CHAPTER 1

INTRODUCTION

"There is only one human language, apart from the lexicon, and language acquisition is in essence a matter of determining lexical idiosyncrasies."

(Chomsky, 1989:44)

1.1 Background

In international business, English has now been firmly established as the *lingua franca*, and the ability to use it is considered one of the basic prerequisites for business people whose first language is other than English. A recent questionnaire survey conducted in Japan by this author shows that an overwhelming majority of the respondents (98%) answered that English is necessary at their workplaces[1], confirming the prevailing perception and reality of English as the *de facto* common language of business.

In countries like Japan where English is learned only as a school subject, people usually do not have a real motive to learn it—other than, perhaps, to get a good grade on their school reports for whatever reasons. But once one graduated from school and started working for a company, he or she suddenly comes to feel a pressing and immediate need for it—because, as mentioned above, it has now become one of the prerequisites to be a capable business person. Naturally, people are investing a lot of time and money to learn English for Business Purposes, or EBP. As a result, there is currently an abundance of EBP teaching/learning materials available everywhere on the market.

The problem is, however, that most of these materials are written on the basis of intuitive judgments of respective authors as to what make EBP unique from other areas of ESP (English for Specific Purposes) or from EGP (English for General Purposes) in its lexical, grammatical and pragmatic aspects. It may be that most of these "intuitive" judgments, after all, properly reflect the reality of the language use, for most of the authors are themselves experienced and expert users of the language. Nevertheless, the fact that they largely lack the support of empirical data and theoretical grounds with regard to the very judgments they make remains a major problem from a pedagogical viewpoint.

This lack of theoretical and data-driven objective perspectives, however, is

simply a reflection of a general scarcity of scholarly research in the area of EBP. Although we had witnessed a sudden proliferation of ESP research during the 1960's through the 1970's (Tazaki 1995, p.233-235), the focus of those research had been either on EAP (English for Academic Purposes) or EST (English for Science and Technology), and research in the business area to data has been largely lagged behind. In reviewing the current state of research and published teaching materials relating to Business English, Tony Dudley-Evans of University of Birmingham, who is also the editor of *English for Specific Purposes*, one of the few scholarly magazines specialized in ESP studies, concludes as follows (Dudley-Evans & St. John 1996, pp. 1-13):

> "All discussions of Business English reflect the paucity of research in the area. . . . It is surprising that relatively little analysis of actual business genres exist. "

As a result, "there is not yet an established common core of business language in the way that there is a relatively well-agreed core of semi-technical lexis and grammar that is widely used in academic, scientific and technological subjects and that has formed the basis of EST textbooks" (*op. cit.*, p. 5. Also see St. John 1996, p. 5). In short, "we know surprisingly little about language usage and performance in business contexts" (Holden, 1989, p. 43).

Given this state of research in the area, it is only natural that material writers have had no choice but to rely heavily on their personal experience and informed intuition in preparing much needed EBP textbooks. Yet, it is indeed curious, as Holden (*ibid.*) writes, that "this matter has been neglected for so long" in view of the fact that how people use language in business contexts undoubtedly represents "one of the most potential social influences on modern life."

1.2 Purposes and scope of the study

The current study has been conceived against this background, with the primary purpose of identifying and describing some of the major lexical and grammatical features of written business English, leaving the syntactic and textual/ discoursal aspects for future studies. It is the author's hope that the various findings and arguments presented in this study will not only form a reliable basis for future research in this important area, but also provide teachers of EBP with much needed objective data with which to guide their educational endeavor—be it classroom instruction or material writing.

Being a practitioner himself, the author's viewpoint in carrying out the current study is more pedagogical than academic. As such, the ultimate goal of this study

is to produce a series of teaching and self-learning materials in the area of EBP, with the Japanese business people as the primary target, rather than trying to deal with some limited sets of academically significant issues in a greater depth. This, however, does not mean that the study ignores academic precision in both study methodology and argument construction. On the contrary, the author intends to make every effort to respect existing academic tradition by, for instance, making clear as much as practically possible the study methods and procedures through which particular sets of data have been obtained and conclusions drawn therefrom. Also, it is the author's intention to refrain from making arguments that are not data-driven, or from presenting a claim that has no empirical basis to substantiate it.

The other purpose of the study the author has in mind is to establish a computer-based study methodology appropriate for the analysis of such large-scale corpus data as those used in the current study. This naturally includes the development of a series of computer programs with which to analyze the corpus data from various viewpoints and for specific purposes. This constitutes an important part of the study, not only because it will largely determine the quality of the current study, but also because the success or failure of the future studies that the author plans to undertake depends on the successful development of such a computer-based study methodology.

Currently, there are many linguistic data-analysis software packages available on the market. Some of them, such as WordSmith Tools and TXTANA, are quite useful and, in fact, they have been instrumental in carrying out the current study. Nevertheless, they are no panacea for all the possible problems: they do only what they are designed to do. Thus, in retrospect, the author ended up in spending much, if not the most, of the time in writing, testing and rewriting various computer programs needed to carry out all the tasks that formed indispensable parts of the current study. It was, however, the time worth spent. The computer programs written for the current study will prove useful for other researchers as well in carrying out otherwise cumbersome and time-consuming, if not error-prone, tasks of large-scale text data analysis. The author hopes to make available these computer programs in a ready-to-run format for interested researchers once the current study is completed, so that they can be tested, modified or otherwise used at the user's disposal.[2]

1.3 Working hypotheses

In carrying out the current study, the author established the following three hypotheses related to lexical and grammatical features of EBP in its *written* domain. It should be noted, however, that these hypotheses are more practical than theoretical in nature in that they function primarily as guidelines around which to

organize the main part of this paper. The three hypotheses and related topics of discussion are as follows:

1) The "lexical closure" hypothesis

Our first hypothesis relates to the relative size of vocabulary needed to carry out business-related everyday discourse, and can be stated as follows:

> "Written Business English is characterized by a high degree of lexical closure, and only a very restricted number of vocabulary is being used in day-to-day written business discourse."

To prove this hypothesis and to illustrate its pedagogical implications in more concrete terms, the following topics will be discussed:

- How closed is the vocabulary of EBP in comparison with those of other genres?
- In terms of POS (parts-of-speech) distribution, what categories demonstrate notable closure?
- Is it possible to quantify the "very restricted number of vocabulary" and to extract the "core" part of the EBP vocabulary for each of the major POS category?
- From a pedagogical viewpoint, what are the most important lexical items among the "core" vocabulary?

