Contents

Acknowledgments xi

Chapter 1. Introduction: the Project 1

1.1. Characterizing a set of infinite size 4
1.2. Computers and linguistics 5
1.3. Levels of formalization 6
1.4. Not applicable 7
1.4.1. Poetry and plays on words 7
1.4.2. Stylistics and rhetoric 9
1.4.3. Anaphora, coreference resolution, and semantic disambiguation 10
1.4.4. Extralinguistic calculations 12
1.5. NLP applications 12
1.5.1. Automatic translation 14
1.5.2. Part-of-speech (POS) tagging 18
1.5.3. Linguistic rather than stochastic analysis 27
1.6. Linguistic formalisms: NooJ 27
1.7. Conclusion and structure of this book 30
1.8. Exercises 31
1.9. Internet links 32

Part 1. Linguistic Units 35

Chapter 2. Formalizing the Alphabet 37

2.1. Bits and bytes 37
2.2. Digitizing information 39
2.3. Representing natural numbers 39
2.3.1. Decimal notation 39

2.3.2. Binary notation . . . 40
2.3.3. Hexadecimal notation . . . 41
2.4. Encoding characters . . . 41
2.4.1. Standardization of encodings . . . 43
2.4.2. Accented Latin letters, diacritical marks, and ligatures . . . 45
2.4.3. Extended ASCII encodings . . . 46
2.4.4. Unicode . . . 47
2.5. Alphabetical order . . . 53
2.6. Classification of characters . . . 56
2.7. Conclusion . . . 56
2.8. Exercises . . . 57
2.9. Internet links . . . 57

Chapter 3. Defining Vocabulary . . . 59

3.1. Multiple vocabularies and the evolution of vocabulary . . . 59
3.2. Derivation . . . 63
3.2.1. Derivation applies to vocabulary elements . . . 63
3.2.2. Derivations are unpredictable . . . 64
3.2.3. Atomicity of derived words . . . 65
3.3. Atomic linguistic units (ALUs) . . . 67
3.3.1. Classification of ALUs . . . 67
3.4. Multiword units versus analyzable sequences of simple words . . . 70
3.4.1. Semantics . . . 72
3.4.2. Usage . . . 76
3.4.3. Transformational analysis . . . 77
3.5. Conclusion . . . 80
3.6. Exercises . . . 81
3.7. Internet links . . . 81

Chapter 4. Electronic Dictionaries . . . 83

4.1. Could editorial dictionaries be reused? . . . 83
4.2. LADL electronic dictionaries . . . 90
4.2.1. Lexicon-grammar . . . 90
4.2.2. DELA . . . 93
4.3. Dubois and Dubois-Charlier electronic dictionaries . . . 94
4.3.1. The Dictionnaire électronique des mots . . . 95
4.3.2. Les Verbes Français (LVF) . . . 97
4.4. Specifications for the construction of an electronic dictionary . . . 99

4.4.1. One ALU = one lexical entry 99
4.4.2. Importance of derivation. 100
4.4.3. Orthographic variation 101
4.4.4. Inflection of simple words, compound words, and expressions 103
4.4.5. Expressions 104
4.4.6. Integration of syntax and semantics 104
4.5. Conclusion 107
4.6. Exercises 108
4.7. Internet links 108

Part 2. Languages, Grammars and Machines 111

Chapter 5. Languages, Grammars, and Machines 113

5.1. Definitions 113
5.1.1. Letters and alphabets 113
5.1.2. Words and languages. 114
5.1.3. ALU, vocabularies, phrases, and languages 114
5.1.4. Empty string 115
5.1.5. Free language 116
5.1.6. Grammars 116
5.1.7. Machines 117
5.2. Generative grammars 118
5.3. Chomsky-Schützenberger hierarchy 119
5.3.1. Linguistic formalisms 122
5.4. The NooJ approach 124
5.4.1. A multifaceted approach. 124
5.4.2. Unified notation 125
5.4.3. Cascading architecture 127
5.5. Conclusion 127
5.6. Exercises 128
5.7. Internet links 129

Chapter 6. Regular Grammars 131

6.1. Regular expressions 131
6.1.1. Some examples of regular expressions 135
6.2. Finite-state graphs 137
6.3. Non-deterministic and deterministic graphs. 139
6.4. Minimal deterministic graphs 141
6.5. Kleene's theorem 142
6.6. Regular expressions with outputs and finite-state transducers 146

