page of the BLC ADJECTIVE LIST 1 (USAGE RANK LIST). The organization of this list is exactly the same as that of the adverb list and, therefore, is not explained here. For a full list, see Appendix E6, Vol. 2.

In Figure 3-13 is the first page of the BLC NOUN LIST 1. What is shown here is a lemmatized frequency comparison table containing a total of 3,136 noun types. The composition of the list is fairly self-explanatory, but several comments as to the nature of this list are nevertheless in order. First, all the minor entries with a total frequency of less than five have been omitted from the list for the reason mentioned in Footnote 9 of this chapter. I have also excluded most of the proper nouns including the names of people, companies, organizations, products, countries, cities, and so on. Abbreviations and acronyms have been included if Freq.≧ 5. In case the same word appears in different graphic forms, such as "slow-down" and "slowdown," whichever is more common has been adopted in the current list. For some entries that require a special comment, I have added a footnote. For instance, the entry *name* (Usage Rank 4) is given the following footnote:

[1] Mostly used as a dummy name as in "Dear < NAME >."

This means that the observed frequency and the resulting usage score of this particular item should not be taken in their face values. The apparent contradiction of including the entry *new* (Usage Rank = 70) in this "noun" list is also explained in Footnote [2] which reads:[22]

[2] Mostly used as part of proper nouns *(e.g.* New York).

As with other categorical wordlists, the entries of this list (See Appendix E8 for a full list) have been sorted in descending order of the usage scores given in the last column of the list. The appropriateness of using the usage score as the primary measure of the relative importance of respective entries is again clearly demonstrated in the current noun list. Compare, for instance, the two entries *appendix* (Rank 2429) and *sulfur* (Rank 3128). These two nouns occur five times each in the entire corpus. However, the former is distributed in five different subcorpora, whereas the latter occurs only in one particular subcorpus. This distributional difference has been properly captured in the following statistics quoted from the list in a summarized form:

Entry	N (Freq.)	Disp.	Usage Score	Ranking
appendix	5	0.57	2.87	2429
sulfur	5	0.02	0.08	3128

Had we adopted the frequency data alone, we would have reached an improper conclusion that these two nouns are of equal importance.

3.4 Summary

In this chapter, some of the major wordlists compiled for the current study have been introduced and their technical composition briefly explained. We have also reviewed the technical procedures of wordlist compilation in some detail because this author believes that *how* we get data is equally important as, or perhaps more important than, *what* data we get.

Although we haven't yet discussed the contents of these wordlists in terms of what they tell us about the lexico-grammatical characteristics of written Business English, it should be mentioned at this point that these wordlists in themselves constitute a major accomplishment considering the fact that no similar data have ever been available to date. Also significant, albeit on a more practical side, is the compilation of a series of computer programs. Of particularly importance is the multi-function wordlist compiler, `mk_list.awk`, without which the current study would have been impossible. Since this author also believe that other researchers interested in lexical analysis would greatly benefit from this program, the full program source has been included in Appendix C2, so that it can be tested, modified or otherwise used at the user's disposal.[23]

Now that we have all the primary data—the Corpora and the Wordlists[24]—we are ready to move on to the next phase of the study; that is, to begin to look more closely into the data and to discuss some of the major characteristics of the BLC lexicon. To this we now turn.

E10-BLC副詞と形容詞リスト.xls

Appendix E4

Business Letter Corpus Adverb List 1 (Usage Rank List)

☞ All the instances of adverbs (RB+) in BLC subcorpora except BZ09, BZ11, and BZ14 are listed in their graphic forms.
Total No. of Word Tokens = 1,092,589 (Based on the tagged BLC. Non-text strings are excluded)
Total No. of Adverbs (RB, RBR, RBS) = 61,949　Ratio to Total Word Tokens = 5.67 %
No. of RB Word Types (after adjustment) = 740　RB Type-Token Ratio ≒ 1:81 (1.23 %)
Cum Freq.= Cumulative frequency　Disp. = 1-SD÷(M×SQRT(n-1))　Usage = Sum×Disp.

