

Contents

Preface	xiii
Contributing Authors	xv
Part I The User's View	
1 Orientation	3
<i>Atro Voutilainen</i>	
1.1 Morphosyntactic tags	3
1.2 Automatic tagging	6
2 A Short History of Tagging	9
<i>Atro Voutilainen</i>	
2.1 Approaches to wordclass tagging	9
2.2 Pioneering work	10
2.3 The breakthrough of data-driven methods	11
2.3.1 N-gram taggers	12
2.3.2 Data-driven local rules	13
2.4 Recent work in the data-driven approach	14
2.4.1 Hidden Markov Models	14
2.4.2 Recent work on data-driven local rules	16
2.4.3 Neural taggers	16
	v

2.4.4	Case-based taggers	17
2.4.5	Combined data-driven taggers	17
2.5	Recent work in the linguistic approach	17
2.5.1	English Constraint Grammar	18
2.5.2	A rule-based tagger of Turkish	18
2.5.3	A finite-state tagger of French	19
2.5.4	A syntax-based tagger of English	19
2.6	The current situation	19
3	The Use of Tagging	23
	<i>Geoffrey Leech and Nicholas Smith</i>	
3.1	Introduction	23
3.2	Tagging in corpus linguistics	24
3.2.1	Adding further annotations	26
3.2.2	Information extraction	28
3.3	Practical applications	31
3.3.1	Uses of tagging software	31
3.3.2	Uses of tagged text	33
4	Tagsets	37
	<i>Jan Cloeren</i>	
4.1	Introduction	37
4.2	Information contents of the tags in the tagset	37
4.2.1	Morphosyntactic tags	38
4.2.2	Syntactic tags	40
4.2.3	Semantic and discourse tags	41
4.2.4	Distributional similarity tags	42
4.3	Special problems in the application of tagsets	44
4.3.1	Multi-unit tokens and multi-token units	44
4.3.2	Underspecification and ambiguity	46
4.4	Notation	49
4.4.1	Class and feature value names	49
4.4.2	Structure of tags	50
4.4.3	Positioning of tags	51
4.4.4	SGML/TEI guidelines for tags	51
5	Standards for Tagsets	55
	<i>Geoffrey Leech and Andrew Wilson</i>	
5.1	Introduction	55
5.2	Recommendations for morphosyntactic (wordclass) categories	58
5.2.1	Reasonable goals for standardization	58
5.2.2	Word categories: tagset guidelines	61
5.3	Intermediate Tagset	70
5.3.1	Basic Structure	70
5.3.2	Underspecification	71
5.3.3	Example tagsets	72

6	Performance of Taggers	81
	<i>Hans van Halteren</i>	
6.1	Introduction	81
6.2	Performance measures	81
6.2.1	Definitions of measures	82
6.2.2	Usefulness of measures	83
6.3	Performance measurements	86
6.3.1	Experimental setup	86
6.3.2	Effects of the tagset	87
6.3.3	Effects of the method of comparison	89
6.3.4	Effects of choice of tokens measured	90
6.3.5	Effects of separation of test and training material	91
6.3.6	Effects of representativity of test material	94
7	Selection and Operation of Taggers	95
	<i>Hans van Halteren</i>	
7.1	Introduction	95
7.2	Selection of a tagger	95
7.2.1	Tagset	96
7.2.2	Documentation	96
7.2.3	The tagging process	97
7.2.4	Performance	98
7.2.5	Combining the factors	98
7.3	User interaction	99
7.3.1	Tokenization	100
7.3.2	Classification of unknown tokens	101
7.3.3	Selection of the contextually appropriate tag	101
7.3.4	Post-processing of tagged text	102
	Appendix: NOT an inventory of taggers	103
 Part II The Implementer's View		
8	Automatic Taggers: An Introduction	109
	<i>Hans van Halteren and Atro Voutilainen</i>	
8.1	General architecture	109
8.1.1	Tokenization	110
8.1.2	Assignment of potential tags	110
8.1.3	Determination of the most likely tag	110
8.2	Corpus resources	110
8.2.1	Form of corpus resources	111
8.2.2	Size of corpus resources	113
8.2.3	Creation of corpus resources	114
9	Tokenization	117
	<i>Gregory Grefenstette</i>	
9.1	Introduction	117
9.2	Regular expressions	119

