THE UNIVERSITY OF TOKYO

A Corpus-based Study of Lexical and Grammatical Features of Written Business English

(Vol. 1/2)

AN

M.A. THESIS

SUBMITTED TO THE GRADUATE DEPARTMENT OF LANGUAGE AND INFORMATION SCIENCES OF THE UNIVERSITY OF TOKYO

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF ARTS

BY

YASUMASA SOMEYA

Chief Supervisor: Professor Hideo Suzuki

Second Supervisors: Professors Hideo Oka and Kazuhiko Matsuno

1999

ABSTRACT

A Corpus-based Study of Lexical and Grammatical Features of Written Business English

Yasumasa Someya

1999

The primary purpose of the study is to identify and describe some of the major lexico-grammatical features of English for Business Purposes (EBP) in its *written* domain. The author also aims at establishing a computer-based study methodology appropriate for the analysis of such large-scale corpus data as those used in this study, including the development of a series of computer programs with which to analyze the corpus data from various viewpoints and for specific purposes.

The study revealed, among other things, that EBP is characterized by a high degree of lexical closure. The overall lexical growth curve of the one-million-word Business Letter Corpus (BLC) compiled for this study reaches its plateau at about the 15,000-20,000 words range. When lemmatized, 90.8% of all the word tokens were covered only by 1,500 word types, showing a marked contrast to the three reference corpora used in the study. It was further found that the first 350 verb-types in the BLC Wordlist cover about 90.8% of all verb occurrences in the BLC, and only 100 adverbs—including all the major "metadiscourse" items—cover 89.4% of all that are needed to construct cohesive and well-expressed messages. The same also holds true, though with differing degrees and realization patterns, in other POS categories. The tendency of high lexical closure is more evident with the Learner BLC consisting of English business messages written by Japanese business people, where only 109 verb-types account for 90% of all the verb occurrences, and 43 adverbs account for 90.2% of all the instances of adverbs appearing in this corpus.

The study also substantiated our second claim that EBP is characterized by a low level of lexical difficulty. It was found that approximately 77.5% of the first 3,000 word-types in the BLC, which cover over 95% of all the BLC word tokens, are within the *4000 Basic Words* as defined by the JACET (Japan Association of College Teachers), plus a fairly standard set of proper nouns, abbreviations and acronyms.

It also produced ample evidence in strong support of our third hypothesis that written business English is characterized by its incorporation of spoken features into written texts.

Another interesting finding is that many of the high frequency lexical items, which are ranked as per their relative importance defined statistically, coincide with the items that the Japanese users of EBP are prone to make errors in. The mean error ratio of the most important 50 "key" verbs, for instance, was found to be as high as 15.69% with SD=12.54. The error analysis conducted on the Learner BLC further revealed that many of these errors show clear systemicity—in other words, errors occur where they tend to occur and for good reason. With regard to verbs, for instance, errors are more likely to occur when a particular verb in English has its apparent semantic counterpart in Japanese but appears in a different argument structure and under different semantic constraints. Representative cases in point are the verbs *discuss* and *require* whose error ratios are as high as 67.69% and 20.51% respectively—ratios that suggest something is going very wrong with the way these lexical items are taught in the classroom. This finding naturally leads to the proposal that the concepts of "argument structure" and "semantic constraints" be included in the teaching syllabus and students be given appropriate instructions thereof.

It is also noteworthy that many of the "core" lexical items are closely associated with particular syntactic patterns. For instance, the adjective *important* appears 367 times in the BLC, of which 74 cases (20.16%) occurred in the "*It is ADJ (for NP) to VB*", "*It is ADJ to NP*" or "*It is ADJ that-clause (or ZERO-that)* formats. This and other similar instances that abound in the BLC indicate the importance of teaching lexical items not in isolation but in reference to the syntactic environments in which they typically appear.