2) The "lexical ease" hypothesis

Our second hypothesis refers to the level of relative difficulty of the "core" part of the EBP lexicon, and can be stated as follows:

> "Written Business English is characterized by a relatively low level of lexical difficulty, and most of the core part of the EBP vocabulary are covered by the *4000 Basic Words* as defined by the JACET (JACET, 1993)[3]— meaning that they are well within the knowledge of the average Japanese adult learners of EGP."

In more specific terms, the following topics will be dealt with concerning this hypothesis:

- What are the criteria of defining the lexical difficulty?
- Given these criteria, how exactly "easy" or "difficult" is the EBP vocabulary?
- Is this general characteristics of "lexical ease" of EBP vocabulary applicable to all POS categories, or is it limited to some particular area(s)?

- Is the lexical ease hypothesis also applicable to the vocabulary used by the Japanese non-native users of EBP?
- What are the most manifest problem areas for the average Japanese learners of EBP with regard to the proper acquisition of the apparently easy EBP lexicon?

3) The "write-as-you-talk" hypothesis

The third hypothesis relates to stylistic features of written business discourse, and can be stated as follows:

> "Written Business English is characterized by its incorporation of spoken features of the language into written texts. In other words, most business messages exhibit the characteristics of one-on-one 'conversation'—albeit it is conducted on a piece of paper or, more recently, on the PC terminal. "

To prove and substantiate this hypothesis, the following topics will be discussed in this paper. As to the first topic, we will be looking specifically at such lexical and grammatical items as *personal pronouns, articles, prepositions* (*of*, in particular), *nominalization, contractions, WH-particles* (*which*, in particular) and *relativization* for reasons that will be made clear in later chapters.

- In what aspects of language use is the oral/informal characteristics of EBP most salient?
- How EBP compares with English used in other genres in its stylistic characteristics?
- What are the pedagogical implications of these particular characteristics of written business discourse?

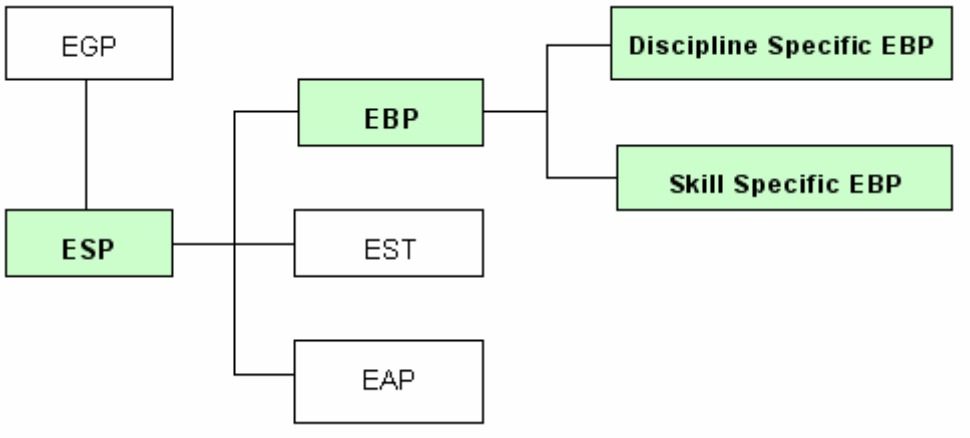
1.4 Definition of "Business English"

In this paper, the term *Business English* is defined as English used for business purposes in international and inter-cultural contexts as a common means of communication by the people who do not share a first language. Thus, it is seen primarily as a branch of ESP in that it specifically refers to the teaching of non-native speakers of English who need to learn to use the language, not for general purposes, but for a specific purpose of achieving business-oriented goals. As such, it is "distinct from the teaching of communication skills related to business for native speakers" (Dudley-Evans & St. John 1996, p. 1).

In the Japanese context, ESP is an extension of EGP in the sense that ESP students are usually an identifiable group of adult learners at the post-EGP level as shown in the following chart (Figure 1-1). Their training in ESP, therefore, is to

supplement the formal English education they received in the school system from a more practical viewpoint. This involves, among other things, a major shift from English as a school subject to English as a means of communication for specific purposes.

Figure 1-1 ESP and its major subcategories[4]



[Key] *EGP* (English for General Purposes = English taught at the Junior High School, High School, and College Levels as a school subject),
ESP (English for Specific Purposes), *EBP* (English for Business Purposes), *EST* (English for Science and Technology), *EAP* (English for Academic Purposes),

The chart shows that ESP can be divided into three major subcategories, EBP, EST and EAP, all of which represent the areas in which the dominance of English as the common language is most conspicuous. EBP, which is the target of the current study, can be further divided into "Discipline Specific EBP" and "Skill Specific EBP." The former refers to the teaching of EBP with a focus on specific business contexts, such as banking, accounting, advertising, marketing, international trade and so on. This approach to EBP reflects the view of ESP as "General English with a set of job-specific terminology" that was prevalent in the 1970s. The later refers to the teaching of EBP with a focus on one or more of the four traditional areas of linguistic skills: writing, reading, listening and speaking (to which the fifth category, "meta-linguistic skills," is sometimes added).

Having thus defined and classified ESP, we can now say that the main topic of the current study relates to the writing section of Skill Specific EBP, with a particular emphasis on its lexical and grammatical components and dealing primarily with the "core" parts thereof.

1.5 Review of previous research

As mentioned earlier, research into Business English has largely lagged behind compared with what have been done so far in other areas of ESP. This doesn't mean, however, that there has been no research in the area that deserves our attention. In what follows I will briefly review some of the major research related to the lexical and grammatical features of Business English. Other aspects of Business English, including syntactic, textual and pragmatic aspects of written business discourse, are not dealt with in the current study, although most of the worth-mentioning research have been conducted in these areas.

1) Overseas research related to EBP vocabulary and usage

According to Fries (1950), one of the first researches conducted on the vocabulary of Business English is Horn (1923). Horn collected 1,125 business letters received by a bank in the state of New York and compared the vocabulary contained in these letters (a total of 2,623 word types) with that of more general, non-business specific correspondence to see if there was any significant differences between these two sets of vocabulary. Horn's conclusion was as follows:[5]

> "This study does not show that the words used in bankers' correspondence apply at all exclusively to the banking business, and yet the list is in a certain sense a vocational list."

Over seventy years having passed since the above study, it is not possible to obtain more accurate data as to the contents of the wordlist he was talking about. Yet, his conclusion that there was no such words that applied exclusively to banking business comes as no surprise. After all, any words used in whatever field can also be used, and they often do, in other fields as well, as long as they form part of the entire English vocabulary. If one wish to discuss genre-specific features of a lexicon, therefore, what one needs to do is not try to find words and expressions that are used "exclusively" to a particular genre, but to try to identify recurring or otherwise significant patterns of lexical use in a given genre.