Usage Rank	Freq. Rank	Entry Word (RB+)	WL	Freq.	Cum. Freq.	% to Total RB	Cum. %	Ave. Freq. per Subcorpus (N=34)	SD	Disp.	Usage
1	1	**sincerely**	6	6,841	6,841	11.0430	11.04	201.21	351.61	0.70	4,760
2	2	**not**	1	4,222	11,063	6.8153	17.86	124.18	165.47	0.77	3,243
3	3	**please**	1	3,957	15,020	6.3875	24.24	116.38	145.00	0.78	3,099
4	4	**very**	1	3,509	18,529	5.6643	29.91	103.21	169.42	0.71	2,506
5	7	**so**	1	1,625	20,154	2.6231	32.53	47.79	55.52	0.80	1,296
6	6	**forward**	1	1,754	21,908	2.8314	35.36	51.59	88.47	0.70	1,230
7	5	**much**	1	1,768	23,676	2.8540	38.22	52.00	106.44	0.64	1,138
8	8	**now**	1	1,576	25,252	2.5440	40.76	46.35	76.87	0.71	1,121
9	10	**as**	1	1,335	26,587	2.1550	42.91	39.26	59.18	0.74	985
10	9	**also**	1	1,459	28,046	2.3552	45.27	42.91	80.18	0.67	984
11	11	**again**	1	1,257	29,303	2.0291	47.30	36.97	75.35	0.65	811
12	15	**just**	1	1,010	30,313	1.6304	48.93	29.71	36.14	0.79	796
13	12	**however**	1	1,230	31,543	1.9855	**50.91**	36.18	75.97	0.63	780
14	14	**soon**	1	1,092	32,635	1.7627	52.68	32.12	55.83	0.70	762
15	17	**only**	1	956	33,591	1.5432	54.22	28.12	36.23	0.78	742
16	16	**well**	1	998	34,589	1.6110	55.83	29.35	44.75	0.73	733
17	13	**here**	1	1,118	35,707	1.8047	57.64	32.88	75.32	0.60	672
18	18	**most**	1	876	36,583	1.4141	59.05	25.76	37.72	0.75	653
19	19	**more**	1	873	37,456	1.4092	**60.46**	25.68	40.13	0.73	636
20	21	**therefore**	2	641	38,097	1.0347	61.49	18.85	28.75	0.73	471
21	20	**truly**	4	644	38,741	1.0396	62.53	18.94	30.49	0.72	464
22	22	**always**	1	533	39,274	0.8604	63.39	15.68	20.47	0.77	412
23	23	**today**	1	532	39,806	0.8588	64.25	15.65	23.23	0.74	394
24	25	**together**	1	471	40,277	0.7603	65.01	13.85	17.79	0.78	366
25	24	**even**	1	484	40,761	0.7813	65.79	14.24	21.32	0.74	358
26	28	**back**	1	443	41,204	0.7151	66.51	13.03	18.25	0.76	335
27	29	**still**	1	417	41,621	0.6731	67.18	12.26	14.98	0.79	328
28	32	**then**	1	382	42,003	0.6166	67.80	11.24	11.38	0.82	315
29	26	**already**	1	464	42,467	0.7490	68.55	13.65	28.05	0.64	298
30	31	**further**	2	386	42,853	0.6231	69.17	11.35	16.13	0.75	291
31	36	**immediately**	4	349	42,816	0.5634	69.11	10.26	12.53	0.79	275
32	37	**too**	1	348	43,164	0.5618	69.67	10.24	12.95	0.78	271
33	40	**once**	1	326	43,490	0.5262	**70.20**	9.59	10.38	0.81	265
34	30	**certainly**	2	408	43,898	0.6586	70.86	12.00	24.55	0.64	263
35	33	**really**	1	370	44,268	0.5973	71.46	10.88	19.53	0.69	254
36	35	**recently**	3	352	44,620	0.5682	72.02	10.35	19.30	0.68	238
37	41	**quite**	1	315	44,935	0.5085	72.53	9.26	14.38	0.73	230
38	43	**perhaps**	1	282	45,217	0.4552	72.99	8.29	9.26	0.81	227
39	27	**unfortunately**	4	452	45,669	0.7296	73.72	13.29	42.73	0.44	199
40	38	**cordially**	11	334	46,003	0.5392	74.26	9.82	23.22	0.59	197
41	48	**yet**	1	232	46,235	0.3745	74.63	6.82	6.45	0.84	194
42	45	**greatly**	4	253	46,488	0.4084	75.04	7.44	13.13	0.69	175
43	51	**ago**	1	224	46,712	0.3616	75.40	6.59	8.81	0.77	172
44	34	**there**	1	358	45,856	0.5779	74.02	6.62	9.49	0.75	169
45	49	**yesterday**	1	228	43,044	0.3680	69.48	6.71	10.61	0.72	165
46	55	**rather**	2	211	43,255	0.3406	69.82	6.21	7.78	0.78	165
47	44	**far**	1	261	43,516	0.4213	70.24	7.68	16.31	0.63	164
48	52	**no**	1	218	43,734	0.3519	70.59	6.41	9.57	0.74	161
49	47	**especially**	2	238	43,972	0.3842	70.98	7.00	12.97	0.68	161

Figure 3-11　Business Letter Corpus Adverb List 1 (MS Excel Screen Shot)
(See Appendix E4 for the complete wordlist)

E10-BLC副詞と形容詞リスト.xls

Appendix E6

Business Letter Corpus Adjective List 1 (Usage Rank List)

☞ All the instances of adjectives (JJ+) in BLC subcorpora except BZ09, BZ11, and BZ14 are listed in their graphic forms.
Total No. of Word Tokens = 1,092,589 (Based on the tagged BLC. Non-text strings are excluded)
Total No. of Adjectives (JJ, JJR, JJS) = 76,820　Ratio to Total Word Tokens = 7.03 %
No. of JJ Word Types = 3,923 (Freq. ≧ 2 = 2,064)　JJ Type-Token Ratio ≒ 1:19 (5.11 %)
Cum. Freq.= Cumulative frequency　Disp. = 1-SD ÷ (M × SQRT(n-1))　Usage = Sum × Disp.