The discussions are not exhaustive by any means, but enough to substantiate the claim that Business English is a "sublanguage" with its unique lexico-grammatical patterns and that the identification and description of which in a systematic way will help the learners of EBP to learn what need to be learned more effectively than it used to be. It is the author's hope that the data presented in this paper and the various findings thereof will provide the teachers of EBP with a solid, data-driven foundation for their classroom instructions and for writing course materials of their own. The author also believes that the various computer programs written by him for the current study will prove useful for those interested in computational analysis of large-scale corpus data. The programs have been written with JGAWK (Ver. 2.11.1+ 3.0) and will be made available in a ready-to-run format for interested researchers at the completion of the current study, so that they can be tested, modified or otherwise used at the user's disposal.

ACKNOWLEDGEMENTS

Many people deserve thanks in helping me complete this MA thesis. Instead of listing names after names, however, I will just say one big thanks to all the people who deserve to be thanked but not mentioned here by name.

My advisor, Professor Hideo Suzuki, deserves my heartfelt appreciation for the warm and constant support he has given me in this seemingly endless endeavor. I also appreciate Professors Hideo Oka and Kazuhiko Matsuno for their critical feedback, which, in retrospect, had a significant impact in shaping my thesis.

My thanks also goes to Professor Hideto Ueda who has given me the opportunity of learning the programming language, AWK. If it were not for his well-organized instruction in one of the graduate courses I took two years ago, I would still be counting the number of words in my one-million-word corpus.

Most of all, I thank my wife, Kiyoko, for her unending support, dedication and, above all, the many sacrifices she made for me. To her, this thesis is dedicated.

Table of Contents

ABSTRACT . . . i

ACKNOWLEDGEMENTS . . . iii

List of Figures . . . vii

List of Tables . . . viii

Chapter 1 Introduction . . . 1

1.1 Background . . . 1

1.2 Purposes and scope of the study . . . 2

1.3 Working hypotheses . . . 3

1) The “lexical closure” hypothesis . . . 4

2) The “lexical ease” hypothesis . . . 4

3) The “write-as-you-talk” hypothesis . . . 5

1.4 Definition of “Business English” . . . 5

1.5 Review of previous research . . . 7

1) Overseas research related to EBP vocabulary and usage . . . 7

2) Japanese research related to EBP vocabulary and usage . . . 9

1.6 Corpora used in the study . . . 11

1) Business Letter Corpus (BLC) . . . 11

2) Reference corpora . . . 12

3) Learner Corpus of English Business Letters (Learner BLC) . . . 17

1.7 Data-analysis software and computer programs . . . 18

Chapter 2 Comparative Analysis of POS Distribution . . . 24

2.1 The main rationale for POS distribution analysis . . . 24

2.2 Study procedure . . . 25

2.3 Relative proportions of the major POS categories in the BLC and the Reference Corpora . . . 27

2.4 Discussions . . . 29

1) Modals . . . 31

2) Conditional *if* . . . 32

3) Pronouns . . . 33

4) Infinitival *to* . . . 33

5) Determiners . . . 35

6) WH-particles . . . 36

7) Existential *there* . . . 37

8) Interjections . . . 42

2.5 Comparison between the Native and Learner BLCs . . . 43

1) Major areas of difference and their implications . . . 44
a) Nouns and Determiners . . . 44
b) Adjectives . . . 46
c) Modals and Conditional *if* . . . *47*
2.6 Summary . . . 47

Chapter 3 The BLC Wordlists . . . 55
3.1 Definitions of technical terms . . . 55
3.2 The BLC General Wordlists . . . 59
1) THE COMPREHENSIVE BLC WORDLIST . . . 59
2) THE BLC KEYWORDS LIST . . . 64
K *See Volume 2 (Appendices D1 and D2) for the complete wordlists*
3.3 The BLC Categorical Wordlists . . . 68
1) THE BLC VERB LIST . . . 68
2) BLC wordlists for adverbs, adjectives and nouns . . . 75
K *See Volume 2 (Appendices E1, E4, E6 and E8) for the complete wordlists*
3.4 Summary . . . 77

