
Contents

1	Introduction	1
2	Tokenization and Segmentation	5
2.1	Methods of Tokenization	5
2.2	Normalization of Forms	7
2.3	Multi-Word Expressions	8
2.4	Word Segmentation	9
2.5	Empty Nodes	13
2.6	Sentence Segmentation	14
3	Part of Speech Tags	15
3.1	Types of Tags	15
3.2	Parallel and Serial Combination of Tags	19
3.2.1	Ambiguity	19
3.2.2	Layered Features	22
3.2.3	Chained Features	24
3.3	Harmonization Efforts	25
3.3.1	EAGLES, PAROLE and MULTEXT-EAST	25
3.3.2	Indian Languages	30
3.3.3	Interset, UPOS and Universal Dependencies	30
3.3.4	UniMorph	32
3.4	How to Define a Part-of-Speech Category	35
3.5	Part-of-Speech Categories	40
3.5.1	Nouns	40
3.5.2	Verbs	43
3.5.3	Adjectives	44
3.5.4	Adverbs	45

3.5.5	Pronouns, Determiners and Quantifiers	47
3.5.6	Adpositions, Conjunctions, Linkers and Particles	50
3.5.7	Interjections and Onomatopoeia	52
3.5.8	Other	52
4	Morphological Features	55
4.1	Gender	56
4.2	Animacy	58
4.3	Noun Class	59
4.4	Number	60
4.5	Case	63
4.5.1	Core Cases	64
4.5.2	Non-core Non-local Cases	66
4.5.3	Local, Temporal and Directional Cases	69
4.6	Definiteness	72
4.7	Degree of Comparison	74
4.8	Polarity	76
4.9	Person	77
4.10	Clusivity	78
4.11	Politeness	79
4.12	Deixis	80
4.13	Cross-reference of Possessor	81
4.14	Cross-reference of Verbal Arguments	82
4.15	Tense	84
4.16	Aspect	86
4.17	Voice	87
4.18	Mood	91
4.19	Evidentiality	94
5	Dependency Trees	95
5.1	Simple Noun Phrases	97
5.2	Quantifiers and Classifiers	103
5.3	Simple Clauses	105
5.4	Verb Groups	111

5.5	Clauses with Non-Verbal Predicates	116
5.6	Subordinate Clauses	120
5.7	Coordination	123
6	Some Concluding Tokens	133
	Summary	135
	List of Figures	137
	List of Tables	141
	Language Index	157