

Contents

1	Introduction	1
1.1	The Corpus as a Model of Linguistic Use	2
1.2	The Internal and External Dimensions in the Corpus	3
1.3	The Predictive Power of the Corpus	5
1.4	Genres and Registers	6
1.5	Linguistic Variation across Genres and Registers	9
2	Language Resources	11
2.1	General Corpora	11
2.1.1	The Brown Corpus and the Brown Family	11
2.1.2	The International Corpus of English (ICE) Family	14
2.1.3	BNC and ANC	17
2.2	Specialised Collections	22
2.2.1	Wall Street Journal	22
2.2.2	PubMed	23
2.3	Lexical Sources	24
2.3.1	WordNet	24
2.3.2	FrameNet	24
3	Corpus Annotation and Usable Linguistic Features	27
3.1	Textual Annotation	28
3.2	Grammatical Annotation	29
3.2.1	The LOB Tagset	30
3.2.2	The ICE Tagset	32
3.2.3	A Comparison of LOB and ICE	35
3.3	Syntactic Annotation	39
3.3.1	The Penn Treebank Scheme	40
3.3.2	The ICE Parsing Scheme	42
3.3.3	Summary	44
3.4	Dialogue Act Annotation	44
3.4.1	Notable DA Schemes	47
3.4.2	ISO DA Scheme	49

3.5	Machine Learning and Linguistic Features.....	51
3.5.1	Machine Learning and Text Classification.....	51
3.5.2	Weka.....	53
4	Etymological Features across Genres and Registers	55
4.1	Research Background.....	55
4.2	Resources	57
4.2.1	Corpus Resource	57
4.2.2	Lexical Resource.....	58
4.2.3	Reference Lists.....	59
4.3	Investigation of Text Categories	60
4.3.1	Descriptive Statistics.....	60
4.3.2	Borrowed Words and Text Categories.....	61
4.3.3	Summary	64
4.4	Investigation of Subject Domains	65
4.4.1	Creation of a Sub-corpus.....	66
4.4.2	Descriptive Statistics.....	66
4.4.3	Borrowed Words and Domains	67
4.4.4	Summary	69
4.5	Conclusion.....	70
5	Part-of-Speech Tags and ICE Text Classification	71
5.1	Research Background.....	71
5.2	Methodology	72
5.2.1	Experimental Setup	72
5.2.2	Corpus Resources.....	73
5.2.3	Machine-Learning Tools	74
5.3	Feature Sets	74
5.3.1	Fine-Grained POS Tags (F-POS).....	74
5.3.2	BOW	75
5.3.3	Impoverished Tags (I-POS).....	75
5.4	Experimental Results	76
5.4.1	Results Obtained From NB Classifier.....	76
5.4.2	Results Obtained from NB-MN Classifier.....	77
5.4.3	Discussion	80
5.5	Conclusion.....	82
6	Verbs and Text Classification	83
6.1	Transitivity Type and Text Categories.....	83
6.1.1	The Distribution of Lexical Verbs	84
6.1.2	The Distribution of Verb Transitivity Types.....	88
6.1.3	Conclusion.....	95
6.2	Infinitive Verbs and Text Categories.....	97
6.2.1	The Overall Distribution of Infinitives	99
6.2.2	Aux Infinitives.....	100

6.2.3	Bare Infinitives.....	103
6.2.4	<i>To</i> -Infinitives.....	105
6.2.5	<i>For/to</i> -Infinitives.....	109
6.2.6	Summary and Conclusion	115
7	Adjectives and Text Categories.....	117
7.1	Adjective and Formality.....	117
7.1.1	Research Background.....	117
7.1.2	Methodology	118
7.1.3	Adjective Use Across Text Categories	119
7.1.4	Adjective Density and Automatic Text Classification.....	124
7.1.5	Conclusion.....	126
7.2	Adjective Phrase (AJP) and Subject Domains	127
7.2.1	Corpus Resource	127
7.2.2	Investigation of Adjective Use.....	131
7.2.3	Conclusion.....	133
8	Adverbial Clauses across Text Categories and Registers.....	135
8.1	Adverbial Clauses Across Speech and Writing.....	136
8.1.1	Adverbial Clauses Across Spontaneous and Prepared Speech	137
8.1.2	Adverbial Clauses Across Timed and Untimed Essays.....	138
8.2	Frequency Distribution of Adverbial Subordinators	139
8.3	Discussions and Conclusion.....	141
9	Coordination across Modes, Genres and Registers	143
9.1	Methodology and Corpus Data	150
9.2	The Distribution of Coordinators	153
9.3	Syntactic Categories of Coordination Conjoins.....	156
9.4	Syntactic Functions of Coordination.....	160
9.5	Conclusion.....	165
10	Semantic Features and Authorship Attribution.....	167
10.1	Corpus Annotated with Ontological Knowledge	171
10.2	Selection and Evaluation of Stylistic Features.....	175
10.3	Discussions and Conclusion.....	180
11	Pragmatics and Dialogue Acts.....	183
11.1	Corpus Resource	184
11.2	Related Research on the SWBD-DAMSL Scheme.....	188
11.3	Methodology	190
11.3.1	Machine Learning Techniques	190
11.3.2	Data Preprocessing.....	190
11.3.3	Research Questions	191
11.4	Classification Results	191

11.5 Qualitative Analysis	194
11.5.1 Hedge	194
11.5.2 Statement-non-Opinion Vs. Statement-Opinion	203
11.5.3 Acknowledge (Backchannel).....	206
11.6 Conclusions	214
12 The Future	217
Appendix A: A List of ICE Part-of-Speech Tags	221
Appendix B: A List of LOB Part-of-Speech Tags.....	229
Appendix C: A List of Penn Treebank Part-of-Speech Tags	233
Appendix D: A List of ICE Parsing Symbols.....	235
Appendix E: A List of Penn Treebank Parsing Symbols	237
Appendix F: A List of Adverbial Subordinators in Speech.....	239
Appendix G: A List of Adverbial Subordinators in Writing.....	243
Bibliography	245
Index	261

Text Genres and Registers: The Computation of
Linguistic Features

Fang, C.A.; Cao, J.

2015, XIII, 267 p. 40 illus., Hardcover

ISBN: 978-3-662-45099-4