As implied in Horn's conclusion, the vocabulary of EBP is much less specialized than those of other fields of ESP and is "a lot nearer to the everyday language spoken by the general public" (Picket 1986, p. 16). Therefore, it is generally considered difficult to identify its characteristics from a lexical viewpoint. Picket (1986), however, showed that such words as *order, invest, invoice, purchase* and *overdue*, to mention only a few for an illustrative purpose, were more "businesslike" than others in terms of their comparative frequencies of occurrence. The lexical items he pointed to are by no means limited to the EBP field but are

used in other areas as well, but the point is that these items exhibit particularly higher frequencies of use in written business discourse—suggesting that they may constitute "keywords" of EBP.

Irgl (1986) conducted an interesting study focusing on the relationship between genres and synonyms. By analyzing a corpus of business and economics texts, Igle concludes that those concepts that are significant to a particular genre tend to be expressed in many synonyms. The term, *price fluctuation*, for instance, were found to occur in 60 different synonyms in Irgl's corpus. This study is very interesting in that it suggests, rather convincingly, the possibility of extracting genre-defining "key" concepts or terminology by looking specifically at the distribution of synonyms.

Stabbs (1996) conducted research on the use of "performative verbs" in business documents, and found that there were striking differences in the patterns of use with regard to the performative verbs observed in real data and those appearing in invented data used in the speech act literature. According to Stabbs, the invented data consist mainly of examples which combine first-person singular with a simple present tense form (i.e. *I promise, I warn, I advise,* etc.). "But a striking feature of real data," writes Stabbs, "is that such forms are rare, and are restricted to certain verbs or to very formal contexts. *I apologize* is the only such form which is common. *I thank you* co-occurs with *Dear Sir* in a letter from a bureaucracy. *I hereby certify* occurs in legal form . . ." (Stabbs 1996, p. 212).[6] Such first-person simple verb forms have particular stylistic implications in themselves and, therefore, cannot constitute what Searle (1969) referred to as a "paradigm device."

This example shows, among other things, that the data invented to explain a particular theory often do not reflect the reality of language use properly and can thus be very misleading. What Stabbs pointed out is quite important from a pedagogical viewpoint as well, since most of the "core" verbs of EBP are performative or illocutional verbs and the proper understanding of how these verbs are used is one of the key points for effective business writing.

Ghadessy (1992) discussed the lexical characteristics of "Singapore English" based on a rather small corpus of 66,240 word-tokens. The corpus, known as "Singapore Written Business Communication Corpus (= SWBCC)," was originally compiled by his colleague, Jonathan Webster in 1983 (See Webster 1984 for details), and consists of a total of 566 samples of business letters, memorandums, telexes and other types of business-related documents collected from various companies, banks, and institutions operating in Singapore. Ghadessy analyzed the corpus data from a viewpoint of Hallidian functional grammar, using the triad analytical framework of "Participant, Process, and Circumstance." The Participant, according to Ghaddesy, is indicated by nominal groups (in subject or object positions), the

Process by verb phrases, and the Circumstance by prepositional phrases and adverbs. He finds that "the participants in the business world are concrete objects," rather than abstract processes, relationships, or resultant states of being as is often the case with scientific and academic discourse. Some of the most frequent "participants" common nouns he cites are *letter, policy, credit, order, payment, account* and *company*. As to the Process, he identifies two types of process: Material and Mental. Some of the most frequent "material process verbs" identified in the SWBCC are *confirm, indicate, render, advise,* and *look forward to*. Among the most frequent "mental process verbs" are included such items as *wish, hope, regret*, and *agree*. With regard to the Circumstance, Ghaddesy mentions 21 most frequent prepositions and adverbs, and goes on to discuss how typically they are used in the corpus.

He also touches upon the differences and similarities between the vocabulary of Singapore English and those of other "standard" varieties of English, and concludes that "there are many more similarities" than differences, and suggests the possibility of establishing "a core vocabulary for international business English" (*op. cit.*, pp. 99-100). Although the corpus he used is too small to be a representative sample of "Singapore English," or of EBP for that matter, this study not only yielded many interesting results, but also suggested a very useful analytical framework for future research in this area.

2) Japanese research related to EBP vocabulary and usage

In Japan, there has been very little corpus-based lexical studies done on EBP to date, except for some small-scale exploratory studies. The only exception, perhaps, is the series of papers written by Mitsuo Nakamura based on his one-million-word-plus Model Business Letter Corpus (MBLC). In one of his most recent papers (Nakamura, 1995),[7] he attempted to identify the "notions" that are most characteristic of Business English within the general framework of the notional-functional approach proposed in Brieger and Comfort (1992). First, he lists up 37 major notions which represent a total of 1,208 different hyponyms. Among those 37 notions, he finds the following 14 notions particularly prominent in business discourse—appearing at least twice more frequently in the MBLC than in his reference corpora of General English:

SATISFIED, REGRET, SUGGEST, WILLING, THANK, HAPPY, APOLOGIZE, REQUEST, SYMPATHY, PROMISE, PRAISE, HOPE, ABLE, ALLOW

Having thus identified the most "businesslike" concepts, he goes on to propose an outline of a model syllabus showing how best these notions be presented to the

Japanese learners of EBP. Although still sketchy, this is a very interesting and thought-provoking proposal and, along with other papers he has written so far, deserves due attention of other researchers in EBP or ESP.

Matsumoto (1983) compiled a frequency wordlist of EBP based on his corpus consisting of 691 business letters (35,634 tokens) collected from Japanese companies. One of the major problems with his study, other than the corpus being far too small to make any generalized statement, is that he didn't use a reference corpus. Consequently, when he says, for instance, that the word *order* was used 784 times in his corpus, we have no clue of evaluating the significant of this number. Another problem is that he didn't make any word-class distinction for such words as *order, request,* and *promise* whose verb forms and noun forms are the same in their bare forms. Despite these and other shortcomings, the study still merits a certain significance as an exploratory attempt in view of the fact that there has been very few similar studies conducted so far.

Kadota (1985) presents an outline of a "semantic frequency list" of EBP with reference to West (1953). His proposal is a very interesting one, but is still too sketchy to make any comments on it and we can only hope that the proposal be further developed as he wrote he would.