Chapter 4 The BLC Lexicon . . . 87
4.1 Verifying the three hypotheses against empirical evidence . . . 87
1) The "lexical closure" hypothesis . . . 87
2) The "lexical ease" hypothesis . . . 93
3) Some lexical evidence in support of the "write-as-you-talk" hypothesis . . . 97
a) Personal pronouns . . . 99
b) Articles . . . 103
c) Preposition "*of*" . . . 106
d) Contractions . . . 107
e) Negative marker -'*nt*/-'*t* . . . 109
f) WH-relativization . . . 110
4.2 Modals as the "key" keywords of the BLC lexicon . . . 116
1) Distribution and frequencies of modals in the BLC . . . 117
2) Syntactic structures of modal verb-phrases . . . 119
a) Study procedure . . . 120
b) Results and discussion – identifying the most frequent patterns of modal verb-phrase structures in business discourse . . . 121
3) Conditional *if* and the types of conditional sentences . . . 125
a) Verb forms and their combinations in *if*-conditional sentences . . . 125

b) Clausal order of the *if*-clause and the main clause . . . 128
4.3 The BLC "Core" Vocabulary . . . 131
1) Defining the core verbs . . . 131
a) BLC core verbs sorted by the *Usage* scores . . . 133
b) BLC core verbs sorted by the *Keyness* scores . . . 136
c) Preliminary error analysis of the Learner BLC for the core verbs . . . 139
2) Defining the core adverbs, adjectives and nouns . . . 141
4.4 Summary . . . 150

Chapter 5 Conclusion . . . 164
5.1 Summary of the thesis . . . 164
5.2 Attainment of research purposes . . . 168
5.3 Contributions of the current study . . . 169
5.4 Further research . . . 169

Appendix A1: A detailed list of data sources of the Business Letter Corpus . . . 171
Appendix A2: A sample excerpt from the Business Letter Corpus . . . 177
Appendix A3: A sample excerpt from the Learner BLC . . . 178
Appendix B1: POS tag set used with the Brill Tagger . . . 179
Appendix B2: LOB Corpus POS tag set . . . 181
Appendix B3: Brown Corpus POS tag set . . . 184
Appendix C1: A List of AWK programs used in the study . . . 187
Appendix C2: Program source of the wordlist compiler, `mk_list.awk` . . . 190
Appendix C3: Sample Image of "Lexical Profiling" for Business Core Verbs . . . 195
Appendix C4-1 (Table 4-40): Business English Core Adverbs (by Usage). . . 198
Appendix C4-2 (Table 4-41): Business English Core Adverbs (by Keyness) . . . 200
Appendix C4-3 (Table 4-42): Business English Core Adjectives (by Usage) . . . 201
Appendix C4-4 (Table 4-43): Business English Core Adjectives (by Keyness) . . . 203
Appendix C4-5 (Table 4-44): Business English Core Nouns (by Usage) . . . 205

Appendices D1-D3 (General BLC Wordlists), E1, E4, E6, E8 (Categorical BLC Wordlists), and F1-F2 (Learner BLC Wordlists) are provided in Volume 2 of this paper.

Bibliography . . . 207

List of Figures

Figure 1-1 ESP and its major subcategories . . . 6

Figure 1-2 Triad relationship between three study corpora . . . 14

Figure 1-3 Relationships between the BLC and the three Reference Corpora . . . 16

Figure 2-1 LOB tags conversion table (excerpt) . . . 50

Figure 2-2 `prn_tag.awk` (for POS tags extraction) . . . 50

Figure 2-3 Sample output of POS tag frequency count by WordSmith (excerpt) . . . 27

Figure 2-4 Comparative distribution of major POS categories . . . 30

Figure 2-5 Comparative POS distributions between the Native and Learner BLCs . . . 45

Figure 3-1 Sample output of WordSmith Wordlist for the BLC (excerpt) . . . 83

Figure 3-2 `matchnew.awk` (for replacing entries of a wordlist with word-level tags) . . . 84

