CHAPTER 2

COMPARATIVE ANALYSIS OF POS DISTRIBUTION

In this chapter, we turn to the analysis of the corpus data from a viewpoint of POS distribution. First, the main rationale for such an analysis will be briefly overviewed. Then, in Section 2.2, the study procedure of the current POS analysis will be describe in some detail. Section 2.3 presents the major results of the analysis, which will be followed by a series of discussions in Section 2.4 on some of the major POS categories whose distribution in the BLC are found statistically significant. Finally, in Section 2.5, we will make a brief comparison between the Native and Learner BLCs to see if there are any noteworthy differences between these two corpora and, if any, to account for such differences.

2.1 The main rationale for POS distribution analysis

The vocabulary of a particular language or a genre can be described and analyzed from a variety of viewpoints. One such viewpoint is to look at it in terms of the frequency and distribution of grammatical word classes, or parts of speech.

The main rationale for analyzing a lexicon or a text from this perspective is that the information thus obtained provide good indicators of the nature of the text in question. Suppose we know, for instance, that the average proportions of nouns and verbs in fairly standardized written English texts is about 28 percent and 11 percent respectively; then, this information can be used as one of the objective yardsticks in judging or predicting the stylistic nature of a given text or in identifying the genre to which it belongs. We can reasonably assume that a text with a considerably higher proportion of nouns than the baseline figure is likely to be some sort of professional writing such as academic and scientific articles, whereas a text that exhibits a much higher proportion of verbs than is normally expected is more likely to belong to the "imaginative" or "personal" genres that are more concerned with human agents and their actions than the description of generalized processes, relationships, or resultant states of being.

The "nominal" versus "verbal" distinction, of course, does not simply mean to count the numbers of nouns and verbs and to compare them, nor is it sufficient to discuss the nature of a given lexicon for whatever purposes with no reference to other variables such as the frequencies and distributional patterns of pronouns, adjectives, adverbs, articles, prepositions and so on. All these grammatical categories are so closely related that no single or selected variables can be discussed meaningfully in isolation. To give a few examples, we know that the nominal style necessarily results in more articles and prepositions than its verbal counterpart.[1] The verbal style, on the other hand, is more likely to be organized around human participants and, therefore, tends to use more personal pronouns as either the external or the internal arguments of the main verb.[2] From a syntactic viewpoint, the verbal style generally means shorter sentences, but it may also mean an added syntactic complexity, with more clauses and, therefore, more predications within a sentence.[3]

Thus, the apparently straightforward nominal-verbal comparison actually entails a host of other linguistic consequences, and researchers have naturally been interested in the study of the frequency and distribution of grammatical word classes and their correlations in order to gain a better understanding as to how languages are structured and used. Most of the earlier research, however, typically focused on only a few variables and the data they used were too small or skewed to make any meaningful generalizations.

It was only in the late 1980s that a large-scale, corpus-based lexico-grammatical study became possible with the compilation of the POS-tagged versions of the Brown and LOB Corpora－almost 15 years after the compilation of the original Brown Corpus in 1967. The first major study was conducted by Francis and Kuc&era (1982), which was soon followed by Johansson and Hofland (1989). Whereas these two studies were done on written English, Altenberg (1990) conducted an analysis of the relative proportions of the word classes in a 50,000-word sample of spoken English taken from the London-Lund Corpus. These three studies have now become the *de facto* standards, as it were, to which subsequent corpus-based studies continue to make reference.

Despite these pioneering works, however, there has been very little serious corpus-based lexical study done on Business English to date, except for some small-scale exploratory studies which we briefly reviewed in Chapter 1. In what follows, I will be presenting some of the major results of one of the first－if not *the* first－such studies, beginning in the current Chapter with an analysis of the frequency and distribution of the major word classes in the one-million-word BLC in comparison with the three reference corpora.

2.2 Study procedure

The current POS analysis has been conducted in roughly the following manner and sequence.

First, POS tags were assigned to the plain text versions of the BLC and two of the three reference corpora, the Brown and TIME Corpora, using the Brill Tagger.[4] As to the LOB Corpus, the tagged version provided by the ICAME was used. Some of the POS tags used in the LOB Corpus, however, have been replaced with corresponding default tags of the Brill Tagger using a special "tag conversion table" that I compiled for this purpose[5], so that all the corpora are assigned with the same tag set. The conversion of tags was done automatically using the lemmatization function of WordSmith.

Next, each of the POS-tagged corpora was run through an AWK program, `prn_tag.awk`, to create POS-tags-only versions of the corpora.[6] Briefly explained, this program takes a POS-tagged text file like ① below as an input, and creates a new file like ②. The field data format of the input file is assumed to be "word_ TAG" (a word joined by a corresponding POS tag via the underbar).

① Original Text (sample input)

```
[BZ01:00005] I_PRP want_VBP to_TO thank_VB you_PRP for_IN all_PDT the_DT
help_NN you_PRP gave_VBD me_PRP on_IN the_DT merger_NN we_PRP concluded_VBD
last_JJ week_NN ._.
[BZ01:00006]
[BZ01:00007] Your_PRPS invaluable_JJ advice_NN smoothed_VBD the_DT way_NN
for_IN our_PRPS success_NN ._.
[BZ01:00008] We_PRP could_MD n't_NEG have_HV done_DON it_PRP without_IN
you_PRP ._.
```

② POS Tags only Text (sample output)

```
[BZ01:00005] PRP VBP TO VB PRP IN PDT DT NN PRP VBD PRP IN DT NN PRP VBD
JJ NN .
[BZ01:00006]
[BZ01:00007] PRPS JJ NN VBD DT NN IN PRPS NN .
[BZ01:00008] PRP MD NEG HV DON PRP IN PRP .
```

The output files were then run through WordSmith to count the frequency and calculate the relative proportion of each word class. The following (Figure 2-3) is a sample output for the POS tagged version of the BLC (Note that the original possessive case marker "`$`" at the end of `PRP$` and `WP$` has been replaced with "`S`"). Finally, the POS frequency outputs for the BLC and the three reference corpora were imported to MS Excel for further processing.

```
==========================================================
WordSmith Tools -- 98/08/28 1:46:26 (for BLC2_id.tag)
all 98 entries

N   Word  Freq.   %        Lemmas
1   NN  276,535  25.3101   NNS(51968), NNP(62838), NNPS(699)
2   IN  145,458  13.3131
3   PRP 133,536  12.2220   PRPS(40604)
4   VB  131,313  12.0185   VBN(23976), VBG(20117), VBP(15140),
                           VBD(12976), VBZ(4479)
5   DT   89,518  8.1932   PDT(632)
6   JJ   71,101  6.5076   JJS(2652), JJR(2365)
7   RB   61,968  5.6717   RBR(1009), RBS(875)
8   BE   39,203  3.5881   BEN(2877), BEG(676)
9   CC   29,685  2.7169
10  TO   26,929  2.4647
11  MD   26,580  2.4328
12  CD   21,010  1.9230
13  HV   15,706  1.4375   HVG(442)
14  WH    9,997  0.9150   WRB(3288), WPS(61), WP(2324), WZH(242)
15  DO    4,436  0.4060   DON(380), DOG(301)
16  IF    4,218  0.3861
17  POS   2,244  0.2054
18  EX    1,246  0.1140
19  NEG   1,046  0.0957
20  FW    437     0.04
[...]
==========================================================
```

Figure 2-3 Sample output of POS tag frequency count by WordSmith (excerpt)

2.3 Relative proportions of the major POS categories in the BLC and the Reference Corpora

The above results have been consolidated into the following comparison table (Table 2-1), which shows relative proportions of major POS categories expressed in percentage to the total number of words/tags in each corpus (numbered 1 to 7). For the BLC, breakdown figures are also given for each of the three subcorpora for a within-group comparison purpose. At the end of each row are the mean ratio (Average of 1, 5 and (6+7)/2), the standard deviation (SD), the difference coefficient (Dif.C.), and the "Z" score of each POS category respectively.