Usage Rank	Freq. Rank	Entry Word (JJ+)	WL	Freq.	Cum. Freq.	% to Total JJ	Cum. %	Ave. Freq. per Subcorpus (N=34)	SD	Disp.	Usage
1	1	**dear**	1	7,684	7,684	10.0026	10.00	226.00	290.01	0.78	5,968
2	2	**new**	1	2,089	9,773	2.7193	12.72	61.44	64.97	0.82	1,704
3	3	**other**	1	1,398	11,171	1.8198	14.54	41.12	50.99	0.78	1,096
4	5	**next**	1	1,245	12,416	1.6207	16.16	36.62	44.96	0.79	1,011
5	4	**best**	1	1,263	13,679	1.6441	17.81	37.15	39.53	0.81	997
6	6	**such**	1	1,080	14,759	1.4059	19.21	31.76	50.15	0.73	783
7	8	**good**	1	980	15,739	1.2757	20.49	28.82	38.71	0.77	769
8	7	**more**	1	993	16,732	1.2926	21.78	29.21	35.59	0.79	764
9	11	**last**	1	897	17,629	1.1677	22.95	26.38	20.76	0.84	739
10	9	**many**	1	904	18,533	1.1768	24.13	26.59	26.68	0.82	702
11	15	**first**	1	751	19,284	0.9776	25.10	22.09	21.83	0.80	628
12	12	**able**	1	876	20,160	1.1403	26.24	25.76	42.13	0.72	627
13	10	**possible**	2	901	21,061	1.1729	27.42	26.50	34.08	0.78	608
14	14	**sure**	1	754	21,815	0.9815	28.40	22.18	24.74	0.81	608
15	17	**available**	3	634	22,449	0.8253	29.22	18.65	30.98	0.73	505
16	13	**happy**	1	786	23,235	1.0232	30.25	23.12	49.57	0.67	503
17	16	**great**	1	670	23,905	0.8722	31.12	19.71	47.79	0.64	487
18	18	**few**	1	552	24,457	0.7186	31.84	16.24	18.03	0.81	445
19	19	**special**	1	547	25,004	0.7121	32.55	16.09	21.29	0.77	421
20	22	**sorry**	1	501	25,505	0.6522	33.20	14.74	25.66	0.70	399
21	20	**due**	2	543	26,048	0.7068	33.91	15.97	17.21	0.80	378
22	21	**recent**	2	506	26,554	0.6587	34.57	14.88	27.80	0.70	354
23	27	**several**	1	441	26,995	0.5741	35.14	12.97	16.25	0.77	351
24	24	**personal**	2	484	27,479	0.6300	35.77	14.24	23.03	0.72	348
25	29	**full**	1	416	27,895	0.5415	36.31	12.24	26.23	0.67	323
26	33	**current**	2	382	28,277	0.4973	36.81	11.24	34.75	0.59	315
27	30	**past**	1	410	28,687	0.5337	37.34	12.06	23.70	0.68	314
28	28	**necessary**	1	428	29,115	0.5571	37.90	12.59	19.39	0.73	313
29	25	**same**	1	467	29,582	0.6079	38.51	13.74	15.71	0.78	312
30	26	**present**	1	445	30,027	0.5793	39.09	13.09	11.36	0.82	305
31	32	**better**	1	386	30,413	0.5025	39.59	11.35	9.52	0.84	304
32	38	**free**	1	351	30,764	0.4569	40.05	10.32	12.59	0.78	295
33	23	**additional**	4	500	31,264	0.6509	40.70	14.71	15.15	0.80	289
34	31	**long**	1	400	31,664	0.5207	41.22	11.76	13.88	0.79	284
35	37	**own**	1	363	32,027	0.4725	41.69	10.68	15.59	0.75	284
36	34	**financial**	4	380	32,407	0.4947	42.19	11.18	19.52	0.71	275
37	36	**important**	1	367	32,774	0.4777	42.66	10.79	17.71	0.72	275
38	39	**excellent**	2	333	33,107	0.4335	43.10	9.79	21.47	0.66	258
39	35	**effective**	3	369	33,476	0.4803	43.58	10.85	13.33	0.78	242
40	41	**complete**	2	322	33,798	0.4192	44.00	9.47	18.07	0.67	241
41	40	**further**	2	328	34,126	0.4270	44.42	9.65	13.73	0.75	221
42	45	**small**	1	267	34,393	0.3476	44.77	7.85	11.04	0.75	216
43	44	**outstanding**	3	280	34,673	0.3645	45.14	8.24	9.27	0.79	212
44	49	**high**	1	257	34,930	0.3345	45.47	7.56	10.84	0.74	211
45	48	**local**	2	258	35,188	0.3359	45.81	7.59	16.83	0.65	203
46	47	**grateful**	2	262	35,450	0.3411	46.15	7.71	8.84	0.78	197
47	55	**large**	1	236	35,686	0.3072	46.45	6.94	6.68	0.81	191
48	53	**major**	3	238	35,924	0.3098	46.76	7.00	20.29	0.59	186
49	43	**technical**	3	285	36,209	0.3710	47.13	8.38	7.74	0.82	185