Nagano (1991) is an attempt to describe some of the most important vocabulary items of Business English from a lexico-grammatical perspective similar to the one proposed in Konishi (1980), albeit on a much smaller scale. In this work, Nagano selected a total of 67 verbs, nouns, adverbs and adjectives which he found most important and/or problematic for the average Japanese learners of EBP, and described each of them in a greater detail including such information as to markedness, collocation, comparative frequency, as well as the syntactic and semantic constraints of each entry. This work is apparently corpus-based, but he didn't mention the size of his corpus, nor did he make clear on what grounds the selection of the 67 entries had been made. Although this work obviously was not meant to be an academic endeavor, but was written primarily as a reference book, such information would have added more credibility to what he has done in this very important piece of work.

Isogai and Hosoda (1978) is a corpus-based dictionary of collocational patterns of nouns, verbs and other lexical items that are used very frequently (but often erroneously by the Japanese writers of EBP) in business-related documents. The dictionary covers approximately 570 entries sorted in alphabetical order, and for each item the authors give a succinct description of how it should be, and should not be, used. They also give a sample sentence or two for each entry to give the readers a fuller picture of the typical syntactic environment in which each entry is used. The authors write that the dictionary (they call it a "manual" for Business

English usage) is based on a corpus of over 180,000 sample sentences they collected from business letters, memorandums, contracts, quotations, minutes of meeting, reports, and other business-related documents. Although it is practical in nature rather than academic, this work deserves a special mention here for its data-based approach. Had it been provided in a CD-ROM, however, it would have been more useful for the busy business people for whom this "manual" was written.

All in all, the above review of previous research only confirmed the unwilling statement made by Dudley-Evans that it reflects "the paucity of research in the area" (p. 2). It is particularly true with regard to the lexical aspect of EBP. We, in fact, know surprisingly little about what we are talking about, with so many things still left to be studied. It is this author's hope that the current study will contribute to adding much needed knowledge towards establishing the "common core" of EBP, particularly in the lexical domain, which will eventually form the basis of more reliable and effective EBP teaching materials within the Japanese context.

1.6 Corpora used in the study

As the title of this paper indicates, the current study is a corpus-based, data-driven study. As such, it heavily draws on corpus data for making any arguments as to the lexical and grammatical characteristics of written business English. The corpora to be used in this study are as follows:

1) Business Letter Corpus (BLC)
2) Reference corpora
3) Learner Corpus of English Business Letters (Learner BLC)

The first corpus, the Business Letter Corpus (BLC), constitutes the main corpus of the study. The reference corpora consist of three corpora of different types, *i.e.* the Brown Corpus, the LOB Corpus, and the TIME Corpus. The third corpus, the Learner Corpus of English Business Letters (Learner BLC) consists of English business letters written by Japanese business people.

1) Business Letter Corpus (BLC)

International written business communication is typically conducted in the form of the exchange of business letters, memorandums, e-mail and other business-related documents. Therefore, in the 1st phase of the study, I compiled a corpus of such documents in a computer-readable format. Most of the data were taken from a variety of EBP textbooks and reference materials published after 1980 (except for one instance whose year of publication is 1979), since "real" business documents are hard to obtain in a large number. The size of this corpus was

targeted at 1,000,000 words (tokens), so that it can match the size of the two main reference corpora to be used in the current study, *i.e.* the Brown and LOB Corpora.

The original version of the BLC (BLC_1) consists of 37 subcorpora from BZ01 to BZ37, as shown in Table 1-1 on the next page. BZ01 to BZ21 represent American English and BZ21 to BZ27 British, while non-native speakers, mostly Japanese, are involved in the writing of sample letters contained in BZ28 through BZ32. Of these 37 subcorpora, however, BZ07, BZ11 and BZ14 consist mainly of personal letters which are not related to business in the normal sense of the term and, therefore, have been excluded from the current research corpus (BLC_2) which is what we refer to as the "BLC" in this paper (it will also be referred to as the "Native BLC" when reference is made to the Learner Corpus of English Business Letters which we will discuss shortly).[8]

The BLC contains a total of over 11,000 business-related documents covering a wide spectrum of day-to-day business communication scenes including, but not limited to, such instances as inquiries, requests, refusals, apologies, complaints, claims, reminders, confirmations, invitations, negotiations, order processing, sales promotions, business proposals, customer relations, seasonal greetings, as well as various kinds of interoffice correspondence. Contractual and legal documents, however, have not been included in the current version of the BLC because they are considered to constitute a different genre in itself and should be treated separately.

Also, the BLC data are not discipline-specific, meaning that they are not confined to particular areas and/or types of business. A sample excerpt from the BLC is given in Appendix A2.

Table 1-1 List of subcorpora included in the Original Version of the Business Letter Corpus (BLC-1)