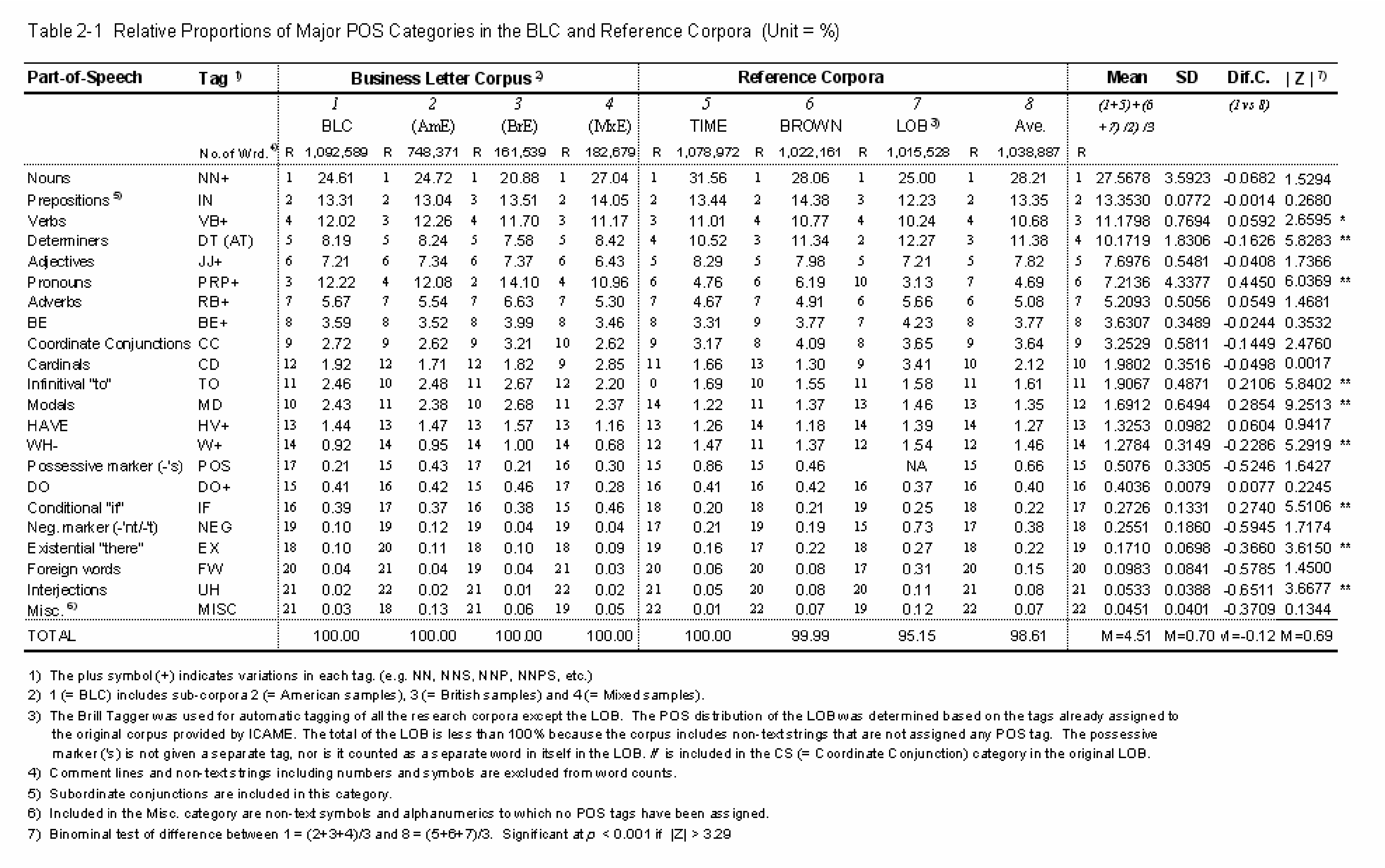


Table 2-1 Relative Proportions of Major POS Categories in the BLC and Reference Corpora (Unit = %)

Part-of-Speech	Tag [1]		Business Letter Corpus [2]								Reference Corpora								Mean	SD	Dif.C.	| Z | [7]	
			1		2		3		4		5		6		7		8		(1+5)+(6		(1 vs 8)		
			BLC		(AmE)		(BrE)		(MxE)		TIME		BROWN		LOB [3]		Ave.		+7) /2) /3				
	No.of Wrd. [4]	R	1,092,589	R	748,371	R	161,539	R	182,679	R	1,078,972	R	1,022,161	R	1,015,528	R	1,038,887	R					
Nouns	NN+	1	24.61	1	24.72	1	20.88	1	27.04	1	31.56	1	28.06	1	25.00	1	28.21	1	27.5678	3.5923	-0.0682	1.5294	
Prepositions [5]	IN	2	13.31	2	13.04	3	13.51	2	14.05	2	13.44	2	14.38	3	12.23	2	13.35	2	13.3530	0.0772	-0.0014	0.2680	
Verbs	VB+	4	12.02	3	12.26	4	11.70	3	11.17	3	11.01	4	10.77	4	10.24	4	10.68	3	11.1798	0.7694	0.0592	2.6595	*
Determiners	DT (AT)	5	8.19	5	8.24	5	7.58	5	8.42	4	10.52	3	11.34	2	12.27	3	11.38	4	10.1719	1.8306	-0.1626	5.8283	**
Adjectives	JJ+	6	7.21	6	7.34	6	7.37	6	6.43	5	8.29	5	7.98	5	7.21	5	7.82	5	7.6976	0.5481	-0.0408	1.7366	
Pronouns	PRP+	3	12.22	4	12.08	2	14.10	4	10.96	6	4.76	6	6.19	10	3.13	7	4.69	6	7.2136	4.3377	0.4450	6.0369	**
Adverbs	RB+	7	5.67	7	5.54	7	6.63	7	5.30	7	4.67	7	4.91	6	5.66	6	5.08	7	5.2093	0.5056	0.0549	1.4681	
BE	BE+	8	3.59	8	3.52	8	3.99	8	3.46	8	3.31	9	3.77	7	4.23	8	3.77	8	3.6307	0.3489	-0.0244	0.3532	
Coordinate Conjunctions	CC	9	2.72	9	2.62	9	3.21	10	2.62	9	3.17	8	4.09	8	3.65	9	3.64	9	3.2529	0.5811	-0.1449	2.4760	
Cardinals	CD	12	1.92	12	1.71	12	1.82	9	2.85	11	1.66	13	1.30	9	3.41	10	2.12	10	1.9802	0.3516	-0.0498	0.0017	
Infinitival "to"	TO	11	2.46	10	2.48	11	2.67	12	2.20	0	1.69	10	1.55	11	1.58	11	1.61	11	1.9067	0.4871	0.2106	5.8402	**
Modals	MD	10	2.43	11	2.38	10	2.68	11	2.37	14	1.22	11	1.37	13	1.46	13	1.35	12	1.6912	0.6494	0.2854	9.2513	**
HAVE	HV+	13	1.44	13	1.47	13	1.57	13	1.16	13	1.26	14	1.18	14	1.39	14	1.27	13	1.3253	0.0982	0.0604	0.9417	
WH-	W+	14	0.92	14	0.95	14	1.00	14	0.68	12	1.47	11	1.37	12	1.54	12	1.46	14	1.2784	0.3149	-0.2286	5.2919	**
Possessive marker (-'s)	POS	17	0.21	15	0.43	17	0.21	16	0.30	15	0.86	15	0.46		NA	15	0.66	15	0.5076	0.3305	-0.5246	1.6427	
DO	DO+	15	0.41	16	0.42	15	0.46	17	0.28	16	0.41	16	0.42	16	0.37	16	0.40	16	0.4036	0.0079	0.0077	0.2245	
Conditional "if"	IF	16	0.39	17	0.37	16	0.38	15	0.46	18	0.20	18	0.21	19	0.25	18	0.22	17	0.2726	0.1331	0.2740	5.5106	**
Neg. marker (-'nt/-'t)	NEG	19	0.10	19	0.12	19	0.04	19	0.04	17	0.21	19	0.19	15	0.73	17	0.38	18	0.2551	0.1860	-0.5945	1.7174	
Existential "there"	EX	18	0.10	20	0.11	18	0.10	18	0.09	19	0.16	17	0.22	18	0.27	18	0.22	19	0.1710	0.0698	-0.3660	3.6150	**
Foreign words	FW	20	0.04	21	0.04	19	0.04	21	0.03	20	0.06	20	0.08	17	0.31	20	0.15	20	0.0983	0.0841	-0.5785	1.4500	
Interjections	UH	21	0.02	22	0.02	21	0.01	22	0.02	21	0.05	20	0.08	20	0.11	21	0.08	21	0.0533	0.0388	-0.6511	3.6677	**
Misc. [6]	MISC	21	0.03	18	0.13	21	0.06	19	0.05	22	0.01	22	0.07	19	0.12	22	0.07	22	0.0451	0.0401	-0.3709	0.1344	
TOTAL			100.00		100.00		100.00		100.00		100.00		99.99		95.15		98.61		M=4.51	M=0.70	M=-0.12	M=0.69	

1) The plus symbol (+) indicates variations in each tag. (e.g. NN, NNS, NNP, NNPS, etc.)

2) 1 (= BLC) includes sub-corpora 2 (= American samples), 3 (= British samples) and 4 (= Mixed samples).

3) The Brill Tagger was used for automatic tagging of all the research corpora except the LOB. The POS distribution of the LOB was determined based on the tags already assigned to the original corpus provided by ICAME. The total of the LOB is less than 100% because the corpus includes non-text strings that are not assigned any POS tag. The possessive marker ('s) is not given a separate tag, nor is it counted as a separate word in itself in the LOB. *If* is included in the CS (= Coordinate Conjunction) category in the original LOB.

4) Comment lines and non-text strings including numbers and symbols are excluded from word counts.

5) Subordinate conjunctions are included in this category.

6) Included in the Misc. category are non-text symbols and alphanumerics to which no POS tags have been assigned.

7) Binominal test of difference between 1 = (2+3+4)/3 and 8 = (5+6+7)/3. Significant at p < 0.001 if |Z| > 3.29

The difference coefficient[7] is calculated for each POS category by the following formula:

$$(\alpha_{FR} - \beta_{FR}) \div (\alpha_{FR} + \beta_{FR})$$

where α_{FR} and β_{FR} stand for the frequency ratios of the two groups, α and β, that are being compared — in this case, the BLC and the Reference Corpora. The coefficient takes a value between −1.0 and +1.0. The positive value means that the distribution of a particular word class is proportionately higher in the BLC, while the negative value means that it is higher in the Reference Corpora. The Z score given for each POS category is the result of the standard two-tailed Binominal Test of difference performed for respective POS categories between the BLC and the Reference Corpora. A Z score greater than 3.29 indicates that the observed difference is statistically significant at $p < 0.001$.[8]