Figure 3-12　Business Letter Corpus Adjective List 1 (MS Excel Screen Shot)

(See Appendix E6 for the complete wordlist)

E11-BLC名詞リスト.xls

Appendix E8

Business Letter Corpus Noun List 1 (Lemmatized Frequency Comparison Table)

☞ All the instances of nouns (NN+) in BLC subcorpora except BZ09, BZ11, and BZ14 are listed in their lemmatized forms.
Total No. of Word Tokens = 1,092,589 (Based on the tagged BLC. Non-text strings are excluded)
Total No. of Nouns (NN, NNS, NNP, NNPS) = 268,851　Ratio to Total Word Tokens = 24.60 %
No. of Noun Word Types = **3,136**　NN Type-Token Ratio ≒ 1:14 (6.93 %)
Disp. = 1-SD ÷(M ×SQRT(n-1))　U (Usage) = Sum ×Disp.

N	NOUN (BaseForm)	BZ01	BZ02	BZ03	BZ04	BZ05	BZ06	BZ08	BZ09	BZ10	BZ12	BZ13	BZ15	BZ16	BZ17	BZ18	BZ19	BZ20	BZ21	BZ22	BZ23	BZ24	BZ25	BZ26	BZ27	BZ28	BZ29	BZ30	BZ31	BZ32	BZ33	BZ34	BZ35	BZ36	BZ37	Sum	Mean	SD	Disp.	U
1	**Mr.**	61	104	2	98	30	28	0	145	145	8	3	25	1	50	12	1716	45	1	131	135	54	413	28	0	124	66	12	81	213	131	256	87	171	27	**4403**	129.50	293.85	0.61	**2663.8**
2	**time**	173	105	14	134	111	128	3	169	149	89	68	188	62	6	305	813	7	48	30	108	15	220	10	61	24	22	25	95	83	36	58	20	33	49	**3461**	101.79	144.13	0.75	**2607.9**
3	**company**	84	80	11	170	105	128	2	274	121	99	31	117	38	14	235	326	38	41	28	83	18	142	13	1	22	16	14	56	9	80	98	52	110	22	**2678**	78.76	78.46	0.83	**2213.6**
4	**name** [1]	233	30	0	24	333	264	1	53	28	170	176	99	150	2	184	41	13	6	13	28	2	25	2	256	0	1	3	13	18	28	24	4	3	10	**2237**	65.79	93.71	0.75	**1682.4**
5	**letter**	37	66	3	43	28	28	8	94	60	34	27	76	28	4	88	760	3	5	53	164	24	141	9	47	50	28	4	11	82	94	154	55	54	29	**2391**	70.32	129.01	0.68	**1627.4**
6	**date**	90	35	16	12	88	61	3	233	113	97	6	181	27	8	322	178	19	72	16	31	7	25	1	206	15	17	12	13	18	11	30	11	20	7	**2001**	58.85	78.14	0.77	**1538.5**
7	**business**	53	73	3	136	86	118	4	106	133	65	54	108	16	14	138	289	17	29	11	81	11	42	10	72	15	20	21	19	6	33	0	0	53	40	**1876**	55.18	59.58	0.81	**1523.3**
8	**order**	58	67	1	11	100	59	7	54	83	80	38	54	20	28	144	167	1	8	76	209	36	38	16	46	12	32	26	77	9	135	37	5	39	20	**1793**	52.74	49.65	0.84	**1499.1**
9	**service**	47	42	9	56	103	92	3	143	212	87	91	163	16	14	134	248	4	20	9	44	9	70	6	30	24	12	6	38	11	26	13	8	10	31	**1831**	53.85	62.29	0.80	**1462.3**
10	**product**	80	37	13	45	41	47	5	70	190	101	94	98	22	8	123	333	5	7	34	41	5	24	1	51	0	10	3	46	1	82	43	2	74	5	**1741**	51.21	65.91	0.78	**1350.9**
11	**office**	26	36	11	24	77	58	2	137	89	67	26	51	9	27	49	196	14	14	7	24	5	69	6	0	20	27	22	42	33	32	140	33	41	20	**1434**	42.18	43.23	0.82	**1178.1**
12	**information**	42	43	10	58	66	33	2	122	109	48	31	81	17	17	77	292	8	6	19	54	4	19	1	17	14	17	19	34	61	27	13	23	14	33	**1431**	42.09	53.13	0.78	**1116.6**
13	**work**	42	15	12	70	48	38	2	145	40	19	11	87	4	5	80	223	2	32	11	76	16	100	3	16	14	10	10	14	52	7	155	16	7	23	**1405**	41.32	50.82	0.79	**1104.2**
14	**day**	83	37	6	41	51	62	5	103	71	45	30	44	16	6	115	237	5	6	23	65	13	84	5	5	17	12	11	61	41	12	31	2	9	19	**1373**	40.38	46.32	0.80	**1098.9**
15	**meeting**	30	30	7	47	62	60	0	159	67	26	14	70	6	17	67	242	13	16	7	14	4	95	6	23	29	42	15	97	42	5	19	9	37	10	**1387**	40.79	49.25	0.79	**1095.5**
16	**sale | sales**	34	56	1	43	61	161	8	82	138	42	59	42	9	17	22	239	8	29	22	69	13	103	6	0	5	3	4	19	3	15	0	8	43	21	**1385**	40.74	52.27	0.78	**1075.7**
17	**year**	30	67	1	133	51	73	3	135	61	27	15	38	5	20	68	184	6	14	11	70	18	138	8	17	4	9	9	20	24	24	2	2	18	25	**1330**	39.12	46.02	0.80	**1057.6**
18	**customer**	56	74	4	21	51	63	6	102	173	66	59	60	17	9	119	99	1	25	42	90	6	109	7	10	2	2	4	12	0	10	0	0	11	6	**1316**	38.71	43.98	0.80	**1055.7**
19	**week**	41	42	8	47	96	113	1	120	81	48	41	62	3	14	35	93	3	23	29	58	18	80	7	4	14	12	13	65	21	20	5	5	8	13	**1243**	36.56	34.17	0.84	**1040.8**
