

---

# CONTENTS

<b>Foreword</b>	v
<b>INTRODUCTION</b>	1
<b>1 TEXTUAL STATISTICS: SCOPE AND APPLICATIONS</b>	5
<b>1.1 APPROACHES TO PROCESSING TEXTS</b>	5
1.1.1 The linguistic viewpoint	6
1.1.2 Content analysis	7
1.1.3 Artificial intelligence	8
<b>1.2 STATISTICS AND TEXTS</b>	8
1.2.1 Pioneering works	9
1.2.2 Information retrieval	9
<b>1.3 STATISTICAL PROCESSING OF TEXTS:         A STATISTICIAN'S VIEWPOINT</b>	10
1.3.1 Processing phases	10
1.3.2 Internal and external information, meta-data	11
1.3.3 A wealth of meta-information	12
<b>1.4 SPECIAL TEXTS: RESPONSES TO OPEN QUESTIONS</b>	14
1.4.1 Open-ended questions: a research tool	15
1.4.2 Manual <i>post-coding</i> of free responses	17
1.4.3 Artificial texts: groups of responses	18
<b>2 THE UNITS OF TEXTUAL STATISTICS</b>	21
<b>2.1 CHOOSING UNITS</b>	21
2.1.1 Computerized text	22
2.1.2 Analyses based on graphical forms	22
2.1.3 Lemmatized analyses	23
2.1.4 Semantically based approaches	24
2.1.5 Brief comparison with other languages	25

<b>2.2 SEGMENTATION AND NUMERIC CODING OF TEXT</b>	26
2.2.1 Numeric coding of <i>Life</i> corpus	27
2.2.2 Corpus P	28
<b>2.3 QUANTITATIVE ANALYSIS OF VOCABULARY</b>	28
2.3.1 Frequencies, frequency distribution	28
2.3.2 Zipf's law	29
<b>2.4 LEXICOMETRIC DOCUMENTS</b>	31
2.4.1 Index of a corpus	32
2.4.2 Contexts, concordances	32
2.4.3 Vocabulary growth	34
2.4.4 Lexical tables	35
<b>2.5 REPEATED SEGMENTS</b>	35
2.5.1 Sentences, sequences	36
2.5.2 Repeated segments table	37
<b>2.6 FINDING CO-OCCURRENCES, QUASI-SEGMENTS</b>	39
2.6.1 Searching around a pivotal word	39
2.6.2 Finding multiple co-occurrences, quasi-segments	40
<b>2.7 DEALING WITH TAGGED CORPORA</b>	40
2.7.1 Appending metadata to free responses in surveys	40
2.7.2 Comparison of main quantitative characteristics	41
<b>3 CORRESPONDENCE ANALYSIS OF LEXICAL TABLES</b>	45
<b>3.1 BASIC PRINCIPLES OF MULTIVARIATE DESCRIPTIVE METHODS</b>	46
<b>3.2 CORRESPONDENCE ANALYSIS</b>	47
3.2.1 Brief historical overview	47
3.2.2 Correspondence analysis - a simple numerical example	47
3.2.3 Validity of the representation	55
3.2.4 Active and supplementary variables	60
3.2.5 A comparison with principal components analysis	63
<b>3.3 MULTIPLE CORRESPONDENCE ANALYSIS</b>	69
3.3.1 Basic structure of a survey sample	71
3.3.2 Validity of the representation	76
3.3.3 Positioning of supplementary variables	78
<b>4 CLUSTER ANALYSIS OF WORDS AND TEXTS</b>	81
<b>4.1 REVIEW OF HIERARCHICAL CLUSTER ANALYSIS</b>	82
4.1.1 The dendrogram	83