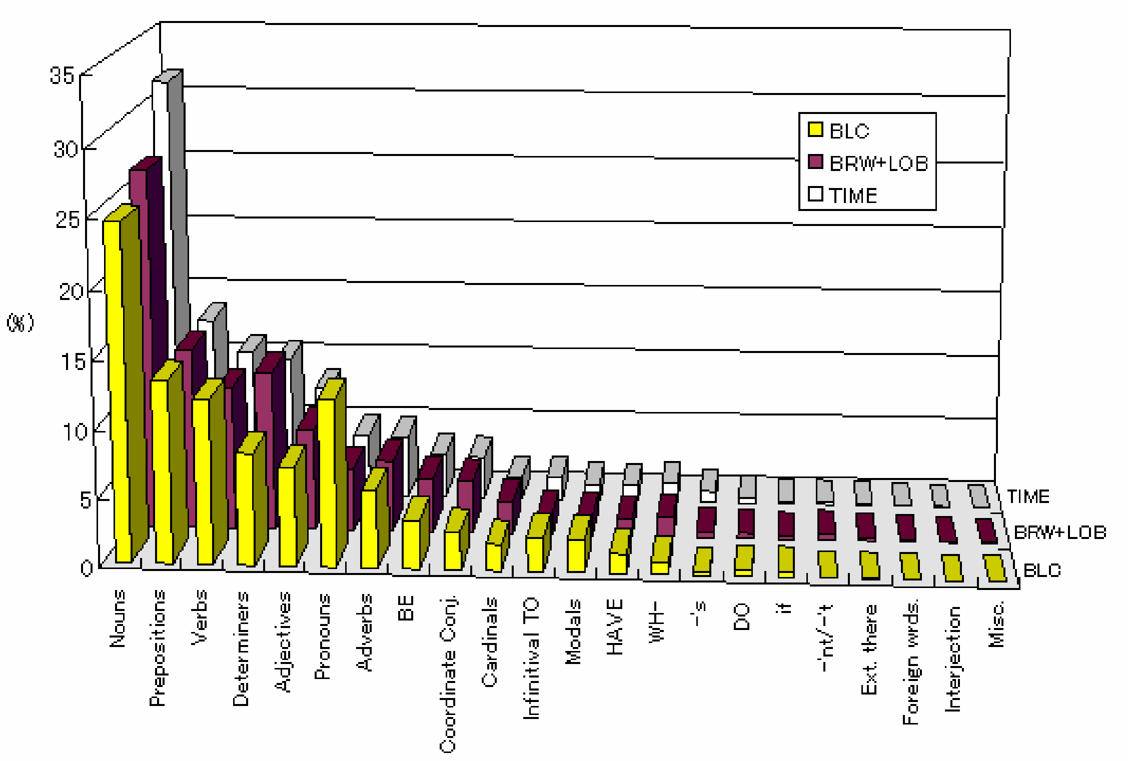
Table 2-2 on the next page gives a triad comparison of major POS categories (sorted in descending order of the mean values) among the BLC, BRW-LOB and TIME Corpora, and Figure 2-4 is a graphic representation of it. The main rationale for producing this comparison table in addition to Table 2-1 is that it gives a better and quick overview as to where the BLC stands in relation to the two other reference corpora: the consolidated Brown and LOB Corpus (BRW-LOB) which represents "general English" and the TIME Corpus representing quite a different type of "special purpose English."

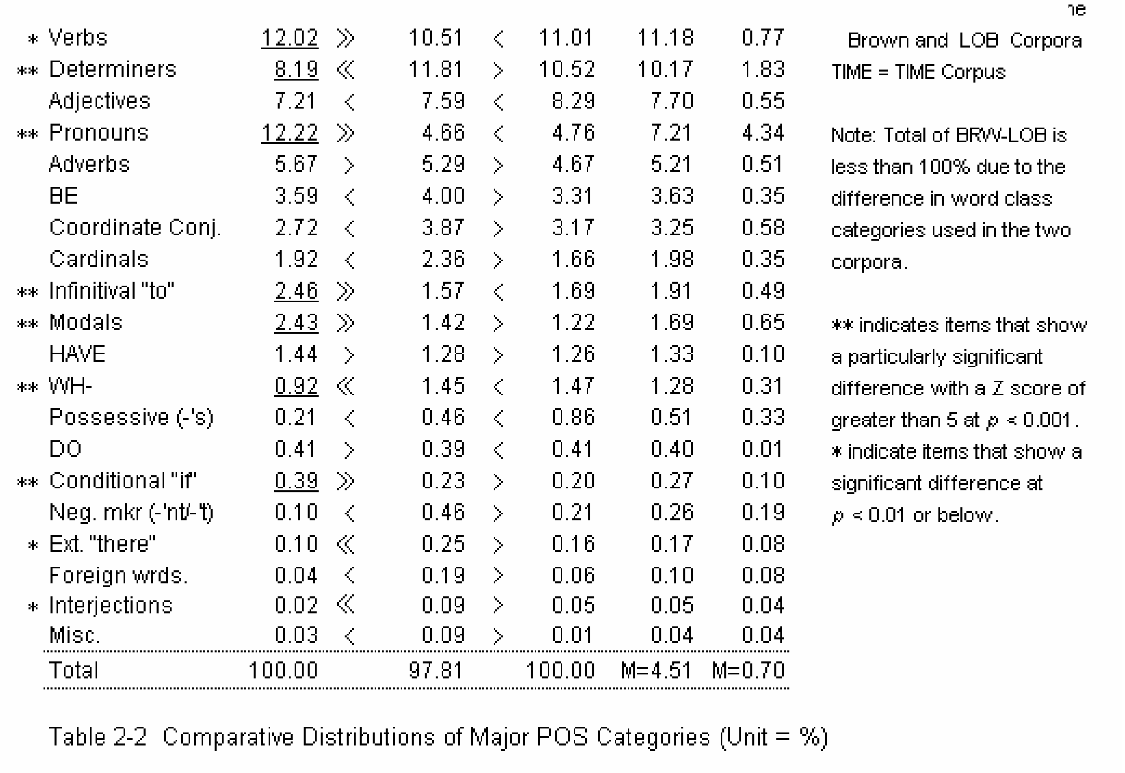
For instance, Table 2-2 shows at a glance that the proportion of NOUNS is the largest in the TIME Corpus (31.56%), followed by the BRW-LOB Corpus (26.53%) and then by the BLC (24.61%), with Mean = 27.57 and S.D.= 3.59. Likewise, we find that the proportion of PRONOUNS is the largest in the BLC (12.22%), followed by the TIME Corpus (4.76%) and then by the BRW-LOB Corpus with a margin of 0.1 percentage point. The "<" symbol in the table indicates a "larger or smaller than" relationship between the adjacent pair, and the "≪" symbol points to the value that is either the largest or the smallest among the three by a relatively large margin. The double-asterisk (**) indicates POS categories that show a particularly signifi-cant difference in the distributional proportions among the three corpora, with a Z score of greater than 5 at $p < 0.001$. The single-asterisk (*) marks the categories that show a significant difference at $p < 0.01$ or below.

2.4 Discussion

As we have seen in Table 2-1, the relative proportions of major POS categories is fairly consistent in both the BLC and the Reference Corpora. In terms of rank order (shown in the "R" columns of the table), NOUNS (NN+) are ranked first in all

Figure 2-4 Comparative Distributions of Major POS Categories





*	Verbs	12.02	»	10.51	<	11.01	11.18	0.77
**	Determiners	8.19	«	11.81	>	10.52	10.17	1.83
	Adjectives	7.21	<	7.59	<	8.29	7.70	0.55
**	Pronouns	12.22	»	4.66	<	4.76	7.21	4.34
	Adverbs	5.67	>	5.29	>	4.67	5.21	0.51
	BE	3.59	<	4.00	>	3.31	3.63	0.35
	Coordinate Conj.	2.72	<	3.87	>	3.17	3.25	0.58
	Cardinals	1.92	<	2.36	>	1.66	1.98	0.35
**	Infinitival "to"	2.46	»	1.57	<	1.69	1.91	0.49
**	Modals	2.43	»	1.42	>	1.22	1.69	0.65
	HAVE	1.44	>	1.28	>	1.26	1.33	0.10
**	WH-	0.92	«	1.45	<	1.47	1.28	0.31
	Possessive (-'s)	0.21	<	0.46	<	0.86	0.51	0.33
	DO	0.41	>	0.39	<	0.41	0.40	0.01
**	Conditional "if"	0.39	»	0.23	>	0.20	0.27	0.10
	Neg. mkr (-'nt/-'t)	0.10	<	0.46	>	0.21	0.26	0.19
*	Ext. "there"	0.10	«	0.25	>	0.16	0.17	0.08
	Foreign wrds.	0.04	<	0.19	>	0.06	0.10	0.08
*	Interjections	0.02	«	0.09	>	0.05	0.05	0.04
	Misc.	0.03	<	0.09	>	0.01	0.04	0.04
	Total	100.00		97.81		100.00	M=4.51	M=0.70

pus
ne
Brown and LOB Corpora
TIME = TIME Corpus

Note: Total of BRW-LOB is less than 100% due to the difference in word class categories used in the two corpora.

** indicates items that show a particularly significant difference with a Z score of greater than 5 at $p < 0.001$.
* indicate items that show a significant difference at $p < 0.01$ or below.

Table 2-2 Comparative Distributions of Major POS Categories (Unit = %)

the research corpora by a large margin, followed by PREPOSITIONS (IN), VERBS (VB+), DETERMINERS (DT), ADJECTIVES (JJ+), ADVERBS (RB+) and so on in roughly this order, with PRONOUNS (PRP+) being ranked somewhere in between these major categories—

generally higher in the rank in the BLC (and its subcorpora) and lower in the three Reference Corpora.[9]

More specifically, it was found that NOUNS comprise approximately 27.57% of all the word tokens in the research corpora[10] with the standard deviation of 3.59; PREPOSITIONS account for approximately 13.35% (S.D.=0.08); VERBS—excluding BE, HAVE and DO—account for 11.18% (S.D.=0.77); DETERMINERS, 10.17% (S.D.=1.83), ADJECTIVES, 7.7% (S.D.=0.55); ADVERBS, 5.21% (S.D.=0.51), PRONOUNS, 7.21% (S.D.= 4.34) and so on. When the BLC and the Reference Corpora are compared binominally for each of these POS categories, the most significant differences (Z > 5) are found in the following six categories:

MODALS (Dif.C.= 0.2854, Z=9.2513)
CONDITIONAL *if* (Dif.C.= 0.274, Z=5.5106)
PRONOUNS (Dif.C.= 0.445, Z=6.0369)
INFINITIVAL *to* (Dif.C.= 0.2102, Z=5.8402)
DETERMINERS (Dif.C.= −0.1626, Z=5.8283)
WH-PARTICLES (Dif.C.= −0.2286, Z=5.2919)

Also found significant, albeit to a lesser degree, are the differences in VERBS (Dif.C.=0.0592, Z=2.6595), EXISTENTIAL *there* (Dif.C.=−0.366, Z=3.615), and INTER-JECTIONS (Dif.C.=−0.6511, Z=3.6677). In sum, the current POS analysis revealed that Business English, in comparison with English used in other genres, is characterized by a significantly heavier use of modals, *if,* pronouns, infinitival *to* and verbs on one hand, and by a comparatively infrequent use of determiners, WH-particles, existential *there* and interjections on the other, as summarized in Table 2-2.