20	**month**	55	20	5	25	72	61	1	92	76	61	36	37	6	17	63	155	3	35	26	95	14	79	10	1	10	6	9	43	24	19	25	0	20	21	**1222**	35.94	34.79	0.83	**1016.1**
21	**number**	104	27	6	25	26	52	5	44	28	42	24	65	40	7	412	80	11	10	17	127	7	13	0	91	6	2	8	25	30	25	12	3	9	15	**1398**	41.12	72.42	0.69	**969.39**
22	**account**	57	16	0	3	65	21	0	48	117	39	32	31	42	16	77	53	3	2	47	128	27	114	5	24	14	4	0	32	1	14	15	2	5	11	**1065**	31.32	34.74	0.81	**859.41**
23	**price**	31	31	2	11	41	42	6	21	54	23	25	25	8	9	84	365	8	7	16	116	18	10	8	13	3	24	3	42	3	49	31	13	39	18	**1199**	35.26	63.01	0.69	**826.05**
24	**subject**	12	23	12	9	59	37	0	173	3	0	1	6	1	8	16	126	29	63	14	13	1	22	1	22	15	24	105	369	7	35	6	8	6	8	**1234**	36.29	70.22	0.66	**818.37**
25	**payment**	34	14	0	1	51	26	5	27	112	18	12	50	26	5	67	138	3	1	21	102	12	7	2	32	15	9	5	52	1	49	48	12	8	8	**973**	28.62	33.65	0.80	**773.82**
26	**copy**	18	60	6	29	36	18	3	58	45	14	23	13	1	6	31	289	8	4	22	66	6	41	5	7	22	35	18	7	38	21	46	25	4	15	**1040**	30.59	48.95	0.72	**750.27**
27	**opportunity**	2	29	0	113	28	65	2	84	99	21	23	33	1	3	49	147	1	5	6	45	1	75	7	5	4	6	3	8	42	15	2	25	1	18	**968**	28.47	36.87	0.77	**749.78**
28	**position**	16	62	1	167	40	31	0	124	18	3	3	52	14	8	107	210	1	10	2	10	9	39	4	11	5	3	6	11	24	12	3	1	5	12	**1024**	30.12	49.54	0.71	**730.81**
29	**program /-mme**	12	27	6	41	68	51	1	105	67	70	17	43	3	5	21	122	1	17	2	5	0	24	1	0	4	3	3	0	182	1	1	22	0	44	**969**	28.50	41.41	0.75	**723.91**
30	**credit**	33	39	1	5	49	27	2	20	57	20	11	31	42	14	120	23	0	6	38	80	4	39	5	17	4	3	4	14	0	27	53	4	0	4	**796**	23.41	26.22	0.81	**640.80**
31	**May**	5	34	3	10	42	15	1	61	36	23	11	2	1	1	3	259	3	10	19	42	10	92	10	3	23	37	11	21	24	24	24	16	16	12	**904**	26.59	45.31	0.70	**635.81**
32	**way**	21	30	7	23	26	40	3	78	61	16	20	28	6	6	43	277	0	12	6	25	3	119	5	13	1	2	2	17	26	6	5	3	1	4	**935**	27.50	50.64	0.68	**635.30**
33	**cost**	26	15	24	18	37	59	2	86	51	22	23	17	1	20	32	164	1	11	5	41	4	5	6	0	15	11	3	28	8	10	42	11	11	5	**814**	23.94	31.15	0.77	**629.62**
34	**problem**	30	20	14	31	43	38	1	104	56	22	16	27	3	2	83	392	0	21	12	10	7	15	0	4	5	3	3	31	7	17	1	3	2	7	**1030**	30.29	68.07	0.61	**627.11**
35	**department**	26	33	37	23	26	26	1	74	15	33	9	26	3	13	42	71	3	20	5	39	11	71	4	0	9	0	5	36	12	28	4	10	21	5	**741**	21.79	20.18	0.84	**621.55**
36	**need**	10	20	6	42	45	32	0	87	93	41	41	53	4	3	51	132	1	14	5	25	1	10	0	3	10	5	3	25	8	10	1	2	2	10	**795**	23.38	30.74	0.77	**613.05**
37	**job**	31	42	7	131	27	64	0	110	33	7	3	112	14	2	83	21	1	28	6	3	5	42	1	1	10	4	11	1	4	1	2	2	0	3	**812**	23.88	35.50	0.74	**601.91**
38	**question**	17	8	6	19	43	34	2	77	55	58	43	42	5	0	83	100	1	8	2	11	0	11	4	2	3	16	8	44	14	12	3	2	6	15	**754**	22.18	26.45	0.79	**597.43**
39	**people**	22	69	1	57	35	48	1	73	19	18	9	64	4	2	34	166	3	15	7	4	2	32	2	1	8	1	2	10	54	6	5	0	2	16	**792**	23.29	33.59	0.75	**593.18**
40	**project**	14	6	1	40	24	20	1	97	50	18	2	52	1	1	86	142	1	8	0	0	0	12	0	15	0	4	1	16	31	12	69	37	4	18	**783**	23.03	32.83	0.75	**588.70**
41	**thank**	12	16	0	37	25	80	2	46	29	65	42	49	9	0	50	40	3	2	1	9	0	131	3	12	3	4	8	22	35	2	1	5	7	3	**753**	22.15	28.57	0.78	**583.88**
42	**area**	6	4	11	40	14	25	2	79	61	21	12	36	2	1	36	183	0	5	7	31	5	12	1	17	3	2	0	16	19	14	102	5	2	11	**785**	23.09	36.51	0.72	**568.93**
43	**interest**	16	13	2	56	22	14	3	56	63	21	10	23	10	8	14	236	4	0	9	24	2	37	2	4	4	5	2	20	56	16	1	15	9	23	**800**	23.53	41.34	0.69	**555.32**
44	**manager**	19	49	5	60	35	39	0	57	32	33	5	13	2	6	20	79	7	35	9	20	9	29	3	1	10	18	2	10	3	5	19	4	17	13	**668**	19.65	19.33	0.83	**553.60**
45	**staff**	30	15	0	33	23	27	0	85	73	20	12	35	0	4	11	114	2	22	13	8	4	32	5	0	5	6	6	6	46	13	17	1	14	6	**688**	20.24	25.66	0.78	**536.11**