9.2.1	Definition	119
9.2.2	Regular expression tools LEX and AWK	121
9.2.3	An example of a tokenizer	121
9.3	Ambiguity in tokenization	125
9.3.1	Splitting graphic tokens	125
9.3.2	Combining graphic tokens	132
10	Lexicons for Tagging	135
	<i>Anne Schiller and Lauri Karttunen</i>	
10.1	Introduction	135
10.2	Morphology-based lexicons	137
10.2.1	Direct mapping	140
10.2.2	Merging morphological classes	141
10.2.3	Refining morphological classes	141
10.2.4	Adding residual wordclasses	143
10.3	Corpus-based lexicons	144
10.3.1	Enlarged Training Corpus	146
10.3.2	External Lexical Resources	146
11	Standardization in the Lexicon	149
	<i>Monica Monachini and Nicoletta Calzolari</i>	
11.1	The initiative for standardization	149
11.2	Interdependence between lexicon and corpus	151
11.2.1	Lexical encoding vs. corpus tagsets	151
11.2.2	Tagsets as collapsed feature specifications	152
11.2.3	Multi-linguality	152
11.2.4	Lexicon specifications as an interface between tagsets	153
11.3	The EAGLES proposal for morphosyntactic encoding	156
11.3.1	Methodology of standardization	157
11.3.2	The proposal	159
11.4	Instantiation in different languages	161
11.5	Guidelines for the validation phase	164
11.5.1	Values pertinent to a given language	165
11.5.2	Logic relationships between values	165
11.5.3	Constraints in the application of attributes and values	167
11.5.4	Semantics of the PoS	168
11.5.5	Semantics of the features	170
11.6	Application in EU projects	171
11.6.1	MULTEXT	171
11.6.2	PAROLE	172
11.6.3	Coverage with respect to languages, users and applications	173
12	Morphological Analysis	175
	<i>Kemal Oflazer</i>	
12.1	Introduction	175
12.2	Morphology	177
12.2.1	Types of morphology	177
12.2.2	Types of morphological combination	178

12.2.3	Computational morphology	179
12.3	Two-level morphology	180
12.3.1	The morphographemic component	181
12.3.2	The morphotactics component	186
12.3.3	Development tools	190
12.3.4	Developing a Morphological Analyser	193
12.4	A morphological analyser for Turkish	194
12.4.1	Requirements	195
12.4.2	System architecture	199
12.4.3	The morphographemic transducer: T_{is-ix}	202
12.4.4	The morphotactics transducer: T_{ix-if}	203
13	Tagging Unknown Words	207
	<i>Eric Brill</i>	
13.1	Introduction	207
13.2	Behaviour of unknown words	207
13.3	Dealing with unknown words	209
13.4	Unknown words in case-based tagging	211
13.5	Unknown words in transformation-based tagging	212
13.6	Lexicon extrapolation	215
14	Hand-Crafted Rules	217
	<i>Astro Voutilainen</i>	
14.1	Introduction	217
14.2	Comparison of paradigms	218
14.3	Rule formalism	219
14.3.1	Overview	220
14.3.2	Operations	220
14.3.3	Targets	221
14.3.4	Context conditions	221
14.3.5	Sample rules	222
14.3.6	Some facts about a large grammar	224
14.4	Writing a disambiguation grammar	226
14.4.1	A sample session	227
14.4.2	Experiences with novices: NorFa'95 CG 'competition'	240
14.5	General observations	242
14.6	Remaining ambiguity	243
14.6.1	Using statistical models	244
14.6.2	Using collocational information	244
14.6.3	Using a syntactic parser	245
14.6.4	Using observed local regularities	245
15	Corpus-Based Rules	247
	<i>Eric Brill</i>	
15.1	Introduction	247
15.2	Learning rules	248
15.3	Parser-based wordclass disambiguation	249

15.4	Transformation-based learning	251
15.5	N-best wordclass tagging	256
15.6	Unsupervised learning	258
15.7	Issues of portability	261
16	Hidden Markov Models	263
	<i>Marc El-Beze and Bernard Merialdo</i>	
16.1	Introduction	263
16.2	HMMs in general	264
	16.2.1 Definition and use	264
	16.2.2 An example	265
	16.2.3 Choosing the underlying topology	267
	16.2.4 Training	268
	16.2.5 Decoding	269
16.3	HMMs for wordclass tagging	271
	16.3.1 The structure of the model	271
	16.3.2 Choice of tagset when using HMMs	273
16.4	Training HMMs for tagging	274
	16.4.1 Training on tagged text	274
	16.4.2 Smoothing the triclass model	276
	16.4.3 Training HMM taggers with Baum-Welch	278
16.5	Tagging with HMMs	280
	16.5.1 Using the Viterbi algorithm	280
	16.5.2 Other forms of decoding	282
16.6	Combining different linguistic levels	282
	16.6.1 Using wordclasses in word models	282
	16.6.2 Using lemma models in wordclass tagging	284
17	Machine Learning Approaches	285
	<i>Walter Daelemans</i>	
17.1	Introduction	285
17.2	Inductive learning from examples	287
	17.2.1 Concepts	287
	17.2.2 Classification of learning methods	288
	17.2.3 Performance evaluations	290
	17.2.4 Overview of methods	290
17.3	Case-based learning	291
	17.3.1 Algorithm	292
	17.3.2 Case-based tagging	293
	17.3.3 Evaluation	296
17.4	Decision tree induction	297
	17.4.1 Algorithm	297
	17.4.2 Decision tree tagging	298
	17.4.3 Evaluation	300
17.5	Neural network methods	300
	17.5.1 Algorithm	301
	17.5.2 Neural network tagging	302

17.5.3 Evaluation	303
17.6 Discussion	303
Appendix A: Example tagsets	305
A.1 The Brown Corpus tagset	305
A.2 The Penn Treebank tagset	307
A.3 The EngCG tagset	309
References	311
Index	327