Having thus identified the grammatical categories that exhibit significant differences between the BLC and the Reference Corpora, we shall now briefly consider possible reasons of these differences and their grammatical and syntactic implications.

1) Modals

As can be seen in Tables 2-1 and 2-2, the overall average proportion of MODALS to the total word tokens is about 1.69% (S.D.=0.65). Their proportion in the BLC, however, is much larger than in the Reference Corpora. In fact, the difference coefficient of 0.2854 means that MODALS are, on average, about 1.77 times more frequent in the BLC than in the Reference Corpora (As we will see later, modals with positive "Keyness" scores, *i.e. will, would, shall, should, can,* and *may,* are 2.3 times more frequent in the BLC). This higher proportion of MODALS in the former may be accounted for by the fact that business discourse is almost always concerned with future events—both possible and imaginative. It wouldn't make much sense in business to talk about the past and the status quo unless the discourse is aimed at some particular goals to be realized in the not-too-distant future. In English, of course, future events are most likely be expressed with a modal sentence containing one or two of the possibility modals (*can, could may, might*), predictive modals (*will, would, shall*), or necessity modals (*ought, should, must*). The politeness feature of modal auxiliaries is also an important factor that contributes to the high frequency of occurrence of modals in business discourse. In any event, modals are very important lexical items in Business English and they warrant a more detailed analysis and discussion, which I shall make in Chapter 4 of this paper.

2) Conditional *if*

The high frequency of occurrence of the conditional marker *if* may be taken as an empirical evidence in support of the above speculation as to the importance of modals in Business English. *If* accounts for 0.39% of all word tokens in the BLC, whereas it accounts for 0.22% of all word tokens in the Reference Corpora, with the difference coefficient of 0.274 in favor of the former. In more practical terms, this means that *if* occurs about 1.77 times more frequently in the BLC than in the Reference Corpora. In other words, it is used at least once every 256-word chunk of text in the BLC. Since the mean sentence length of the BLC is 16.68 words[11], this means that at least one out of every 15 sentences in the BLC contains this particular lexical item as its major constituent.[12] In short, it is almost impossible to write business messages without saying “if” somewhere in the discourse.

From a pedagogical viewpoint, the importance of *if* rests on its direct relationship with the conditional construction, which is one of the key syntactic variations of Business English. It should also be mentioned that a close analysis of the Learner BLC indicates that the Japanese business people are having much difficulty in writing well-formed conditional sentences to express future events. This tendency is not peculiar to the business people, but is common to Japanese learners of English in general as attested by many college and high school teachers.[13]

More detailed discussions as to the syntactic and semantic environments in which *if* occurs in the BLC, as well as the types of errors Japanese writers of Business English make with regard to *if*-sentences will be made in Chapter 4. For the moment, it is sufficient to confirm the importance of *if*—from both quantitative and qualitative perspectives—as a major grammatical and syntactic device with which to express conditional meanings.

3) Pronouns

As we have already seen, the difference in the distributional proportions of PRONOUNS between the BLC and the Reference Corpora is highly significant with Z= 6.0369. The difference coefficient of 0.445 obtained for this category indicates that, on average, pronouns are about 2.6 times more frequent in the BLC than in the Reference Corpora. The current POS analysis, however, does not show which pronouns we are talking about, nor does it reveal the exact distributional pattern of individual pronouns. Again, we have to postpone our discussion on these matters until we will have obtained in Chapter 3 more detailed information on individual lexical items including pronouns (discussion on personal pronouns will be made in Chapter 4).

At this stage, we can only hypothesize that it may well be the personal pronouns, particularly the 1st- and 2nd-person pronouns, that are responsible for the much higher proportion of the category PRONOUNS in the BLC, and that, if such is the case, this finding may be interpreted as a good lexical evidence to support the view that a business message is, and should be, basically a "conversation on the paper"—meaning that "*I/we (me/us)*" talking to "*you*" and vice versa, rather than an anonymous writer talking to nobody in particular.

Further, if the "conversational" style is what characterizes written business discourse, we can also expect that we will have a much higher proportion of contractions—the combination of a subject NP and one of the reduced forms of BE, HAVE, WILL/SHALL, and the negative marker NOT—in the BLC than in the Reference Corpora. We will consider these issues in Chapter 4 in more detail in relation to the "write-as-you-talk" hypothesis set out at the outset of this paper.

4) Infinitival *to*

The infinitival *to* occurs 26,929 times in the BLC (approximately 2.46% of the total word tokens), while it occurs an average of 16,711 times in the three Reference Corpora (1.61%, S.D.=0.07)[14], with the difference coefficient of 0.2102 in favor of the former. This difference in the observed frequencies between the two corpora is highly significant, and the reason why the infinitival *to* is more frequent in the BLC needs to be properly accounted for.

One possible explanation is that simplicity is highly valued in business writing and, other things being equal, the writer tends to choose *to*-infinitive complemen-tation over syntactically more complex *tha*t-clause complementation, as long as the main verb of the sentence permits both of the constructions. This assumption predicts that sentences like (5a) and (6a) below are more likely in business messages than (5b) and (6b):[15]

(5a) I *hope to* hear from you soon. [BZ01:00527]
(5b) We *hope that* we may hear from you again soon. [BZ29:00245]
(6a) We have *decided to* review our entire marketing effort. [BZ12:00023]
(6b) We have *decided that* we must terminate this program. [BZ32:01529]

The assumption, however, is still very much intuitive, and other factors are likely be involved in the process of choosing one form over the other. Also, it may be that the choice is not entirely free but is dependant upon which verb to use. A quick look over the BLC data indicates, for instance, that the verb *hope* is strongly associated with *that*-complementation and that the ZERO-*that* form is three times more likely than the full form (Table 2-3). The verb *decide*, on the other hand, tends to co-occur with *to*-infinitive than with *that*-clause as its complement (Table 2-4).

Syntactic Pattern	Freq.	%
HOPE + that COMP	409	20.17
HOPE + ZERO-that COMP	1,300	64.10
HOPE + to COMP	319	15.73
TOTAL	2,028	100.00

Table 2-3 Major syntactic patterns of HOPE and their frequencies in the BLC

Syntactic Pattern	Freq.	%
DECIDE + that COMP	30	7.11
DECIDE + ZERO-that COMP	11	2.61
DECIDE + to COMP	250	59.24
DECIDE + NP (incl. passive)	50	11.85
DECIDE + on/upon NP	31	7.35
DECIDE + against NP	2	0.47
DECIDE (no surface object NP)	7	1.66
others	41	9.72
TOTAL	422	100.00

Table 2-4 Major syntactic patterns of DECIDE and their frequencies in the BLC

These data point to the need of looking into the actual usage of individual verbs in more detail before making any generalization as to whatever the factors that may be at work in the mind of the writer in making the syntactic choice between the two constructions.

5) Determiners

There is no doubt that DETERMINERS, particularly the definite and indefinite articles, are among the most frequent lexical items in English. In his survey of the six major English corpora (*i.e.,* Birmingham Corpus, Brown Corpus, LOB Corpus, Wellington Corpus, American Heritage Corpus, and London-Lund Corpus), Kennedy (1998, p. 101) reports that the definite article *the* is ranked 1st in all the six corpora and *a* is ranked 5th in the first four corpora, 4th in the AHC, and 6th in the LLC. The tendency is quite the same with other more genre-specific corpora, such as those of economics or academic texts, as reported in Sutarsyah *et al*. (1994, p. 41).

In the BLC, DETERMINERS were found to comprise approximately 8.19% of all the word tokens and are ranked 5th in the frequency order following NOUNS, PREPOSITIONS, PRONOUNS, and VERBS. Although their frequency as a category is still very high, the current POS analysis show that DETERMINERS occur about 1.38 times *less* frequently in the BLC than in the Reference corpora, with the difference coefficient of −0.1626 (Z=5.4946).

As to the definite and indefinite articles, they are the 1st and 7th most frequent words in the BLC, but what makes the matter look quite different is that they get extremely high *negative* "Keyness" scores[16] in the BLC, indicating that these two items occur a lot less (approximately 37% less for *the* and 29% less for *a*) in business writing than in other genres, as we shall see in more detail in Section 2.2.

This rather low profile of the definite and indefinite articles in the BLC can only be accounted for by the assumption that business messages are basically action-oriented, being more concerned with human participants and their actions in a particular event—namely, with "Who does/did/is to do what, in what way and why, etc."—than the description of abstract states, objects, or generalized processes.

This assumption predicts that the verbal style is more preferred in business writing than the nominal style, which will naturally results in a greater frequency of lexical verbs on one hand, and a lesser frequency of nouns on the other, than are normally expected. As we have already seen in Table 2-1, the proportion of VERBS was found to be greater in the BLC than in the combined Reference Corpora (*i.e.* 12.02% vs 10.68%, excluding BE, HAVE, and DO), whereas the proportion of NOUNS is much less in the former than in the latter (*i.e.* 24.61% vs 28.81%), partly confirming the above assumption. To be more precise, however, we may need to check up on the number of nominalizations rather than nouns in general, since it is the use of nominalization that requires additional articles and certain propositions. The following quotes from the Brown and LOB Corpora will make the point clear:[17]

(7) The first is a negative warning: there is no point **in the creation of** faculty committees and advisory boards with high-sounding titles but no real authority. [BRW:03159]

(8) It does, however, give positives when certain other antibodies are present so that care must be taken **in the establishment of** the specificity of any antibody detected by this method. [LOB:29825]

In the above sentences, the nouns *creation* and *establishment* are derived from the verbs *create* and *establish* respectively and, when used in the nominalized form, they can only appear in the syntactic environment of "P Det N PP" (*e.g.* in the creation of NP; in the establishment of NP), whereas their verbal counterparts require only one infinitival *to* (*e.g.* to create NP; to establish NP) to express the same intended propositional meanings. Thus, it seems clear that there is a close relationship between the amount of nominalizations and the frequencies of articles and certain propositions (*of*, in particular).[18]

In Chapter 4 (Section 4.3), we will see in more detail whether nominalizations are in fact less common in business correspondence than in the language of other genres by checking up on the numbers of typical noun-forming suffixes such as *-tion/-sion, -ment, -ness,* and *-ity* in the BLC and comparing them with those in the three Reference Corpora. Should this be confirmed as expected, then we would have at least explained one of the major reasons why the definite and indefinite articles occur a lot less in business writing.

