
Contents

Introduction	ix
Chapter 1. Linguistic Resources for NLP	1
1.1. The concept of a corpus	1
1.2. Corpus taxonomy	4
1.2.1. Written versus spoken	4
1.2.2. The historical point of view	5
1.2.3. The language of corpora	5
1.2.4. Thematic representativity	7
1.2.5. Age range of speakers	8
1.3. Who collects and distributes corpora?	8
1.3.1. The Gutenberg project	9
1.3.2. The linguistic data consortium	9
1.3.3. European language resource agency	9
1.3.4. Open language archives community	10
1.3.5. Miscellaneous	10
1.4. The lifecycle of a corpus	10
1.4.1. Needs analysis	12
1.4.2. Design of scenarios to collect data for the corpus	12
1.4.3. Collection of the corpus	12
1.4.4. Transcription	16
1.4.5. Corpus annotation	18
1.4.6. Corpus documentation	22
1.4.7. Statistical analysis of data	22
1.4.8. The use of corpora in NLP	23

1.5. Examples of existing corpora	23
1.5.1. American National Corpus	23
1.5.2. Oxford English Corpus	23
1.5.3. The Grenoble Tourism Office Corpus.	24
Chapter 2. The Sphere of Speech	25
2.1. Linguistic studies of speech	25
2.1.1. Phonetics	25
2.1.2. Phonology	46
2.2. Speech processing	61
2.2.1. Automatic speech recognition	62
2.2.2. Speech synthesis	80
Chapter 3. Morphology Sphere	89
3.1. Elements of morphology	89
3.1.1. Morphological typology	90
3.1.2. Morphology of English	91
3.1.3. Parts of speech	95
3.1.4. Terms, collocations and colligations.	99
3.2. Automatic morphological analysis	100
3.2.1. Stemming	101
3.2.2. Regular expressions for morphological analysis.	104
3.2.3. Informal introduction to finite-state machines	108
3.2.4. Two-level morphology and FST	112
3.2.5. Part-of-speech tagging	117
Chapter 4. Syntax Sphere	127
4.1. Basic syntactic concepts	127
4.1.1. Delimitation of the field of syntax	127
4.1.2. The concept of grammaticality	128
4.1.3. Syntactic constituents	129
4.1.4. Syntactic typology of topology and agreement.	139
4.1.5. Syntactic ambiguity	140
4.1.6. Syntactic specificities of spontaneous oral language	141
4.2. Elements of formal syntax	145
4.2.1. Syntax trees and rewrite rules.	145
4.2.2. Languages and formal grammars.	152
4.2.3. Hierarchy of languages (Chomsky–Schützenberger)	154

4.2.4. Feature structures and unification	162
4.2.5. Definite clause grammar	169
4.3. Syntactic formalisms	171
4.3.1. X-bar	171
4.3.2. Head-driven phrase structure grammar	178
4.3.3. Lexicalized tree-adjoining grammar	193
4.4. Automatic parsing	201
4.4.1. Finite-state automata	202
4.4.2. Recursive transition networks	203
4.4.3. Top-down approach	207
4.4.4. Bottom-up approach	212
4.4.5. Mixed approach: left-corner	215
4.4.6. Tabular parsing (chart)	221
4.4.7. Probabilistic parsing	225
4.4.8. Neural networks	233
4.4.9. parsing algorithms for unification-based grammars	237
4.4.10. Robust parsing approaches	238
4.4.11. Generation algorithms	242
Bibliography	245
Index	275