Subcorpus No.	Author ID	English Type [1]	Document Type	No. of Documents	No. of Word Tokens	No. of Word Types	TTR [2]	Standard TTR	No. of Sentences	MSL [3]	SD [4]
BZ01	Holt	AmE	Business Letters and Memos	316	24,643	2,754	11.18	38.15	1,744	13.52	7.33
BZ02	Pce	AmE	Business Letters and Memos	245	39,260	4,842	12.33	43.37	2,329	16.01	9.55
BZ03	Shipley	AmE	Business Letters, Memos and Others	24	7,461	1,865	25.00	42.99	479	14.03	8.30
BZ04	Frank	AmE	Business Letters and Memos	216	42,841	4,507	10.52	44.77	2,651	15.38	9.51
BZ05	B_Works	AmE	Business Letters and Memos	395	45,335	5,350	11.80	42.87	2,597	16.24	8.83
BZ06	S_Works	AmE	Business Letters and Memos	316	49,944	5,718	11.45	44.74	3,195	14.96	8.80
BZ07	P_Works	AmE	Personal Letters *	405	42,880	5,474	12.77	43.01	2,957	13.68	7.51
BZ08	MSOffice	AmE	Business Letters and Memos	15	1,733	591	34.10	40.20	111	14.89	7.62
BZ09	ModelOffice	AmE	Business Letters, Memos and Others	558	87,932	7,530	8.56	44.06	4,598	17.88	9.42
BZ10	ModelOffice	AmE	Business Letters and Memos	399	53,411	5,279	9.88	42.85	3,011	16.21	8.82
BZ11	ModelOffice	AmE	Personal Letters *	432	57,193	6,142	10.74	42.81	3,333	16.35	9.37
BZ12	B_PLUS	AmE	Business Letters and Memos	230	29,597	3,544	11.97	40.70	1,824	15.26	7.20
BZ13	S_PLUS	AmE	Business Letters and Memos	171	20,172	2,586	12.82	40.64	1,289	14.98	6.92
BZ14	P_PLUS	AmE	Personal Letters *	186	18,859	2,784	14.76	38.16	1,302	13.99	6.44
BZ15	LetterExpert	AmE	Business Letters and Memos	449	38,163	3,308	8.67	36.12	2,521	14.31	8.04
BZ16	Moscor	AmE	Business Letters	99	7,404	1,307	17.65	38.40	555	12.53	6.20
BZ17	Geffner	AmE	Business Letters, Memos and Reports	70	8,827	1,828	20.71	42.99	530	14.96	9.35
BZ18	Griffin	AmE	Business Letters and Memos	575	56,893	4,177	7.34	35.40	3,543	15.12	8.53
BZ19	Kudyra_1	AmE	Business Letters	3,039	225,748	4,827	2.14	18.24	15,229	14.11	8.07
BZ20	Kudyra_2	AmE	Business Memos and E-mail	42	4,811	1,334	27.73	43.20	413	10.04	8.06
BZ21	Baugh	AmE	Business Memos	62	11,320	2,459	21.72	45.25	623	15.65	7.13
BZ22	Ashley	BrE	Business Lettes and Memos	146	19,109	2,521	13.19	38.41	917	19.49	12.26
BZ23	Gartside	BrE	Business Letters	432	50,989	4,009	7.86	36.59	2,309	20.57	9.62
BZ24	Saville	BrE	Business Letters	74	7,242	1,441	19.90	38.89	350	19.01	10.73
BZ25	Bosticco	BrE	Business Letters	838	72,404	5,335	7.37	36.27	3,572	18.88	9.79
BZ26	Naterop	BrE	Business Letters	35	4,719	1,181	25.03	40.95	279	15.51	9.84
BZ27	Maitland	BrE	Business Letters	200	11,027	1,214	11.01	30.56	748	13.57	6.43
BZ28	Alex	MxE	Business Letters and Memos	100	11,853	1,652	13.94	34.29	707	15.83	10.46
BZ29	WM	MxE	Business Letters and Memos	76	9,035	1,501	16.61	36.31	568	14.50	10.31
BZ30	AlcEmail	MxE	Business E-mail	100	10,511	1,609	15.31	35.26	720	13.54	9.41
BZ31	RyanEmail	MxE	Business E-mail	369	24,898	2,855	11.47	36.76	1,827	12.68	6.59
BZ32	Wada	MxE	Business Letters	227	35,959	3,632	10.10	39.85	2,192	15.84	9.25
BZ33	Ikezaki	MxE	Business Letters	142	20,261	2,558	12.63	36.62	1,240	15.29	8.80
BZ34	Anonymous_1	MxE	Business Letters	206	31,175	2,997	9.61	35.96	1,566	16.43	12.85
BZ35	Anonymous_2	MxE	Business Letters	77	10,729	1,391	12.96	30.44	572	17.30	13.05
BZ36	IBC	MxE	Business Letters and Memos	254	20,532	2,874	14.00	39.80	1,212	15.55	11.19
BZ37	SoneyaFile	MxE	Business Letters	90	18,080	3,093	17.11	42.41	1,061	16.24	10.21
TOTAL				11,610	1,232,950	M=3,191	M=14.11	M=38.87	72,930	M=15.41	M=8.97
BLC-1 [5]				11,610	1,238,650 [6]	26,727	2.16	36.10	74,915	15.47	9.01

1) AmE =American English; BrE = British English; MxE = Mixed (Japanese writers involved) 2) TTR = Type-Token Ratio; Standard TTR = Standardized TTR per 1000 words 3) MSL = Mean Sentence Length 4) Standard Deviation of MSL 5) "BLC-1" includes all the subcorpora from "BZ01" to "BZ37." The three subcorpora indicated by the asterisk (*e.g.* BZ07, BZ11, BZ14) are personal letters and, therefore, have been excluded from the current research corpus, "BLC-2" which is what we refer to as the "BLC" in this paper. 6) The difference in total word tokens is due to comment lines inserted in the original data. (☞ See Appendix A1 for detailed information about each subcorpus.)

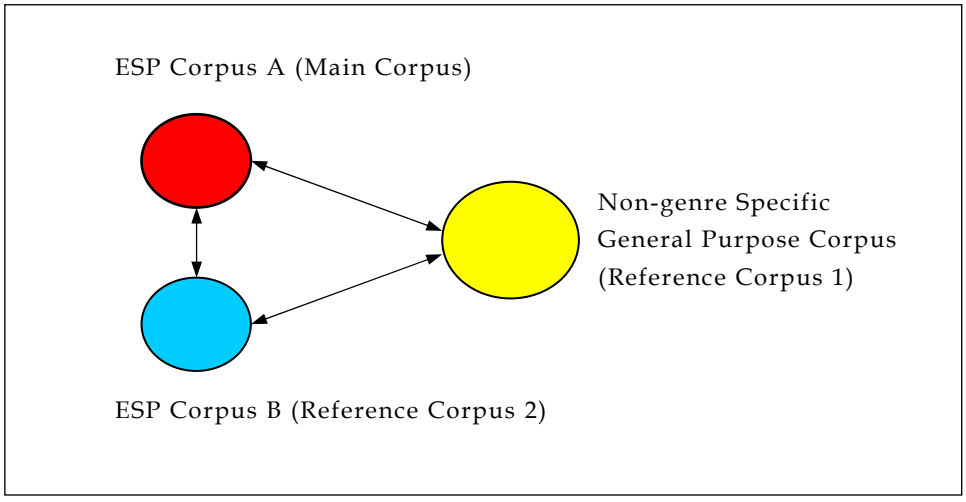
2) Reference corpora

A "reference corpus" as the term is used in this paper is a corpus, or a set of corpora, used for comparative purposes. Since the primary purpose of the current study is to extract lexico-grammatical features of the language used in a particular genre, *i.e.* EBP, it is desirable that we have another set of corpus data that represents non-genre specific General English as a reference corpus, so that we can have a comparative perspective in evaluating whatever features of EBP that we may identify.

Preferably, a reference corpus should have an attested reliability and represen-tativity as a general purpose corpus. It should also be easily accessible to other researchers to guarantee what Popper called falsifiability — to make it possible for them to test the validity of any statements this author makes based on his analysis conducted on the reference corpus. From these viewpoints, the Brown Corpus and the LOB Corpus have been selected as the primary reference corpora to be used in the current study. These two corpora constitute ideal reference corpora, not only because they are currently considered the "standard" corpora in the field of corpus linguistics yielding many important scholarly papers related to the current study, but also because they have been assigned highly accurate POS (parts-of-speech) tags which is indispensable in conducting the current study.