6) WH-particles

The WH words were found to occur significantly fewer in the BLC than in the Reference Corpora, with the difference coefficient of −0.2286. One possible reasons for this may be that *WH-relativization* tends to be avoided in business corres-pondence when, of course, there is a choice not to use such a construction in a given context. The resultant effect of this is shorter and syntactically less complex sentences—which is in perfect accord with the communicative purpose of business correspondence.

In order to prove this assumption, however, we need to find out (1) whether the sentences are indeed shorter and less complex in the BLC than in the Reference Corpora, and (2) the exact distribution and frequencies of such major WH-relativisors as *which, where, when, who* and *that* in the BLC in comparison with the Reference Corpora. As to the first question, we know that the mean sentence length (MSL) of the BLC is about 16.68 with the average number of predications per sentence (MNP) of 3.08, whereas the MSL and MNP of the three Reference Corpora are about 20.35 and 3.56 respectively.[19] These data indicate that the sentences are in fact much shorter and syntactically less complex in the BLC than in the Reference Corpora. The second question I will discuss in Section 4.1 of Chapter 4 in more detail, focusing on *which* that is by far the most frequently used relative pronoun in English.

7) Existential *there*

The use of *there* as a dummy (or preparatory) subject of a sentence is very common in English. Thus, it appears 2,280 and 2,789 times in the Brown and LOB Corpora respectively, comprising roughly 0.22% of the total word tokens in the former and 0.27% in the latter. It is also used in the BLC very frequently, but the total frequency of existential-*there* (hereinafter, EX-*there*) in the BLC is only 1,115 (see Rank 86 of the COMPREHENSIVE BLC WORDLIST (KAppendix D1)), which is less than a half of those in the Brown and LOB Corpora, and about 34% less than that in the TIME Corpus. The Z score obtained for this grammatical category is 3.615, meaning that the observed difference between the BLC and the three Reference Corpora is statistically significant at p < 0.001.

The reason why EX-*there* is less common in business messages is not as clear as it may seem. However, if business messages are "more concerned with human participants and their actions," than with the description of "things out there," as I have mentioned in the previous discussion, then it seems natural for the business writer to choose an animate subject as a starting point of his message when he refers to the existence of something, someone or a particular state of being, rather than presenting it in an uninvolved EX-*there* construction. Thus, (9a) is more likely than (9b) in business correspondence.

(9a) We have a sales representative in Japan.
(9b) ?There is a sales representative in Japan.

Having said this, however, I must again refer back to the fact that the EX-there construction is nevertheless a very important syntactic variation available to the business writer, comprising roughly one out of every 57 sentences in the BLC. The EX-there construction, therefore, deserves a special attention and its communicative function and patterns of usage in written business discourse need to be explored in some detail.

a) Communicative function of the EX-*there* construction

In English, the unmarked position of new information is toward the end of a sentence (Takami and Kamio, 1998, p. 121). A sentence like (10a) below sounds a bit awkward because in this example the new information "A dog" is placed at the sentence-initial position which is usually reserved for the given information, or *Theme*, of a sentence. To avoid this awkwardness, we move the subject NP that carries new information to its unmarked clause-end position as in (10b) and, instead, insert expletive *there* to the now empty subject position to get (10c):

(10a) [A dog] is in the garden.
(10b) [ϕ] is [a dog] in the garden.
(10c) [There] is [a dog] in the garden.

From a functional viewpoint, the main communicative function of the sentence-initial "there" as a dummy subject is to indicate that what follows is a new piece of information, thereby properly directing the attention of the reader/hearer to what it should be focused on. Therefore, this construction is best understood as a type of a focus sentence much like the "*It is . . . that*" construction, or the cleft sentence. From this function follows the selectional constraint of the postponed psychological subject[20] of an EX-*there* sentence that it must be an indefinite NP. In other words, it cannot be a pronoun or accompanied by the definite article. Sentences like (11) and (12) below, therefore, are considered ill-formed:[21]

(11) *There is the book on the table. [*sic*]
(12) *There is it on the table. [*sic*]

b) Major syntactic patterns of the EX-*there* construction in the BLC

According to Murata *et al.* (1996, p. 42), the EX-*there* construction is used in "a sentence whose main verb denotes either the existence or appearance (of something, someone or a particular state of being)." Thus, the most typical verb used in this construction is a form of the verb BE (*e.g.* is, are, was, were, be, been, being), including a combination of "HAVE + BE," "MODAL (+ HAVE) + BE," and "SEEM/APPEAR TO + BE." When a full verb is used in this construction, the verb by definition must permit replacement with the abstract verb EXIST (or COME TO EXIST) to be semantically acceptable, as in the following examples:

(15) He gave them half-an-hours talk on the Synagogue [...], and **there followed** a period for questions and answers. [LOB:25192]
(16) **There comes** a time in the lives of most of us when we want to be alone. [BRW:12979]

(17) With 4 months left in our fiscal year, **there remain** significant staff-reduction quotas to meet. [BZ10:02962]

In the BLC, we have identified 1,115 instances of the EX-*there* construction. Approximately 83.05% (N=926) of them occur in the "there is/are/was/were NP" or "there has/have been NP" frames, and about 30.56% (N=283) of which are headed by the conditional *if* (including four instances of *whether*). Among the combinations with modals (N=149) , by far the most frequent pattern is "there will/would be NP" which comprises about 76.51% (N=114) of all the instances of the "there + MODAL BE + NP" frames in the BLC, as shown in Table 2-5.

Syntactic Patterns	Present	Past	Present Perfect	Past Perfect	Interrogative	Subjunctive	Participle	TOTAL
THERE is NP	419	26	50	-	27	-	-	522
THERE are NP	74	23	21	-	3	-	-	121
if THERE is NP	235	7	1	-	-	-	-	243
if THERE are NP	34	1	1	-	-	-	-	36
whether THERE is NP	3	-	-	-	-	-	-	3
whether THERE are NP	1	-	-	-	-	-	-	1
THERE can be NP	1	1	-	-	-	-	-	2
THERE may be NP	11	3	2	1	-	-	-	17
THERE must be NP	3	-	2	-	-	-	-	5
THERE should be NP	10	-	1	-	-	-	-	11
THERE will be NP	99	15	-	-	-	-	-	114
THERE seems to be NP	4	-	-	-	-	-	-	4
THERE seem to be NP	1	-	-	-	-	-	-	1
THERE appears to be NP	3	-	1	-	-	-	-	4
THERE continues to be NP	1	-	-	-	-	-	-	1
THERE comes NP	1	-	-	-	-	-	--	1
THERE remain NP	1	-	-	-	-	-	-	1
should THERE be NP	-	-	-	-	-	24	-	24
let THERE be NP	-	-	-	-	-	2	-	2
THERE being NP	-	-	-	-	-		2	2
TOTAL	901	76	79	1	30	26	2	1115

Table 2-5 Major syntactic patterns of Ex-*there* construction in the BLC and their frequencies

a (216), no (186), any (140), anything (106), some (81), other (67), many (30), an (26), little (21), more (21), nothing (20), further (17), something (14), another (14), much (13), certain (13), several (12), specific (8), additional (8), every (7)

Table 2-6 Most frequent determiners and qualifiers co-occurring with *there* within three words to the right (N>5)

Table 2-6 shows the major types of indefinite determiners and qualifiers that co-occurs with *there* within three words to the right (*e.g.* "There BE **Det** N" or "There BE (Det) **QFY** N"). The figures in the brackets indicate their respective frequencies. Determiners and qualifiers that occur outside the three-words-to-the-right search scope are not included (e.g. "there might have been **Det/QFY** N"). This table indicates that the indefinite NP in the EX-*there* construction can accompany a variety of determiners and qualifiers, but they may be lumped together into the following six patterns:

there is/are
- a, an [N]
- any, some, no, every, certain [N]
- anything, something, nothing [COMP S]
- (any/some) other, another [N]
- many, little, more, much, several [N]
- further, additional, (any) specific [N]

Some of the prototypical examples of the EX-*there* construction containing one of these determiners and/or qualifiers are as follows:

(18) We are convinced that **there is** a considerable market here for your products. [BZ23:03859]

(19) **There are**, however, other projects we are now pursuing that would work to increase AAA's exposure in the Japanese market. [BZ19:02051]

(20) **If there are** any problems, please contact us immediately. [BZ22:01258]

(21) However, **there will be** some changes in terms of the handling of claims, requests for information and the processing of paperwork. [BZ12:02001]

(22) To date, **there have been** no accidents or injuries reported by our tenants, but we want to notify you formally of your responsibility in this situation. [BZ09:00947]

In all of these instances, the NP that follows the dummy subject *there* carry new information to which the attention of the reader is being drawn. When read aloud or spoken, therefore, the underlined NP is naturally given tonic prominence in each of these sentences.