4.1.2 Cutting the dendrogram	84
4.1.3 Appending supplementary elements	85
4.1.4 Filtering on first principal axes	86
<b>4.2 CLASSIFICATION OF ROWS AND COLUMNS OF A LEXICAL TABLE</b>	86
4.2.1 Cluster analysis of words	87
4.2.2 Cluster analysis of texts	90
4.2.3 Notes on cluster analysis of words	91
<b>4.3 CLUSTER ANALYSIS OF SURVEY DATA SETS</b>	94
4.3.1 Mixed clustering algorithms	95
4.3.2 Sequence of operations in survey analysis	96
4.3.3 Application example: working demographic partition	97
<b>5 VISUALIZATION OF TEXTUAL DATA</b>	101
<b>5.1 CORRESPONDENCE ANALYSIS OF LEXICAL TABLES</b>	102
5.1.1 Basic lexical tables	102
5.1.2 Aggregated lexical tables	103
5.1.3 Frequency threshold for words	104
5.1.4 Example	104
5.1.5 Construction of aggregated lexical and segmental table	108
5.1.6 Analysis and interpretation of lexical tables	111
5.1.7 Illustration of displays using repeated segments	115
5.1.8 Stability vis-à-vis an internal lemmatization	115
<b>5.2 WORKING DEMOGRAPHIC PARTITIONS</b>	118
<b>5.3 DIRECT ANALYSIS OF RESPONSES OR DOCUMENTS</b>	121
5.3.1 How are distances interpreted?	122
5.3.2 Analysis of sparse matrix T	123
5.3.3 Application example	124
<b>6 CHARACTERISTIC TEXTUAL UNITS, MODAL RESPONSES AND MODAL TEXTS</b>	129
<b>6.1 CHARACTERISTIC ELEMENTS</b>	130
6.1.1 Computation of characteristic elements	130
6.1.2 List of characteristic units	134
<b>6.2 MODAL RESPONSES</b>	136
6.2.1 Selection of modal responses using characteristic elements	137

6.2.2 Selection of modal responses using chi-square distances	140
6.2.3 Implementation and examples	141
<b>7 LONGITUDINAL PARTITIONS, TEXTUAL TIME SERIES</b>	<b>147</b>
7.1 LONGITUDINAL PARTITIONING OF A CORPUS	147
7.1.1 Longitudinal partitioning example	148
7.1.2 Analysis of <i>age category</i> gradation	149
7.1.3 Adjacent characteristic elements	150
7.2 TEXTUAL TIME SERIES	153
7.2.1 <i>Speeches</i> time series	154
7.2.2 Chronological characteristic elements	155
7.2.3 Characteristic increments	157
7.2.4 Parallel analysis of a lemmatized corpus	161
<b>8 TEXTUAL DISCRIMINANT ANALYSIS</b>	<b>163</b>
8.1 TWO MAJOR AREAS OF CONCERN IN TEXTUAL ANALYSIS	164
8.1.1 Discriminant analysis based on patterns: stylometry	164
8.1.2 Global discriminant analysis: information retrieval, coding, validation	165
8.2 UNITS AND INDICES OF STYLOMETRY	166
8.2.1 Function words, speech parts	167
8.2.2 Richness of vocabulary	168
8.3 STATISTICAL MODELS IN STYLOMETRY: AN EXAMPLE	169
8.3.1 Modelling a frequency distribution	169
8.3.2 Available data for attribution problems	170
8.3.3 Other approaches to the problem	173
8.4 GLOBAL DISCRIMINANT ANALYSIS	174
8.4.1 General principles	175
8.4.2 Units for global discriminant analysis	177
8.4.3 Discriminant analysis and modal responses	177
8.4.4 Discriminant analysis regularized through preliminary correspondence analysis	179
8.4.5 Validation of a discriminant analysis	179

<b>8.5 GLOBAL DISCRIMINATION AND VALIDATION</b>	<b>181</b>
8.5.1 Example and problem	181
8.5.2 Vocabulary and analysis for Tokyo	185
8.5.3 Reality of patterns	192
8.5.4 Discriminant analysis and confusion matrices	193
8.5.5 Conclusions to section 8.5	199
 <b>Appendix 1: Singular value decomposition and             correspondence analysis</b>	 <b>200</b>
<b>Appendix 2: Clustering techniques</b>	<b>211</b>
<b>Appendix 3: More details about the nonparametric             estimation model</b>	 <b>219</b>
<b>Appendix 4: Search for repeated segments in a corpus</b>	<b>221</b>
<b>Glossary</b>	<b>224</b>
<b>References</b>	<b>229</b>
<b>Author Index</b>	<b>238</b>
<b>Subject Index</b>	<b>242</b>
<b>Symbols</b>	<b>246</b>