名詞(使用度順) | English | ABC | 累積比率 | コア名詞 | コア名詞(ABC) | Sheet8 | Sheet9 | Sheet10

Figure 3-13　Business Letter Corpus Noun List 1 (MS Excel Screen Shot: See Appendix E8 for the complete wordlist)

Endnotes to Chapter 3

[1] The type-token ratio is sometimes referred to as the lexical density. Ure (1971), for instance, used the latter term to mean what is usually meant today by the type-token ratio. *Longman Dictionary of Applied Linguistics* (1985) also adopts this definition of the lexical density following Ure. For Stubbs (1996: p.72), however, the lexical density of a text or a corpus is "the proportion of *lexical words* (*i.e.* nouns, adjectives, adverbs, and main verbs) expressed as a percentage (to the total number of words in a given text)." If N is the number of words in a text, and L is the number of lexical words, then the lexical density of that text can be obtained by the formula, $100 \times L/N$. Generally speaking, there is a strong tendency "for written texts to have a lexical density of over 40%, and for spoken texts to be under 40%." (Stubbs, *ibid.*). Halliday gives still another definition of the lexical density from a functional sentence perspective. He states that the lexical density of a sentence, or a text, is the proportion of the number of lexical words to the number of clauses in that sentence/text (Halliday 1994, p. 351). In other words, the lexical density as defined by Halliday is a measure of how densely a clause as a basic unit of a message is packed with information-carrying lexical words. My use of the term in this paper, however, is close to that of Ure; namely, it refers to the type-token ratio of a text or a corpus, unless otherwise mentioned expressly.