Figure 3-3 Business Letter Corpus Comprehensive Wordlist (MS Excel screen shot) . . . 63

Figure 3-4 Sample output of WordSmith Keywords List, comparing the BLC and the combined Reference Corpus (BROWN+LOB+TIME) . . . 65

Figure 3-5 BLC Keywords List screen shot (Last page: Keyness Ranks 5352-5412) . . . 67

Figure 3-6 `vb.awk` (for extracting verbs from a POS-tagged wordlist) . . . 85

Figure 3-7 A flow chart showing steps to produce the Lemmatized BLC Verb List 1 . . . 69

Figure 3-8 A sample output of the consolidated word frequency comparison table for the BLC subcorpora, produced via `mk_list.awk` (excerpt) . . . 71

Figure 3-9 Business Letter Corpus Verb List 2 (MS Excel Screen Shot) . . . 72

Figure 3-10 Business Letter Corpus Verb List 1 (MS Excel Screen Shot) . . . 74

Figure 3-11 Business Letter Corpus Adverb List 1 (MS Excel Screen Shot) . . . 78

Figure 3-12 Business Letter Corpus Adjective List 1 (MS Excel Screen Shot) . . . 79

Figure 3-13 Business Letter Corpus Noun List 1 (MS Excel Screen Shot) . . . 80

Figure 4-1 Comparison of lexical growth curves . . . 89

Figure 4-2 The numbers of lemmatized word types and their cumulative percentages to the total word tokens of the BLC . . . 91

Figure 4-3 Word level distribution of the first 1500 word-types of the BLC Wordlist . . . 94

Figure 4-4 Word level distribution of the first 1500 word-types of the BLC Wordlist . . . 94

Figure 4-5 Word level distribution of the BLC lexicon up to the 6408th entry (Freq.≧ 5) . . . 95

Figure 4-6 Sample MS-DOS output of *Word Level Checker* (Ver. 1) for the first 1500 word-types of the Learner BLC . . . 97
Figure 4-7 Sample excerpt from the modal verb-phrase dictionary (MD.DIC) . . . 156
Figure 4-8 MD_match.awk (for identifying specified MD verb structures, with "leftmost shortest matching" algorithm) . . . 157
Figure 4-9 prn_MD.awk (for printing the strings marked by MD_match.awk) . . . 158
Figure 4-10 cnt_freq.awk (for frequency counting of the data file created via MD_match.awk and prn_MD.awk) . . . 158
Figure 4-11 Corpus-wise comparison of the numbers of verb types and percentages to the total verb tokens . . . 132
Figure 4-12 Sample KWIC concordance with error tags . . . 162
Figure 4-13 Lexical growth curves of adverbs, adjectives and nouns in the BLC . . . 142

List of Tables

Table 1-1 List of subcorpora included in the Original Version of the Business Letter Corpus (BLC-1) . . . 13
Table 1-2 Required tasks and their performability with the three existing software packages chosen for the current study . . . 19
Table 2-1 Relative proportions of major POS categories in the BLC and Reference Corpora . . . 28
Table 2-2 Comparative distributions of major POS categories (Unit =%) . . . 30
Table 2-3 Major syntactic patterns of *hope* and their frequencies in the BLC . . . 34
Table 2-4 Major syntactic patterns of *decide* and their frequencies in the BLC . . . 34
Table 2-5 Major syntactic patterns of *Ex-there* construction in the BLC and their frequencies . . . 39
Table 2-6 Most frequent determiners and qualifiers co-occurring with there within three words to the right (N>5) . . . 40
Table 2-7 Frequencies of INTERJECTIONS in the BLC and the Reference Corpora . . . 42
Table 2-8 Comparative POS distributions between the Native and Learner BLCs . . . 45
Table 2-9 Comparative data for ADJECTIVES in the Native and Learner BLCs . . . 46
Table 3-1 Number of words contained in the WL-tag dictionary (wrdlvl-2.dic) . . . 58
Table 3-2 Comparison of the numbers and percentages of word types and tokens at different frequencies in the COMPREHENSIVE BLC WORDLIST . . . 61
Table 4-1 Type-Token comparison of the BLC and the Reference Corpora before lemmatization . . . 89

