

Contents

1	Introduction	1
1.1	Objectives and Rationale of the Book	1
1.2	Why Do We Need to Go Beyond Raw Corpora	3
1.3	What Is Corpus Annotation	5
1.4	Organization of the Book	6
	References	7
2	Text Processing with the Command Line Interface	9
2.1	The Command Line Interface	9
2.2	Basic Commands	11
2.2.1	Notational Conventions	11
2.2.2	Printing the Current Working Directory	11
2.2.3	Listing Files and Subdirectories	12
2.2.4	Making New Directories	12
2.2.5	Changing Directory Locations	13
2.2.6	Creating and Editing Text Files with UTF-8 Encoding	14
2.2.7	Viewing, Renaming, Moving, Copying, and Removing Files	16
2.2.8	Copying, Moving, and Removing Directories	20
2.2.9	Using Shell Meta-Characters for File Matching	21
2.2.10	Manual Pages, Command History, and Command Line Completion	21
2.3	Tools for Text Processing	22
2.3.1	Searching for a String with <code>egrep</code>	22
2.3.2	Regular Expressions	24
2.3.3	Character Translation with <code>tr</code>	29
2.3.4	Editing Files from the Command Line with <code>sed</code>	30
2.3.5	Data Filtering and Manipulation Using <code>awk</code>	31
2.3.6	Task Decomposition and Pipes	35
2.4	Summary	38
	References	38

3	Lexical Annotation	39
3.1	Part-of-Speech Tagging	39
3.1.1	What is Part-of-Speech Tagging	39
3.1.2	Understanding Part-of-Speech Tagsets	42
3.1.3	The Stanford Part-of-Speech Tagger	46
3.2	Lemmatization	54
3.2.1	What is Lemmatization and Why is it Useful	54
3.2.2	The TreeTagger	55
3.3	Additional Tools	58
3.3.1	The Stanford Tokenizer	58
3.3.2	The Stanford Word Segmenter for Arabic and Chinese	59
3.3.3	The CLAWS Tagger for English	61
3.3.4	The Morpha Lemmatizer for English	61
3.4	Summary	64
	References	64
4	Lexical Analysis	67
4.1	Frequency Lists	67
4.1.1	Working with Output Files from the TreeTagger	68
4.1.2	Working with Output Files from the Stanford POS Tagger and Morpha	72
4.1.3	Analyzing Frequency Lists with Text Processing Tools	73
4.2	N-Grams	76
4.3	Lexical Richness	80
4.3.1	Lexical Density	80
4.3.2	Lexical Variation	82
4.3.3	Lexical Sophistication	84
4.3.4	Tools for Lexical Richness Analysis	84
4.4	Summary	90
	References	91
5	Syntactic Annotation	95
5.1	Syntactic Parsing Overview	95
5.1.1	What is Syntactic Parsing and Why is it Useful?	95
5.1.2	Phrase Structure Grammars	96
5.1.3	Dependency Grammars	102
5.2	Syntactic Parsers	106
5.2.1	The Stanford Parser	106
5.2.2	Collins' Parser	110
5.3	Summary	112
	References	113

6 Syntactic Analysis	115
6.1 Querying Syntactically Parsed Corpora	115
6.1.1 Tree Relationships	115
6.1.2 Tregex	121
6.2 Syntactic Complexity Analysis	130
6.2.1 Measures of Syntactic Complexity	130
6.2.2 Syntactic Complexity Analyzers	136
6.3 Summary	142
References	142
7 Semantic, Pragmatic and Discourse Analysis	147
7.1 Semantic Field Analysis	147
7.1.1 The UCREL Semantic Analysis System	147
7.1.2 Profile in Semantics-Lexical in Computerized Profiling	152
7.2 Analysis of Propositions	154
7.2.1 Computerized Propositional Idea Density Rater	154
7.2.2 Analysis of Propositions in Computerized Profiling	157
7.3 Conversational Act Analysis in Computerized Profiling	158
7.4 Coherence and Cohesion Analysis in Coh-Metrix	160
7.4.1 Referential Cohesion Features	160
7.4.2 Features Based on Latent Semantic Analysis	161
7.4.3 Features Based on Connectives	162
7.4.4 Situation Model Features	163
7.4.5 Word Information Features	164
7.5 Text Structure Analysis	164
7.6 Summary	169
References	170
8 Summary and Outlook	175
8.1 Summary of the Book	175
8.2 Future Directions in Computational Corpus Analysis	177
8.2.1 Computational Analysis of Language Meaning and Use	178
8.2.2 Computational Analysis of Learner Language	178
8.2.3 Computational Analysis Based on Specific Language Theories	180
References	182
Appendix	185