In addition to the above canonical patterns, there are 12 instances of “There VB (to be) NP” in the BLC. The range of the verbs used in this construction is very limited, with only the following five verbs:

(23) There **seem to be** several areas in which we can work together. [BZ19:18606]

(24) There **appears to be** some confusion over our vacation policy. [BZ01:01278]

(25) There **continues to be** no charge for pick-up or redelivery. [BZ10:01566]

(26) With each new promotion, there always **comes** a new challenge, along with the responsibility. [BZ09:03997]

(17) With 4 months left in our fiscal year, there **remain** significant staff-reduction quotas to meet. [BZ10:02962]

All of these verbs, or combinations of “VB + to be,” permit replacement with the abstract verb EXIST (or COME TO EXIST) as predicted earlier. These sentences are, therefore, well-formed both syntactically and semantically.

c) The “should THERE be NP” construction

Table 2-5 also shows that there are 24 instances of the “should there be NP” construction. This is a formal variation of “if there is/are NP” (or, to be more precise, a formal inversion-structure of “if there should be NP”[22]) and is usually considered a British variety. A closer look at the corpus, however, indicates that this construction occurs two times in the British samples (BrE), 21 times in the American samples (AmE), and only one in the Mixed samples (MxE). The following are some of the representative samples:

(27) **Should there be** any matter in which we may have given you cause to be dissatisfied, we hope you will give us the opportunity to put it right so that our custom can be renewed. [BZ23:01627] (BrE)

(28) **Should there be** any reason why this invoice can not be processed for payment, please contact me immediately at 456-9986. [BZ10:02030] (AmE)

(29) Please do not hesitate to ask **should there be** any way in which I can be of further assistance. [BZ19:24142] (AmE)

(30) I wish you every success with your program and invite you to call on me **should there be** some other way in which I could be of assistance.
[BZ19:25271] (AmE)

Of the 21 instances in the American samples, 17 are from the same source (BZ19) by the same author and, therefore, the number can be boiled down to about five. Also, a similar, but more usual "should you have NP" pattern occurs a total of 52 times, of which 37 occur in the American samples, six in the British samples, and nine in the Mixed samples. (The number in the American samples may be reduced to 32 for the same reason stated above.)

This means three things: that this marked construction is rather rare in business correspondence, and that its use is motivated by the notion of formality (if not hyper-formality) as is also the case with the "should you have NP" version of this construction.[23] Also, the data suggest that it may be more American than British, although the cultural attribute seems less of a deciding factor than the stylistic intention of the individual writer regardless of his or her linguistic backgrounds.

8) Interjections

The category INTERJECTIONS includes such lexical items as *Oh, Well, Sorry, Alas, Gee, Hello, Howdy, Yeah, Yes, No* and so on (usually at the sentence-initial position). These very casual expressions are characteristic of informal spoken discourse and, therefore, understandably less common in normal written business discourse. The following comparison table shows that there are only 183 instances of INTERJECTIONS in the one-million-word BLC.

Corpus	Freq.	% to Total Word Tokens
BLC	183	0.0167
BROWN	815	0.0797
LOB	1,099	0.1082
TIME	531	0.0492

Table 2-7 Frequencies of INTERJECTIONS in the BLC and the Reference Corpora

The comparatively larger numbers in the Brown and LOB Corpora can be accounted for by the fact that both of them include many texts taken from novels and short stories as mentioned earlier in this section. The TIME Corpus contains many direct quotations from interviews or witnesses' accounts of news events; therefore, it has much larger proportion of INTERJECTIONS than the BLC. Examples of the sentences that include INTERJECTIONS are as follows:

(32) "**Yeah**", he said. [BRW:39695]

(33) **Alas**, no such thing happened. [BRW:06567]

(34) **Well**, I thought I'd make him an offer that would tempt him. [LOB:39753]

(35) "**Hello**, Willie, where did you come from?" the judge said in mild surprise. [LOB:44463]

(36) Clinton: **Oh**, **ye**s, we disagree. [TIME:01142]

(37) "**Gee**, I agree with you, Sam." [TIME:01168]

In the BLC, most of the 183 instances of INTERJECTIONS are found in "direct mail" (*e.g.* sales promotion letters to prospective customers), and interoffice memorandums and e-mail. They are seldom used in company-to-company business correspondence. The first two of the following sample sentences are taken from interoffice e-mail sent to a very close colleague of the writer. The other two samples are from sales promotion letters:

(38) **Sorry** I'm late in replying to your e-mail of Monday, April 1. [BZ30:00069]

(39) **Bye** for now. [BZ30:00074]

(40) Do keep in touch, and whenever you are in town I hope you drop by to say "**hello**". [BZ01:00473]

(41) **Well**, the new A-10 Timescan will solve the mystery of the disappearing telephone number. [BZ33:00215]

In each of these instances, the writer is writing the message as if he/she was "talking" to the reader, and this particular style of language is in itself an overt message to the reader that the writer thinks he is (or he wants to be) a close friend of the reader. The somewhat inflated use of language in (41) has of course been chosen intentionally on the part of the writer to make otherwise uninteresting and usually "pushy" sales talk sound a little entertaining.

In sum, the main function of INTERJECTIONS in written discourse is to give a "spoken" flavor to a piece of writing, thereby making it somewhat "informal" which, in turn, serves as a sign and assurance of closeness between already close friends. In the business context, INTERJECTIONS are therefore mainly used in interoffice correspondence, and their use is less common in "formal" business letters – unless the writer intentionally adopts a negative politeness strategy (Green, 1990. pp. 192-201) to get closer to the target customer or client.

2.5 Comparison between the Native and Learner BLCs

In this section, we compare the POS distribution of the BLC with that of the Learner Corpus (Learner BLC) to see if there are any significant differences between the two corpora.[24]

1) Major areas of difference and their implications

Figure 2-5 and Table 2-8 on the next page show that the overall patterns of POS distribution of two corpora are very similar, indicating that native and non-native English are after all not that different as to the proportions of nouns, verbs, prepositions and other parts of speech in the surface linguistic structure. A closer look at the data, however, points to some noticeable, if not statistically significant, differences in NOUNS, PRONOUNS, VERBS, DETERMINERS, and ADJECTIVES. In what follows, I will briefly discuss some of these differences focusing on nouns, determiners, and adjectives.

a) Nouns and Determiners

The largest difference is observed in NOUNS, with a relatively higher proportion of this category in the Learner BLC (27.51% vs 24.61%). This can be accounted for by the assumption that Japanese learners of Business English tend to use NPs in cases where native speakers of English would have used VPs. (I would call this assumption the "nominalization hypothesis"[25]). The lower proportion of VERBS in the Learner BLC (11.16% vs 12.02%) seems to support this observation. The assumption also predicts that we will have somewhat higher proportions of DETERMINERS and PREPOSITIONS in the Learner BLC, for the obvious reason that they are the necessary constituents of NPs. The data, however, indicate otherwise.

Table 2-8 indicates that the proportions of these two POS categories are actually less in the Learner Corpus than in the Native BLC, although the observed difference in PROPOSITIONS is almost negligible. This can only be explained in terms of "error effect." In general, the Japanese learners of English tend to omit the necessary articles (*the* or *a/an*) in NPs or in VPs (particularly in the light-verb construction in which verb-derived nouns are used in combination with such operant verbs as *make, get, take* and *have*). The following are some of such examples found in the Learner BLC:[26]

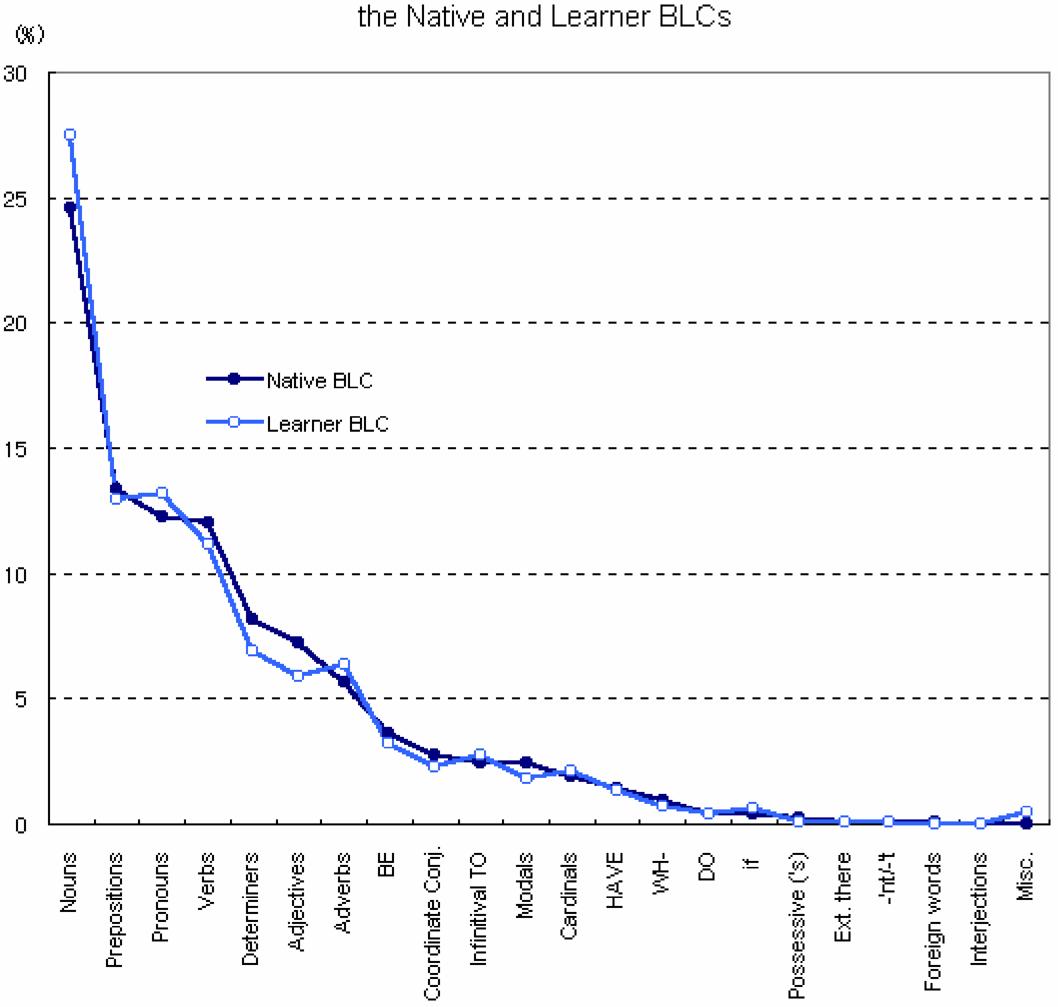
(42) *I requested to all the managers of each department **for submission of** budget estimate for first quarter of 1998 by the letter dated November 1. [WM 98:01799]