[2] To do this, however, a lemma dictionary has to be created by the user. Also, for this operation to be meaningful, the dictionary should contain as many words as practically possible. The dictionary file compiled by the author for the current study (`e_lemma.dic`) contains 40,569 graphic words in 14,762 lemma groups. Although this is far from complete, it nevertheless is the largest and the most comprehensive lemma dictionary available to date in the knowledge of this author. The dictionary has been compiled in the following format. The full list can be downloaded via the Internet from Mike Scott's Website: http://www.liv.ac.uk/~ms2928/index.htm (as of Aug. 22, 1999).

```
Sample excerpt from e_lemma.dic (Ver.1)
=======================================================================
[ e_lemma.dic (Ver.1)
[ Compiled by Yasumasa Someya, September 1, 1998.
[ This lemma list is provided "as is" and is free to use for any
[ research and/or educational purposes.
(...several lines omitted)
a -> an
A-bomb -> A-bombs
abacus -> abacuses
abandon -> abandons,abandoning,abandoned
abase -> abases,abasing,abased
abate -> abates,abating,abated
abbreviate -> abbreviates,abbreviating,abbreviated
```

```
abbreviation -> abbreviations
abdicate -> abdicates,abdicating,abdicated
abdomen -> abdomens
abduct -> abducts,abducting,abducted
aberration -> aberrations
abet -> abets,abetting,abetted
abhor -> abhors,abhorring,abhorred
abide -> abides,abided,abode,abiding
ability -> abilities
(...)
=====================================================================
```

[3] The JACET list has been compiled in reference to a number of studies related to both general and educational lexical studies, including Kuc&era and Francis (1967), Carroll *et al.* (1971), Francis and Kuc&era (1982), Pheby (1981) and *Zen Eiren* (1981, 1988), and is considered one of the most authentic educational English wordlist compiled to date, despite the fact that it largely ignores informal lexical items that are often used in conversational English.

[4] For further details of the "*D4000*" wordlist, see Kaneda (1991, pp. 41-53). The development of the AWK program package, the *Word Level Checker (Ver. 1)* which I shall refer to shortly, was inspired by Professor Kaneda's work.

[5] The WL-tag dictionary file is a plain text file in one-record-per-line format. Each record is terminated by a return code (¥n). Each record consists of three data fields, $1 to $3, delimited by space. In Field 1 ($1) is contained the entry word, followed by a corresponding WL tag in Field 2 ($2) and by a POS tag in Field 3 ($3) respectively, as shown below:

```
Sample excerpt from "wrdlvl-2.dic"
=====================================================================
a 01 DT
a-bomb 17 NN
a.c. 05 NN
a.d.r. 05 NN
a.m. 05 NN
aback 17 RB
abaft 30 NN
abandon 04 VB
abandoned 06 VBD
abandoning 06 VBG
abandonment 06 NN
abandons 06 VBZ
[...]
```

==========

[6] A word-level profile of a given text can be obtained by first assigning a "word-level (WL) tag" to each of the lexical items in that text, and then counting and cross-tabulating the numbers of respective WL tags for further computation of the necessary data. All the tasks required of this operation can be carried out automatically by using the *Word Level Checker (Ver. 1)* — an AWK program package developed by this author. A sample computer output of this program can be seen in Figure 4-6, Chapter 4. For further details, see Someya (1998c and 1998d).

[7] This produces an outcome which looks roughly as follows (For more details about lemmatization with WordSmith, See Scott (1998, pp. 83-85)):

```
(WordSmith Wordlist for the BLC) ====================================
N     Word Freq.     %    Lemmas
1     the  44,923 4.0125
2     to   41,165 3.6768
3     be   38,369 3.4271 am (2335), 'm (1530), are (6822), is (9250),
                         was (3166), were (1306), being (677), been (2878)
4     you  30,130 2.6912
5     of   27,577 2.4632
6     and  24,890 2.2232
7     a    21,446 1.9155 an (3351)
8     your 20,148 1.7996
9     in   19,904 1.7778
10    I    18,550 1.6569
11    we   18,036 1.6110
12    for  17,863 1.5955
13    have 16,222 1.4489 has (3230), having (442), had (1332), 'd (675),
                         've (1097)
[...]
=========================================================================
```

Figure 3-1 Sample output of WordSmith wordlist for the BLC (excerpt)

[8] The program source of *matchnew.awk* is as follows:

```
# ========================================================================
# matchnew.awk (Yasumasa Someya, 7 Oct. 1998)
# Usage: jgawk -f matchnew.awk -v dic=******* INFILE > OUTFILE
# where "******" is the file name of the WL-tag dictionary to be used (The
# default dictionary file is "wrdlvl-2.dic").
# Function: Scan the input file (a word list in "one-entry-per-line"
# format) and if any of the entries matches that of the dictionary file,
# it will  be replaced with a corresponding word-level tag. Non-match
```

```
# data will be printed out as is in the specified output file.
# ===================================================================
{
  printf "¥rReplacing entries with word-level tags. Please wait... %5d
",NR
  > "CON"
  if (flag == 0) {
    word_0=$1
    word_1=$1
    word_3=$2
    while (word_1 > tag_1) {
      if (getline < dic == 0) {
        flag=1; break
      }
      tag_1=$1; tag_2=$2   # $1=word, $2=WLtag (e.g. 01,02...)
    }
    if (word_1==tag_1) {
      gsub(word_1,tag_2,word_1)
      print word_1, word_3
    }
    else print word_0, word_3
  }
   else print $0
}
END {
printf "¥nJob Completed.¥n"> "CON"
}
# ===================================================================
```

Figure 3-2 matchnew.awk (for replacing entries of a wordlist with word-level tags)

[9] Hofland and Johansson (1982) also omitted from their wordlist compiled for the Brown and LOB Corpora the words whose combined expected frequencies on both corpora are less than a total of 10. In the standard χ^2 test of significance, which Hofland and Johansson adopted in their 1982 study, expected frequencies (*ef*) of an item in question must be five or above, since a smaller figure yields unduly large χ^2 values. In other words, any calculation based on $ef < 5$ is considered statistically unreliable. (Saito 1998, p. 83).

