CONTENTS

List of Appendices v
List of Tables vi
List of Figures vii
List of Abbreviations viii

1 Language variety and lexical structure 1
1.1 The structured nature of language 1
1.2 Word frequency counts of French, general and specialised 4
1.3 Preliminary presentation of the word frequency lists for French business correspondence 7

2 Composition of the Corpus 9
2.1 General description of the corpus 9
2.2 Detailed description of the corpus 10
2.3 Branches of industry 15
2.4 Assessment of the effect of large sets of letters from single senders 17
2.5 Pre-editing the letters included in the corpus 18

3 Preliminary considerations regarding the status of items forming frequency lists 22
3.1 Lemmatized and unlemmatized frequency lists 22
3.2 Macrostructure and microstructure 22
3.3 Importance of an unambiguous macrostructure for frequency counts 23
3.4 Weaknesses in the macrostructures of existing frequency lists 24
3.5 General criteria for lemmatizing the FBC corpus 27

4 Segmentation 29
4.1 Character strings and orthographic words 29
4.2 Strings corresponding to less than one complete orthographic word 29
4.3 Strings corresponding to two or more orthographic words separated by spaces 31
4.4 Strings corresponding to two or more orthographic words connected by hyphens and apostrophes 38
4.5 Postscript regarding segmentation 40
4.6 Overall conclusion regarding segmentation 45

5 Lemmatization 47
5.1 Lexical homography 47
5.2 Word-class assignment and lemmatization 49
5.3 The verb 52
5.4 The uninflected words 58
5.5 Pronoun, noun and adjective - differentiation 60
5.6 Pronoun, noun and adjective - lemmatization 63
5.7 Miscellaneous items 66
5.8 The output from the word-class assignment and lemmatization procedure 66
5.9 *The procedure in action* 67
5.10 Conclusion 68

6 The structure of the vocabulary 69
6.1 Vocabulary and lexicon 69
6.2 The frequency distribution table 70
6.3 The Waring-Herdan Distribution 73
6.4 Vocabulary richness 80

7 The growth of the vocabulary and the size of the lexicon 83
7.1 The growth of V 83
7.2 The growth of V_1 87
7.3 Estimating the size of $\mathcal{L}$ 89
7.4 The nature of $\mathcal{V}_0$ 90

8 The predictive power of F 96
8.1 Confidence intervals for F 96
8.2 Overall evenness of distribution 98

9 Dispersion 101
9.1 Measures of dispersion 101
9.2 Comparing the performance of D, D_2 and S 104
9.3 Weighing the criticisms levelled at D 114
9.4 Conclusion regarding the relative merits of D, D_2 and S 117
9.5 A postscript regarding chi-square 117
9.6 Evenly distributed versus unevenly distributed items 119
9.7 The relationship between evenness of distribution and the 'commercial' versus 'technical' item dichotomy 122

10 Usage coefficients 125
10.1 Juilland's usage coefficient, 'U' 125
10.2 Proposals for usage coefficients with a bottom limit greater than zero 126
10.3 Differences in ranking by usage attributable to the use of D, D_2 and S 128
10.4 Conclusion regarding the relative merits of U, U_m and KF 129

11 The effectiveness of simple Range vis-à-vis Juilland's D 131

12 The limits of usefulness of Range and Modified Frequency as a corrective to Raw Frequency in the FBC corpus 145
12.1 R and lemmatization 145
12.2 R as an aid in identifying the 'core' vocabulary of FBC 145
12.3 Modified Frequency 148

13 Range as a basis of comparison across language varieties 150
13.1 Prerequisites for comparing corpora on the basis of R 150
13.2 A comparison of FBC and Engwall 151
13.3 The characteristically evenly distributed items of FBC 153

14 Frequency in a comparative perspective - Registral Value 155
14.1 In search of the characteristic vocabulary of FBC 155
14.2 Preliminary adjustments to the macrostructures of FBC and JUI 156
14.3 The 100 most frequent items of FBC and of JUI 157
14.4 Measuring overall similarity and difference between vocabularies 161
14.5 Registral Value 164
14.6 Modified Registral Value 171
14.7 In search of the characteristic vocabulary of FBC - conclusion 174

15 **Description of the characteristic vocabulary of FBC** 175

15.1 A conceptual framework for the qualitative description of modified PRI's 175

15.2 The textual function 176

15.3 The interpersonal function 181

15.4 The ideational function 195

15.5 Conclusion 214

16 **Overview** 215

16.1 Theoretical and methodological issues 215

16.2 Description of FBC 219

16.3 Some future perspectives 221

Bibliography 225

Appendices

A The alphabetically-ordered list 231

B The lists ranked by Frequency and Registral Value 255

C Various supplementary lists 275

D Frequency distribution table (text coverage) 277

E Goodness of fit of Waring-Herdan Distribution to FBC 281

F 'Technical' and ' Commercial' items in the sample 'Tenth' 286

G Subfrequencies (per 'Fifth') and various dispersion measures 287

H Some thoughts on developing a frequency-independent measure of dispersion based on chi-square 293