Table 4-2 The numbers of lemmatized word types and their cumulative percentages to the total word tokens . . . 91
Table 4-3 Text coverage and approximate vocabulary size . . . 92
Table 4-4 Word level distribution of the first 1500 and 3000 word-types of the BLC Wordlist . . .94
Table 4-5 Word level distribution of the first 800 word-types of the Learner BLC . . . 96
Table 4-6 The first 20 most "businesslike" verbs, adverbs, adjectives, and noun in the BLC Keywords List . . . 98
Table 4-7 Keyness scores of personal pronouns . . . 100
Table 4-8 Comparative frequencies of the major personal pronouns in the Learner and Native BLCs . . . 102
Table 4-9 Keyness scores of the definite and indefinite articles . . . 103
Table 4-10 Comparison of the numbers of "Det|ϕ (Adj) N PP" sequences . . . 104
Table 4-11 Keyness scores of prepositions *of, by* and *in* . . . 106
Table 4-12 Comparative proportions of the major forms of contraction in the BLC and the Brown Corpus . . . 108
Table 4-13 Comparative frequencies of the negative marker -*'nt/-'t* in the BLC, Brown and TIME Corpora . . . 109
Table 4-14 Comparative frequencies and keyness scores of major WH-words . . . 111
Table 4-15 Comparative frequency ranking of major WH-words . . . 111
Table 4-16 Comparative frequencies of the relative pronoun *which* . . . 112
Table 4-17 Comparative normalized frequencies of "pied-piping" *which* relatives . . . 115
Table 4-18 Keyness scores of major modals . . . 117
Table 4-19 Relative proportions of major modals in the BLC and the average number of words and sentences per occurrence . . . 118
Table 4-20 Modal structures and their distribution in the BLC . . . 122
Table 4-21 Use of modals in the "MD+be+VBG" structure in the BLC . . . 123
Table 4-22 Use of modals in the "MD+have+VBN" structure in the BLC . . . 124
Table 4-30 Verb form combinations in conditional sentences with *if*-clause . . . 126
Table 4-31 Comparative frequencies of "*were*" subjunctive and hypothetical "*was*" in the BLC . . . 128
Table 4-32 Frequencies of the two clausal orders of *if*-conditional sentences in the five study corpora . . . 129
Table 4-33 Corpus-wise comparison of the numbers of verb types and percentages to the total verb tokens . . . 132
Table 4-34 Business English core verbs (first 350 entries in order of Usage ranking) . . . 134-136
Table 4-35 Business English core verbs sorted by K-score (first 100 entries in

Keyness order) . . . 137
Table 4-36 Word level distribution of the first 350 core verbs . . . 138
Table 4-37 Classification of verb error types . . . 139
Table 4-38 Learner BLC error ratios for 50 “core” verbs chosen at random . . . 140
Table 4-39 Numbers of word types and percentages to the total word tokens of adverbs, adjectives and nouns in the BLC . . . 142
Table 4-40 (Appendix C4-1): Business English Core Adverbs (by Usage). . . 198
Table 4-41 (Appendix C4-2): Business English Core Adverbs (by Keyness) . . . 200
Table 4-42 (Appendix C4-3): Business English Core Adjectives (by Usage) . . . 201
Table 4-43 (Appendix C4-4): Business English Core Adjectives (by Keyness) . . . 203
Table 4-44 (Appendix C4-5): Business English Core Nouns (by Usage) . . . 205
Table 4-45 Functional classification of metadiscourse . . . 144
Table 4-46 Typical complementation patterns of some of the core adjectives and their frequencies in the BLC . . . 146
Table 4-47 Some of the core nouns and their acceptability of TO-infinitive and THAT-clause as a complement . . . 149