[10] These data, however, will be made available via this author's Internet Website for those interested in conducting further analysis of these least frequent items.

[11] Other major tagging errors include those in "*name*" (Rank 42) and "*new*" (Rank 49). Of the 3,225 instances of "*name*," for instance, 973 were identified as verbs. About a half of this, however, are found to be tagging errors (in these instances, the word "*name*" are

used as the header in the front matter of memorandums or e-mail, and the tagger erroneously identified them as sentence-initial imperative VBs. These errors are concentrated in Subcorpus BZ18). Also, many of the noun instances of "*name*" (N=2,237) are the instances of a dummy use of the term as in "*Dear <name>.*"

[12] The size of the combined Reference Corpus is 3,165,931, which is about 2.8976 times larger than the BLC.

[13] The asterisk (*) on the last column of the KEYWORDS LIST indicates that the difference in the frequencies of each entry between the two corpora is statistically significant. For an easy comparison, compare the figures between the 5th column (=normalized BLC frequency) and the 7th column (= Reference Corpus frequency).

[14] See Appendix C2 for the full program source.

[15] The program source of *vb.awk* is as follows:

```
# ================================================================
# vb.awk (by Yasumasa Someya, 10 August, 1998)
# Usage: jgawk -f vb.awk INFILE > OUTFILE
# Function: Extract verbs from a POS-tagged wordlist
# ================================================================
BEGIN {
# FS="¥t"      # Use this option if TAB is used as field delimiter.
IGNORECASE=1   # Ignore case distinction
}
{ printf "¥rExtracting Verbs. Please wait...%5d ",NR > "CON"
  if (match($0,/_VB|_VBD|_VBG|_VBN|_VBZ|_VBP/))
  print $0 | "sortf"
}
# ================================================================
```

Figure 3-6 *vb.awk* (for extracting verbs from a POS-tagged wordlist)

[16] This list, as well as the NORMALIZED FREQUENCY COMPARISON TABLE (BLC VERB LIST 3) which was prepared to make possible a more accurate frequency comparison among the 34 subcorpora, has not been included in the Wordlist Appendices (Vol. 2 of this paper) since these two lists are of secondary importance in view of the main purpose of the current research and since they would have added another 170 pages to the already very thick volume of appendices.

[17] The coefficient of dispersion (= Disp.) is calculated by the following formula:

$$(1\text{-}SD) \div (M \times SQRT(n\text{-}1))$$

The product of this formula takes any value between 0 and 1, where "0" means that the item in question is maximally skewed and "1" evenly distributed. For more details, see Ueda (1998a, pp.44-48)

[18] Of the remaining 754 cases, 753 are used as nouns (See Figure 3-13) and one as an

adjective.

[19] More specifically, the 6th line of the *vb.awk* program (Figure 3-6) was changed to:

```
if (match($0,/_RB|_RBR|_RBS/))
```

to extract adverbs (RB+). Similarly, the same line was changed to:

```
if (match($0,/_JJ|_JJR|_JJS/))
```

to extract adjectives (JJ+), and to:

```
if (match($0,/_NN|_NNS|_NNP|_NNPS/))
```

to extract nouns (NN+) respectively.

[20] I have also compiled the corpus-wise FREQUENCY COMPARISON TABLES for adverbs and adjectives in the same format as the one shown in Figure 3-9, but they are not included in the Wordlist Appendices, Vol. 2, for the same reason mentioned in Footnote 16.

[21] This assumption has been proven correct in my 1998 research paper. For more details, see Someya (1998b).

[22] This entry, however, should have been excluded from the list to be consistent with the stated policy of omitting all the proper nouns including the names of people, companies, organizations, products, countries, cities, and so on.

[23] The *Word Level Checker* referred to in Section 3.1 is also a quite useful tool, but the size of relevant programs and documentation prohibits its inclusion in the current paper as an appendix. For those interested in this program package, I encourage to refer to Someya (1998c and 1998d). The paper and the AWK programs described therein are available upon request, by sending e-mail to the author (ysomeya@gol.com).

[24] I have also compiled several wordlists from the three Reference Corpora, but they are used only for reference purposes and, therefore, are not included in the current paper. The Learner BLC and the wordlists compiled therefrom will be referred to later in Chapter 4.