6.7. Extensions of regular grammars . . . 151
6.7.1. Lexical symbols . . . 151
6.7.2. Syntactic symbols . . . 153
6.7.3. Symbols defined by grammars . . . 154
6.7.4. Special operators . . . 155
6.8. Conclusion . . . 159
6.9. Exercises . . . 159
6.10. Internet links . . . 159

Chapter 7. Context-Free Grammars . . . 161

7.1. Recursion . . . 164
7.1.1. Right recursion . . . 166
7.1.2. Left recursion . . . 167
7.1.3. Middle recursion . . . 168
7.2. Parse trees . . . 170
7.3. Conclusion . . . 173
7.4. Exercises . . . 173
7.5. Internet links . . . 174

Chapter 8. Context-Sensitive Grammars . . . 175

8.1. The NooJ approach . . . 176
8.1.1. The $a^n b^n c^n$ language . . . 177
8.1.2. The language a^{2^n} . . . 180
8.1.3. Handling reduplications . . . 181
8.1.4. Grammatical agreements . . . 182
8.1.5. Lexical constraints in morphological grammars . . . 185
8.2. NooJ contextual constraints . . . 186
8.3. NooJ variables . . . 188
8.3.1. Variables' scope . . . 188
8.3.2. Computing a variable's value . . . 189
8.3.3. Inheriting a variable's value . . . 191
8.4. Conclusion . . . 191
8.5. Exercises . . . 192
8.6. Internet links . . . 192

Chapter 9. Unrestricted Grammars . . . 195

9.1. Linguistic adequacy . . . 197
9.2. Conclusion . . . 199
9.3. Exercise . . . 199
9.4. Internet links . . . 199

Part 3. Automatic Linguistic Parsing 201

Chapter 10. Text Annotation Structure 205

10.1. Parsing a text 205
10.2. Annotations 206
10.2.1. Limits of XML/TEI representation 207
10.3. Text annotation structure (TAS) 208
10.4. Exercise 211
10.5. Internet links 212

Chapter 11. Lexical Analysis 213

11.1. Tokenization 213
11.1.1. Letter recognition 214
11.1.2. Apostrophe/quote 217
11.1.3. Dash/hyphen 219
11.1.4. Dot/period/point ambiguity 222
11.2. Word forms 224
11.2.1. Space and punctuation 224
11.2.2. Numbers 226
11.2.3. Words in upper case 228
11.3. Morphological analyses 229
11.3.1. Inflectional morphology 230
11.3.2. Derivational morphology 234
11.3.3. Lexical morphology 236
11.3.4. Agglutinations 239
11.4. Multiword unit recognition 241
11.5. Recognizing expressions 243
11.5.1. Characteristic constituent 244
11.5.2. Varying the characteristic constituent 245
11.5.3. Varying the light verb 246
11.5.4. Resolving ambiguity 247
11.5.5. Annotating expressions 251
11.6. Conclusion 254
11.7. Exercise 255

Chapter 12. Syntactic Analysis 257

12.1. Local grammars 257
12.1.1. Named entities 257
12.1.2. Grammatical word sequences 262
12.1.3. Automatically identifying ambiguity 263
12.2. Structural grammars 265

12.2.1. Complex atomic linguistic units 266
12.2.2. Structured annotations 268
12.2.3. Ambiguities 270
12.2.4. Syntax trees vs parse trees 273
12.2.5. Dependency grammar and tree 276
12.2.6. Resolving ambiguity transparently 279
12.3. Conclusion 280
12.4. Exercises. 281
12.5. Internet links 281

Chapter 13. Transformational Analysis 283

13.1. Implementing transformations 286
13.2. Theoretical problems 292
13.2.1. Equivalence of transformation sequences 292
13.2.2. Ambiguities in transformed sentences 293
13.2.3. Theoretical sentences 294
13.2.4. The number of transformations to be implemented 295
13.3. Transformational analysis with NooJ 297
13.3.1. Applying a grammar in “generation” mode 298
13.3.2. The transformation’s arguments 299
13.4. Question answering 303
13.5. Semantic analysis 304
13.6. Machine translation 305
13.7. Conclusion 309
13.8. Exercises. 309
13.9. Internet links 310

Conclusion 311

Bibliography 315

Index 327