However, these two general purpose corpora alone are not sufficient to properly measure the degree of significance or idiosyncrasy of a particular linguistic phenomena observed in a genre-specific corpus such as the BLC. For this purpose it is necessary to introduce a third corpus representing another ESP genre, so that any features that we discuss as characteristic of EBP, or of the BLC, can be viewed from a triad perspective as shown in the following diagram (Figure 1-2):

Figure 1-2 Triad relationship between three study corpora



I have therefore added a corpus consisting of various articles taken from the TIME magazine as a third reference corpus. This corpus will be referred to as the TIME Corpus. More details about these three reference corpora are given below.

a) Brown Corpus

The Brown Corpus, formally referred to as the *Standard Corpus of Present-day Edited American English*[9], contains a total of 1,000,000 words (or 1,022,161 to be more precise) of written American English. It consists of 500 text samples of approxi-mately 2,000-word long distributed over 15 text categories. The text data were collected from among the books, newspapers, magazines, official documents and so on published in the U.S. in 1961. The corpus is divided into two sections: infor-mational and imaginative. The former includes such informational prose as news reports, editorials, biographies, essays, scientific papers and so on. The latter part of the corpus consists of imaginative prose such as those taken from various novels, mysteries, science fictions, adventure stories, romance and love stories and so on. For more details about the Brown Corpus, see Kuc&era and Francis (1967) and Francis and Kuc&era (1982).

The Brown Corpus is the first large-scale corpus ever compiled in a machine-readable format, and has become the *de facto* standard for subsequent general purpose linguistic corpora. It took the original compilers of the corpus about four years to complete the first version of the corpus since the work was commenced in 1961. The corpus was subsequently annotated with parts-of-speech (POS) tags, which took them another eight years from 1970 to 1978 to complete, even with the help of the automatic tagging program, TAGGIT, they developed for this purpose.[10]

In the current study, however, it was decided to use a plain text version of the Brown corpus and assign POS tags anew using the Brill Tagger which will be described shortly. This is because the CD-ROM, *ICAME Collection of English Language Corpora*, which I obtained from the International Computer Archive of Modern and Medieval English (ICAME) of the Norwegian Computing Centre for the Humanities did not contain a tagged-version of the corpus for one thing. And for another, there are several different versions of the tagged corpus under different tag sets, which could potentially cause problems in terms of consistency.[11]

b) LOB Corpus

The LOB Corpus, formally the Lancaster-Oslo/Bergen Corpus of British English, was compiled as a British equivalent of the Brown Corpus, containing approxi-mately 1,000,000 words (or 1,015,528 to be more precise) of written British English. As in the Brown Corpus, it consists of 500 text samples of approximately

2,000-word long distributed over 15 categories.[12] The samples were selected from printed sources published in the U.K. in 1961 under the same sampling principles as those applied to the Brown Corpus, so that a direct comparison between the two can be possible.

The compilation of the LOB Corpus started in 1970 at Lancaster University under the direction of Geoffrey Leech and was completed as the first electronic corpus in British English in 1978 in collaboration with Stig Johannson of Oslo University and Knut Hoflan of Norwegian Computing Center for the Humanities, Bergen; hence the name, the Lancaster-Oslo/Bergen (= LOB) Corpus. The first major research with the LOB Corpus, *Frequency Analysis of English Vocabulary and Grammar based on the LOB Corpus,* was published by Johannson and Hoflan in 1989, which corresponds in its contents and academic significance to Francis and Kuc□era (1982).

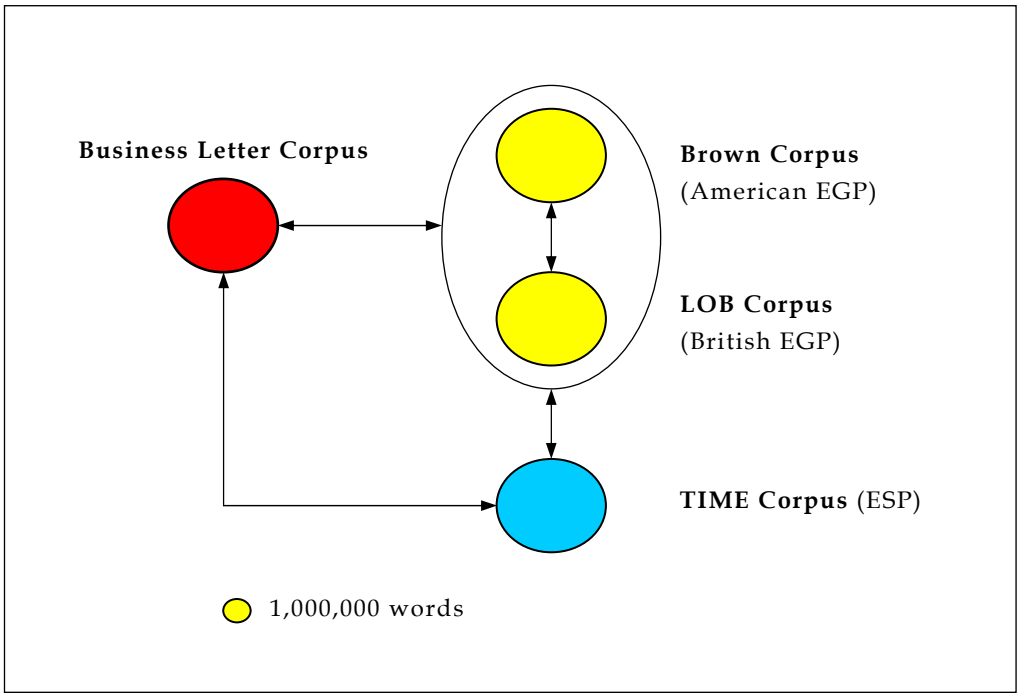
POS tagging of the LOB Corpus was conducted between 1978 and 1983 at Lancaster University using CLAWS, or *C*onstituent *L*ikelihood *A*utomatic *W*ord-tagging *S*ystem (Saito *et al.* 1988, p. 9), which was also used in the POS tagging of the British National Corpus, with a reported "success rate in the region of 96%-97%."[13] In the current study, a POS-tagged version of the corpus provided by the ICAME have been used. Some of the POS tags used in the LOB Corpus, however, have been replaced with corresponding default tags of the Brill Tagger, so that all the research corpora are assigned with the same tag set.[14]

c) TIME Corpus

This corpus was compiled from the text data contained in *1993 TIME Almanac,* a CD-ROM published by Compact Publishing, Inc. This CD-ROM contains all the articles of the TIME magazine published in 1992. Of over two million words of news articles included in this CD-ROM, about 1,000,000 words (or 1,078,972 to be more precise) were extracted to compile the corpus used for this study. The size of this reference corpus, the TIME Corpus, was determined to match the size of the Brown and LOB Corpora.