(43) *We are planning **issuance of** a bilingual monthly magazine for company employees. [WM98:25279]

(44) *I got in touch with Mr. Kawai and he has **made agreement** about this. [WM 98 :01532]

(45) *In view of the previous paragraph, I **make suggestion** as follow: [WM98: 27876]

Figure 2-5 Comparative POS distributions of the Native and Learner BLCs



Native BLC Tokens = 1,092,589
Learner BLC Tokens = 210,404

(Unit = %)

POS Category	Native BLC	Learner BLC	POS Category	Native BLC	Learner BLC
Nouns	24.61	27.51	Cardinals	1.92	2.16
Prepositions	13.31	12.98	HAVE	1.44	1.35
Pronouns	12.22	13.20	WH-	0.92	0.68
Verbs	12.02	11.16	DO	0.41	0.41
Determiners	8.19	6.94	Conditional "if"	0.39	0.65
Adjectives	7.21	5.38	Possessive ('s)	0.21	0.08
Adverbs	5.67	6.10	Ext. "there"	0.10	0.07
BE	3.59	3.21	Neg. mkr ('nt/'t)	0.10	0.08
Coordinate Conj.	2.72	2.26	Foreign wrds	0.04	0.01
Infinitival "to"	2.46	2.77	Interjections	0.02	0.00
Modals	2.43	1.84	Misc.	0.03	*1.16
				100.00	100.00

* Includes *List* Item Markers (LS = 0.26%) and *Symbols* (SYM = 0.2%)

Table 2-8 Comparative POS distributions of the Native and Learner BLCs

These and other similar pieces of evidence suggest that it is quite reasonable to assume that we would have a much larger proportion of the definite and indefinite articles in the Learner Corpus if it were not for such errors. The same can also be said about PREPOSITIONS, although it is not necessarily clear whether errors in this category result in the increase or decrease of the actual number of prepositions.[27]

b) Adjectives

The lower ratio of ADJECTIVES in the Learner BLC (5.38% vs 7.21%) implies that the business messages written in English by the Japanese business people is less "colorful" than those of native speakers of English. Table 2-9 shows that the total number of adjective types used in the Learner Corpus is 719. About 44% of which are so-called *hapax legomena*, or words that appear only once. This means that the L2 users of English tend to use same adjectives repeatedly, so that only 172 adjective types cover 90% of all the adjectives used in the corpus. In contrast, native speakers of English make use of about 4.24 times more adjectives (N=730) to achieve the same coverage.

	Native BLC	Learner BLC
Total No. of Words in the Corpus	1,092,589	210,404
Total Adjective Tokens	76,820 (7.21%)	11,330 (5.38%)
Total Adjective Types	3,923	719
Type-Token Ratio	1: 10	1: 16
No. of Adj. Types with Freq. = 1	937 (23.88%)	318 (44.23%)

No. of Adj. Types needed to achieve:	Native BLC	Learner BLC
50 % Coverage of Total Adj. Tokens	59	30
60 % "	101	42
70 % "	174	59
80 % "	314	90
90 % "	730	172
95 % "	1,420	288

Table 2-9 Comparative data for ADJECTIVES in the Native and Learner BLCs

This, however, does not come as a surprise at all considering the fact that English, after all, is not even a second language for most of the Japanese business people. Nevertheless, all indications are that the limited amount of adjectives as such are not causing any serious problems for the Japanese writers of business English in saying what need to be said in the messages they write. In fact, the data in the Learner Corpus indicate that we only need a certain amount of adjectives to write well-expressed business messages. The problems obviously lie elsewhere.

c) Modals and Conditional *if*

It is also worth mentioning that the proportions of MODALS and Conditional *if* are equally high in the Learner BLC, indicating their importance in business writing for both native and non-native writers of English for business purposes. The observed differences in terms of percentage figures in these three categories are rather small; however, the difference of 0.59 percentage points in MODALS may indicate that the Japanese business people are not making full use of this grammatical category. The difference of 0.29 percentage points in Conditional *if*, on the other hand, indicates that the Japanese writers of business messages use this item much heavily than do their native-speaker counterparts, most likely due to their inability to use other lexical items and/or grammatical devices than *if* that are available to us to express conditional meanings. The major pedagogical implication of these observations seems clear: that the Japanese learners of business English need to be encouraged to learn more about modal auxiliaries and conditional constructions without which business messages are almost impossible to write.

2.6 Summary

In this chapter, we analyzed the corpus data from a viewpoint of POS distribution. We found that, on average, NOUNS comprised approximately 27.57% of all the word tokens in the research corpora, followed by PREPOSITIONS (13.35%), VERBS (11.18%), DETERMINERS (10.17%), ADJECTIVES (7.7%), ADVERBS (5.21%), and PRONOUNS (7.21%). When the BLC and the Reference Corpora were compared binominally for each of these POS categories, the most significant differences were found in MODALS, CONDITIONAL *if*, PRONOUNS, INFINITIVAL *to*, DETERMINERS, and WH-PARTICLES, followed at a lesser degree by VERBS, EXISTENTIAL *there*, and INTER-JECTIONS.

We then continued our discussion on each of these statistically significant areas to consider possible reasons that account for these differences and to identify possible topics for further discussion in later chapters, particularly in Chapter 4. As to EXISTENTIAL *there*, we made a slightly more detailed discussion than those of other categories since this particular item was considered not likely to be dealt with in later chapters.

In Section 2.5, we compared the Native and Learner BLCs in terms of POS distribution, and found that the overall patterns of POS distribution of two corpora were very similar and, in fact, very much the same. This indicates that native and non-native English are after all not that different as to the proportions of nouns, verbs, prepositions and other parts of speech people use in their writing.

A closer look at the data, however, indicated some noticeable differences in several areas including, but not limited to, NOUNS, DETERMINERS, and ADJECTIVES. As to NOUNS, we argued that the relatively higher proportion of this category observed in the Learner BLC was largely due to the effect of "nominalization"— which, we presumed, will also explain both the lower proportion of VERBS and the higher proportions of DETERMINERS and PREPOSITIONS in the Learner BLC. This topic will be explored further in Chapter 4.

With regard to ADJECTIVES, the data indicated that the Japanese users/learners of EBP use much less adjectives in their writing than their native counterparts. Nevertheless, the fact that most of the business messages the Japanese writers write actually do what they are supposed to do in every practical sense with all the grammatical errors and stylistic awkwardness, means that we only need a fairly limited amount of adjectives to write well-expressed business messages — some-where between 170 to 730, and most likely around 300, as we will see in more detail in Chapter 4. We have also touched upon MODALS and Conditional *if*, and concluded that the Japanese EBP learners need to be encouraged to learn more about these two most important lexical devices. Again, we will come back to this issue in later chapters.

With all these findings and new topics for discussion that derived therefrom in mind, we now move on to the next chapter, in which we discuss the BLC lexicon in more detail focusing on individual lexical items rather than simply treating them as members of respective POS categories.

Endnotes to Chapter 2

[1] Compare the following examples, in which (1a) represents the nominal style and (1b) the verbal style:

(1a) We are engaged in *the manufacturing of* precision instruments.

(1b) We *manufacture* precision instruments.

[2] The following examples will make the point clear, in which (2a) represents the nominal style with non-human subject, and (2b) the verbal style with human subject (actor) and object (beneficiary).

(2a) Provision of the detailed information about the product is strictly prohibited.

(2b) *We* cannot provide *you* with the detailed information about the product.

[3] Compare the following examples:

(3a) I need one week *for the completion of* this report.

(3b) I need one week *to complete* this report.

(4a) I requested *his coming* to my office *for further discussion of* the problem.

(4b) I requested him *to come* to my office *to discuss* the problem further.

In the standard theory of the transformational grammar, (3a) may be analyzed as having an underlying deep structure consisting of one *S*, and thus not much different from the actual sentence, or the surface structure. (3b), however, would be analyzed as having an underlying phrase marker consisting of two *S*'s, representing the two underlying predications, [I need one week.] and [I complete this report.]. Similarly, (4a) has an underlying deep structure consisting of one *S*, whereas (4b) has an underlying phrase marker consisting of three *S*'s, representing the three underlying predications, which could be informally given as [I $_{\mathrm{PAST}}$ request him.] [He $_{\mathrm{3PerSg}}$ come to my office.] [We discuss the problem further.]. Thus, from a syntactic viewpoint, (3b) and (4b) are more complex than their nominal-style counterparts, (3a) and (4a) .

[4] Prior to the tagging, the original Brill Tagger was tested for its accuracy and it was found that the accuracy rate was not as high as expected (less than an average of 90% accuracy) when applied to a text on which the tagger was not " trained." I have, therefore, modified the tagger by adding some 200 new rules to the original "contextual rule file," a rule file that reads the preliminary (or "start-state" to use the wording of Eric Brill) tagging data and rewrites inappropriate tags according to the given syntactic environment of each word (See Brill (1993b) for more details about the contextual rule file). This, in effect, boosted the tagging accuracy to an average of 96.26% (SD=4.44). Although this accuracy rate is considered satisfactory from a statistical viewpoint, a caution nevertheless should be taken in interpreting the results of the current POS tagging and the arguments presented in this paper based on them, that there is always the chance of an error at an approximate probability of $p < 0.04$.