I Subfrequencies (per 'Tenth') with Juilland's D and simple Range 296

J The 36 'Technical' items among the first 1020 ranks of the list by Modified Frequency 301

K Items instantiated in all ten 'Tenths' of the FBC corpus (R=10) 302

L Index to Chapter 15: items highly characteristic of FBC 303

TABLES

1 Sample paths through the word-class assignment and lemmatization procedure 67

2 Text coverage 71

3 Goodness of fit between Waring-Herdan Distribution and actual frequency distributions for four versions of FBC count 74

4 Goodness of fit between W-HD and actual frequency distributions for various French texts 76

5 Goodness of fit between W-HD and actual frequency distributions for 3 Giraudoux novels and FBC 77

6 Goodness of fit between W-HD and actual frequency distributions for sections of FBC corpus of varied sizes 79

7 Vocabulary richness in FBC and in Engwall's corpus of modern French novels 82

8 Cumulative sub-totals for 'geog.orthog.' words alone in the shuffled Tenths 86

9 Cumulative sub-totals for orthographic words, excluding all proper nouns, in the shuffled Tenths 86

10 Growth of V_1 87

11 New F 1-4 items in Sample Tenth 91

12 'Commercial' hapax items in the Sample Tenth 92

13 'Commercial' items with F=2 to 4 in the Sample Tenth 94

14 Subfrequencies of items with F=39 97

15 Correlations between D, D_2 and S 104

16 Effect on dispersion measures of F not being a multiple of n 115

17 Correlation of values of D, D_2 and S with F for equiprobable distributions 119

18 D values of all FBC items having $F \geq 57$ and Range < 10 121

19 The link between the EDI/UDI distinction and that between 'commercial' and 'technical' items 123

20 'Lexical connexion' between adjacent Tenths 132

21 Hypothetical values for D(10) and R assuming a perfect linear functional relation 134

22 Range as an indicator of 'commercial' or 'technical' status 148

23 Characteristically evenly distributed items (CED) 154

24 The 100 most frequent items in FBC and in JUI 158

25 Vocabulary correlation FBC $\sim$ JUI 162

26 Items with positive R-val. seen in relation to frequency band 168

27 Significant PRI's seen in relation to frequency band 169

28 The 100 items ranked highest by Modified Registral Value 172

29 'Politeness' items 188

30 The most frequent closing formulas 190

31 Favourite segments filling various slots in the closing formula pattern 191

32 Frequencies of the favourite pattern closing formulas 193

FIGURES

1 Growth of V, V_1 and V_2. Orthographic words (including 'geographical' ones). Natural order of Tenths 85
2 Growth of V_1 and V_2. Orthographic words (excluding all proper nouns). Shuffled order of Tenths 88
3 D_2 compared with D for F=10. All possible distributions 106
4 D compared with D_2 for F=10. All possible distributions 108
5 S compared with D for F=10. All possible distributions 110
6 D_2 compared with D for F=10. 46 actual distributions (scatter diagram) 111
7 S compared with D for 38 F 41. 30 actual distributions (scatter diagram) 112
8 S compared with D for F=5. 30 actual distributions (scatter diagram) 113
9 *Values of D for 30-item samples drawn from five frequency bands* 118
10 Values of χ^2 for 30-item samples drawn from five frequency bands 120
11 Hypothetical values for D(10) and R assuming a perfect linear functional relation 135
12 *Correlation between D(10) and R for the 150 items listed in* Appendix I(i-iv) 136
13 Correlation between D(10) and R for the 30 items listed in Appendix I(i) ($72 \geq F \geq 57$) 137
14 Correlation between D(10) and R for the 30 items listed in Appendix I(ii) ($41 \geq F \geq 38$) 138
15 Correlation between D(10) and R for the 30 items listed in Appendix I(iii) ($18 \geq F \geq 17$) 139
16 Correlation between D(10) and R for the 30 items listed in Appendix I(iv) (F=10) 140
17 Correlation between D(10) and R for the 30 items listed in Appendix I(v) (F=5) 141
18 $V_{R=1}$, $V_{R=2}$, ..., $V_{R=10}$ in FBC and in Enwall's corpus of modern French novels 152
19 Distribution of Registral Values 167
20 Closing formula patterns 191