As mentioned earlier, the rationale for introducing this corpus as one of the three reference corpora is that the lexico-grammatical properties of the BLC data can better be measured by reference to another ESP corpus of a comparable size in addition to the two general purpose corpora, the Brown and LOB Corpora— the former of which representing non-genre specific American English and the latter British English. Using and modifying the diagram in Figure 1-2, the relationships between and among these four corpora can be recapitulated as follows:

Figure 1-3 Relationships between the BLC and the three Reference Corpora



3) Learner Corpus of English Business Letters (Learner BLC)

In addition to the above four corpora, I have also compiled a corpus of English business letters written by Japanese business people (hereinafter referred to as the "Learner BLC"). The rationale for including this non-native corpus of written business English is that, because the ultimate goal of the current study is to produce data-based and data-driven EBP teaching and self-learning materials for Japanese business people that are hopefully more effective and reliable than what are available at the moment, we need to know major problem areas and patterns of errors common to Japanese learners of EBP and that such knowledge can only be obtained through the analysis of business documents actually written by Japanese business people.

The data used for this corpus are the assignment letters submitted by the students enrolled in a correspondence course in business letter writing known as the *Writing Marathon*. This correspondence course started in 1986 and is run by ALC Inc., a publishing company based in Tokyo. I have been involved in this corres-pondence course since the beginning both as an advisor to the publisher and the supervisor of the course.

In this four-month course, the participants are required to write and submit a total of seven assignments [15], which are then corrected for errors, added hand-written comments for possible improvement, and graded for eight different

areas such as format, organization, grammar, tone and style, clarity, fluency and so on, by an expert staff of native speakers of English. Included in the current version of the Learner BLC are a total of 1,484 assignment letters received between May 1997 and July 1998 (The total number of assignments received during this period was 2,367. About 63% of which, however, were excluded because they were not suitable for OCR processing). The total number of words of the Learner Corpus before adjustment is 210,404, which is about one-fourths of that of the Native BLC.

The data obtained from these assignments are not necessarily exhaustive in their coverage, nor may they constitute representative samples of all the possible business correspondence. Nevertheless, they are considered sufficient for the purpose of identifying the possible problem areas and general patterns of errors with regard to the core part of EBP vocabulary. Also, most of the participants of the course are business people needing to learn EBP or to improve on their writing skills. Their writing, therefore, may very well represent the major part, if not in every aspect, of the English writing ability of the entire population of the Japanese business people in need of EBP.

An excerpt from the Learner Corpus is shown in Appendix A3. Note that data correction has been made for OCR reading errors only, and all the spelling and other errors in the original data are left as they are.[16]

1.7 Data-analysis software and computer programs

In order to properly handle and analyze a corpus as large as those mentioned in the previous section, the use of a computer is indispensable. In the first phase of the current study, therefore, I made investigation into the availability of relevant software packages and computer programs. In doing so, I first compiled a list of tasks for which some sort of computer application software is essential, and tried to find commercially available or "free" software packages and computer programs that can perform these tasks to a satisfactory level. The following are some of the major tasks included in the list:

1. Automatic parts-of-speech (POS) tagging.
2. Compilation of a comprehensive wordlist with various statistical information.
3. Automatic lemmatization of wordlist entries.
4. Compilation of a consolidated word frequency comparison table from multiple input files, or subcorpora, of plain running text.
5. Compilation of a consolidated word frequency comparison table from multiple input files, or subcorpora, of plain running text, for specified lexical items (to be specified in the form of a reference wordlist, or a "basefile").

6. Compilation of a frequency wordlist for major POS categories; *i.e.* verbs, adverbs, adjectives, and nouns.
7. Consolidation of multiple wordlists.
8. Automatic calculation of relative importance of individual lexical items, using such statistical indices as "Keyness" and "Usage" (to be defined in Chapter 3)
9. Automatic assignment of "Word Level Tags" (to be defined in Chapter 3) both to wordlist entries and to each token of a plain running text, and compilation of a summary report as to the level of lexical difficulty of the input data.
10. Automatic compilation of a frequency (list and a) comparison table for a specified set of collocations and idioms based on a "leftmost shortest matching" algorithm.
11. Production of KWIC concordance lines for specified lexical items.
12. Regular expression search, or GREPing, of specified text strings.

The investigation was mainly conducted on the Internet. Excluded from the candidates were those that are too expensive to get, unduly difficult or time-consuming to learn to use, or those intended to be used on the UNIX system (in other words, the candidates have to be PC-compatible software running on the MS-DOS or MS-Windows systems). As a result, the following three software packages (including one freeware computer program) were selected among others:

- Brill Tagger
- WordSmith
- TXTANA

The *Brill Tagger* (formally, RULE-BASED TAGGER, Ver. 1.14), is a freeware computer program written by Eric Brill for automatic POS tagging of a plain running text.[17] *WordSmith* (formally, WordSmith Tools), developed by Mike Scott, is considered one of the best linguistic data analysis software currently available on the marker. As Sardinha (1996) called it the "Swiss Army knife of lexical analysis," WordSmith is a multi-function software package capable of such tasks as wordlist compilation, KWIC concordancing, and keywords identification and/or plotting.[18] *TXTANA*, which was developed by Shiro Akasegawa, is a KWIC concordancing software with added, and excellent for that matter, capabilities of collocation analysis and GREPing.[19]

These computer software were found quite useful and, in fact, the current study would not have been possible if it were not for them. Yet, as I mentioned earlier, they are no panacea for all the possible problems. As to the tasks I listed up in the

previous page, only eight out of the 12 tasks[20] can be covered by them—about a half of which only partially—as shown in the following table:

Task No.	Task description	Brill Tagger	WordSmith	TXTANA
1	POS tagging	○		
2	Comprehensive wordlist with statistics		○	
3	Lemmatization		△	
4	W. freq. comp. table for multiple corpora			
5	*Ditto*, with "basefile" option			
6	W. freq. list and comp. table by word class			
7	Merge multiple wordlists		△	
8	Evaluation of "Keyness" and "Usage"		△	
9	Word Level analysis			
10	Collocations freq. comparison table		△	△
11	KWIC concordancing		○	○
12	GREPing		○	○

Task Nos. correspond to the list of tasks on the previous page.
○ = can perform the task to a satisfactory level.
△ = can perform the task only partially, or with a substantial amount of additional work.