[5] The following conversion table was used (excerpt). Note that "`NN -> NN$, NNP, NNP$,`

NNPS, etc." means that the tags NN$, NNP, NNP$, NNPS, etc. are converted, or lemmatized, to NN. The PN tag in the table corresponds to the PRP tag of the Brill Tagger (For details of the LOB tag set, see Appendix B2):.

```
---------------------------------------------------------------------
[loblemma.dic
[Aug 26, 1989
[LOB tags conversion table to be used with WordSmith
[Table begins
NN -> NN$, NNP, NNP$, NNPS, NNPS$, NNS, NNS$, NNU, NNUS, NP, NP$,
      NPL, NPL$, NPLS, NPLS$, NPS, NPS$, NPT, NPT$, NPTS, NPYS$, NR,
      NR$, NRS, NRS$
PN -> PN$, PP$, PP$$, PP1A, PP1AS, PP1O, PP10, PP1OS, PP10S, PP2,
      PP3,PP3A,PP3AS, PP3O, PP30, PP3OS, PP30S, PPL, PPLS
DT -> DT$, DTI, DTS, DTX
AT -> ATI, AP$, AP, APS, AP$, ABL, ABN, ABX
VB -> VBD, VBG, VBN, VBZ
BE -> BED, BEDZ, BEG, BEM, BEN, BER, BEZ
HV -> HVD, HVG, HVN, HVZ
DO -> DOD, DOZ
JJ -> JJB, JJR, JJT, JNP
RB -> RB$, RBR, RBT, RI, RN, RP, QL, QLP
WH -> WDTR, WP, WP$, WP$S, WPA, WPO, WPOR, WPR, WRB
CD -> CD$, CD-CD, CD1, CD1$, CD1S, CDS, CS, OD, OD$
[...]
---------------------------------------------------------------------
```

Figure 2-1 LOB tags conversion table (excerpt)

[6] The program source of prn_tag.awk is as follows:

```
# ====================================================================
# prn_tag.awk (Yasumasa Someya, 1 June 1998; Revised 21 Aug 1998)
# Usage: jgawk -f prn_tag.awk INFILE > OUTFILE
# Function: Extracts POS tags from a tagged corpus.
# ====================================================================
{
s=$0
gsub("_", "_ ",s)
gsub(/[A-Za-z0-9<({[¥]})>.,:;'¥-?!¥¥$£%+=@]+_ /," ",s)   # See note
printf "¥rExtracting POS Tags. Please wait... %7d",NR > "CON"
print s
}
# Note: This script assumes that the specified non-alphanumeric symbols
# and punctuation marks have been tagged self-recursively (e.g. <_<, (_(,
# [_[, ._., :_:, etc.), rather than having been assigned the default SYM
# tag of the Brill Tagger.
# ==========================================================
```

Figure 2-2　prn_tag.awk (for POS tags extraction)

[7] Statistical measure first proposed in Yule (1944) and subsequently adopted in Hofland and Johansson (1982).

[8] See Kiyokawa (1990, pp. 63-74).

[9] This rank order generally corresponds to the one reported in Kennedy (1998: p. 123) for the original Brown and LOB Corpora with some minor differences reflecting the difference in the tagging schemes already mentioned.

[10] The large proportion of NOUNS in the TIME Corpus can be accounted for the fact that this corpus contains many proper nouns (*e.g.* names of people, organizations, places and so on)－many of which are so called *hapax legomena*－or words that appear only once. The observed differences in this category between the TIME Corpus and other two corpora, however, are not statistically significant.

[11] The MSL of the three Reference Corpora is 20.35 words. Note that the official MSL figures as reported in Francis and Kuc&era (1982) for the Brown Corpus is 21.06 words for the informative genre, whose text type is roughly comparable to that of the BLC, and 13.28 words for the imaginative genre.

[12] The figure is calculated in the following formula:

(TWT÷(TWT×Ptg.))÷MSL

where

TWT = Total Word Tokens of the BLC

MSL = Mean Sentence Length of the BLC (= 16.68 words: K see Endnote 19)

Ptg. = Percentage of the item in question to TWT (=0.39% for *if*)

Thus, we get

(1092589÷(1092589×0.39%))÷16.68 = 15.37

[13] See, for instance, Suwabe (1985) and Shiozawa (1985).

[14] The breakdown figures for each corpus is as follows: 15,816 times in the Brown Corpus (≒1.55%); 16,090 times in the LOB Corpus (≒1.58%); and 18,229 times in the TIME Corpus (≒1.69%). These figures suggest that the distribution of the infinitival to is not influenced by the American-British distinction, but is more dependant upon the text genre, although the observed difference between the Brown (or the LOB) Corpus and the TIME Corpus is not statistically significant. The possible reason for this is that both the Brown and LOB Corpora are sort of a "waste basket" containing texts from as many as 15 different categories (roughly divided into two main genres: informative and imaginative), so when taken as a whole, any genre-specific features tend to be concealed by an offset effect.

[15] Sample sentences are quoted from the BLC. At the end of each sentence is the Sentence ID ([BZ01:00527] means that it is the 527th sentence of BLC subcorpus BZ01).

[16] The "Keyness" scores are −9,160 for *the* (Rank 5412) and −2,341 for *a* (Rank 5409) respec-tively. For further details, see Appendix D3: BLC KEYWORDS LIST.

[17] Also see sample sentences quoted in Endnote 1 of this chapter.

[18] The syntactic function of the preposition *of* is to assign objective case (or, oblique case) to the NP that follows immediately after it. In a verb phrase like "create NP," the NP is automatically assigned objective case by the transitive verb that precedes it. When the verb is nominalized, however, the NP would be syntactically left stranded without the case-assigning preposition *of* as in "(the) creation *of* NP." (The Case Filter theory states that an overt NP must have Case at S-structure; otherwise, the entire sentence is blocked as being ill-formed (Nakamura *et al.,* 1989)). Thus, the number of *of* will increase in proportion to that of nominalization in a given text.

[19] Having said this, I quote the following data to substantiate my claim. I will not, however, make any extended argument here with regard to the data since the current study excludes syntactic analysis of the corpus data from its scope.

	Mean Sentence Length	Ave. No. of Predications per Sentence 1)	Ave. No. of Words per Predication
BROWN 2)	20.6726	3.6918	5.5438
LOB 3)	20.4871	3.5834	5.6927
TIME	19.8970	3.4031	5.6838
BLC (Native BLC)	16.6838	3.0830	5.7327
Learner BLC	15.3409	2.8334	5.1095
Mean	18.6163	3.3190	5.5525
SD	2.4409	0.3562	0.2577

1) Predication is defined as "any verb or verbal group with a tensed verb having a grammatical subject . . . and infinitives, gerunds and participles" (Francis and Kuc&era, 1982: p. 550). In this definition, VPs like "ask me," "will go there," "would have gone by now" and "was asked by someone" are equally counted as one instance of predication. `predcnt1.awk` was used for automatic counting of predications.

2) Official figures as reported in Francis and Kuc&era (*ditto*, p. 552) for the Brown Corpus are: MSL = 18.4 (Informative Text = 21.06; Imaginative Text = 13.28); Pred/Sent = 2.64 (Infor-mative Text= 2.78; Imaginative Text = 2.38); Wrd/Pred = 6.96 (Informative Text = 7.57 Imaginative Text = 5.62).

3) Analysis of the LOB Corpus is based on the tagged version of the corpus provided by the ICAME.

[20] Halliday distinguishes three types of Subject: psychological, grammatical and logical. The psychological subject is "that which is the concern of the message." The grammatical subject is "that of which something is predicated," and the logical subject refers to the "doer of the action." (Halliday 1994, pp. 30-33)

[21] "the+NP" can also occur in this construction if what is being predicated is still indefinite in nature, as in the following examples.

(13) In our office, there is *the one individual* whom everyone gets along with and enjoys having around, no matter the occasion. [BZ09:03557]

(14) There is also *the problem of our retailers*, who will be very unreceptive to a price increase so soon after introduction. [BZ19:12310]

[22] It is, however, possible to consider the use of *should* in this construction simply as a replacement of *if* without assuming the inversion operation. The use of non-finite *be* in "should there be" is a natural consequence of using the modal *should* in place of *if*.

[23] In the following example, the use of *should* in the embedded conditional clause is considered syntactically motivated. By using *should* in the second clause, the writer avoided a simplistic repetition of *if* which is already used in the first clause.

(31) *If* this summary is not consistent with your understanding, or *should* you have any questions, please contact me as soon as possible. [BZ29:00421]

[24] POS tags were assigned to the Learner BLC and relevant data obtained in the same manner as was described in Section 2.2.

[25] The hypothesis can be reformulated as follows: "Business messages are basically action-oriented, being more concerned with human participants and their actions in a particular event. The verbal style, therefore, is generally more preferred in business writing than the nominal style. This will naturally results in a greater frequency in business writing of lexical verbs on one hand, and lesser frequencies of verb-derived nouns (nominalizations), determiners and prepositions on the other, than are normally expected. The Japanese learners of business English, however, tend to use NPs in cases where native speakers of English would have used VPs (largely as a result of 'negative transfer' from their L1) and, as a result, use more determiners and prepositions in their writing than are otherwise required－although they are often omitted by error."

[26] My comments and possible revisions of these four sentences are as follows:

(42): The NP "for submission of" should have been "for the submission of" to be at least grammatically correct. The entire sentence, however, sounds still awkward even after all the other surface errors are corrected. A possible revision (a verbal-style version) of this sentence would be: *In my November 1 memorandum, I requested all the department managers to submit their budget estimates for the first quarter of 1998 (by the end of December).*

(43) The NP "issuance of" should have been "the issuance of" to be grammatically correct. A possible revision of this sentence would be: *We are planning to issue a bilingual*

monthly magazine for our employees.

(44) The VP "made agreement" should have been "made an agreement," and a much simpler and grammatically correct verbal-style alternative of this sentence would be: *I contacted Mr. Kawai and he agreed to this (proposal).*

(45) The VP "make suggestion" should have been "made a suggestion" or "made the following suggestion." A much better version of this sentence, however, would be: *In view of (what I discussed in) the previous paragraph, I suggest as follows:*

[27] What the data suggest is that the negative and positive errors are likely to offset each other, showing no significant deviation in either direction as if there were no errors at all.