Table 1-2 Required tasks and their performability with the three existing software packages chosen for the current study

For what they are not capable of, therefore, I had to write computer programs by myself using AWK, a computer programming language originally developed by Alfred V. Aho, Peter J. Weinberger, and Brian W. Kerninghan. AWK is an interpreter language, whose grammar structure is similar to that of the C language, and is particularly suited for processing text data. The version of AWK used in the current study is JGAWK (Japanaized Gnu *AWK* for MS-DOS) Ver. 2.11.1 + 3.0.[21] For more information about AWK and JGAWK, see Aho *et al.* (1988), Close and Robbins *et al*. (1995), Uemura (1993), Ito (1991 and 1992) and Ueda (1998a and 1998b). For a reference purpose, I have included an abridged list of AWK programs which I wrote for the current study as Appendix C1.

Endnotes to Chapter 1

All the Internet addresses (URLs) quoted in this paper are at the time of writing and may have since been changed.

[1] This survey was conducted between August and October 1998 to collect basic data as to the current state of written business communication in English in the Japanese workplace. A total of 140 Japanese business people responded to a two-page questionnaire containing 13 questions. For more details, see Someya (1999a).

[2] For more details of these computer programs, see Appendix C1.

[3] To be discussed in more detail in Chapter 3 (See Endnote 3 of Chapter 3).

[4] The chart is based on Robinson (1991). He, however, uses the term EOP (English for Occupational Purposes) instead of EBP. Also, he divides EAP into two separate categories: EAP and EEP (English for Educational Purposes). EST is not referred to at all in his classification.

[5] Quotation cited in Masuyama (1958, p. 120)

[6] According to Stabbs, one of the most common usage patterns of performative verbs is in the form of "modal plus lexical illocutionary verb" which he calls "hedged performative" (*op. cit.*, p. 214), as shown in the following samples:

- I *would advise* you that . . .
- I think we *should decline* your offer.
- I *would like to extend* to you our thanks for . . .

If what he says is correct, it is expected that the number of modals will increase in proportion to the number of lexical illocutionary verbs. This assumption, as we will see in more detail in later Chapters of this paper, has been in part confirmed valid in case of written business discourse.

[7] Most of Nakamura's papers (in Japanese) are available as online papers via his personal Internet Website at the following URL: http://www.heian.ac.jp/col/KS/nakamura/

[8] For more details about the data sources of the BLC, see Appendix A1.

[9] The compilers of the original Brown Corpus, Nelson Francis and Henry Kuc&era were the staff members of Brown University at the time of corpus compilation; hence the name, the Brown Corpus.

[10] A manual of the POS-tagged version of the Brown Corpus (= *Brown Corpus Manual: Manual of Information to accompany A Standard Corpus of Present-day Edited American English Corpus, for use with Digital Computer, 1979)* is available from the Internet Website of the ICAM. [Online] http://www.hd.uib.no.icame/bcm.html#tc

[11] For a complete list of POS tags used in the Brown Corpus (Ver. 1979), see Appendix B3.

[12] As to a critical discussion about the appropriateness of the 15 text categories (nine in the Informative Section and six in the Imaginative Section), see Saito *et al.* (1998, pp. 39-40).

[13] For more details about the CLAWS system, see the following Internet Website (URL): http://tina.lancs.ac.uk/computing/research/ucrel/claws (*or*, http://tina.lancs.ac.uk/crel/annotation.html#POS). In addition to the CLAWS system, there is another POS tagger used to annotate the LOB Corpus, known as the *TOSCA/LOB Tagger* (for MS-DOS) which was developed by the TOSCA Research Group of the Department of Language and Speech, University of Nijmegen. The TOSCA Tagger has many different versions including *TOSCA Tagger for ICE, TOSCA Tagger-Lemmatizer,* and *TOSCA Tree Editor.* For more details, consult TOSCA Homepage at the following URL: http://iris1.let.kun.nl/TSpublic/tosca

[14] For a complete list of POS tags used in the LOB Corpus (Ver. 1986), see Appendix B2.

[15] The following is a very simplified outline of the seven assignments (For Assignments 1, 2, 4, 6 and 7, students are given two options from which to choose one):

Assignment No.	Option No.	Description
1	1-1	Writing a letter of request, asking for information (L)
	1-2	Writing a letter of request, asking for free materials and offering to pay expenses (L)
2	2-1	Writing a letter of general inquiry, with itemized questions (F)
	2-2	Writing a letter of inquiry involving price negotiation (F)
3	--	Writing a polite reminder, requesting overdue payment (L)
4	4-1	Writing a thank-you letter after a visit (F)
	4-2	Writing a formal letter of recommendation (L)
5	--	Writing a letter declining a request and proposing an alternative (F)
6	6-1	Writing a follow-up memo requesting overdue reports (M/EM)
	6-2	Writing a letter asking for an appointment. (F)
7	7-1	Writing a short report at the completion of the course, summarizing the contents of the course. (R)
	7-2	Writing to the teaching staff, thanking them for their help and making a suggestion for possible improvement of the course. (L)

L = Letter, F = Fax, M/EM = Interoffice memorandum or e-mail, R = Report

[16] This, however, causes a problem with regard to the accuracy of word frequency count. To solve this problem, wordlist compilation and frequency count for the Learner BLC has been conducted on a separate file with all the detectable spelling errors corrected by an automatic spelling checker.

[17] For further information about the Brill Tagger, see Brill (1993a), Ph.D. Dissertation by Eric Brill describing theoretical backgrounds of the tagger, and Brill (1993b) which contains a series of "Instruction Manuals." The Brill Tagger can be downloaded from

Eric Brill's Internet Website (See bibliography for URL information). Someya (1997b) describes how to download and unzip the compressed tar-file, as well as the procedure for converting the program for use on the MS-DOS platform. For an overview of past and current research into automatic POS tagging, see Lager (1995).

[18] For more details about WordSmith, refer to Mike Scott's Internet Website at the following URL: http://www.liv.ac.uk/~ms2928/index.htm

[19] For more details about TXTANA, see Okada (1998) and Someya (1999c). A trial version of TXTANA can be downloaded from the following URL:
http://www.biwa.or.jp/~aka-san /index.html

[20] What these tasks (and the rather abstract and much too short descriptions thereof) mean in a more concrete sense will be made clear in later chapters.

[21] The reason why I used AWK (JGAWK) among other programming languages is that I had an opportunity to learn it in one of the graduate courses offered at the University of Tokyo. For this, I thank Professor Hiroto Ueda for his well-organized instruction. The course handouts used in the class of 1997 have now been published as Ueda (1998a) and Ueda (1998b), both of which include JGAWK and many useful AWK scripts in a floppy disk attached thereto. JGAWK can also be downloaded from the following URL:
http:// www.vector.co.jp/vpack/browse/person/an000012.html