Contents

1 Introduction **21**

2 Formal Morphology of Czech **25**

2.1 Formal Morphology 25

2.1.1 Basic Notions and Notation 25

2.1.2 Full Morphology 26

2.1.3 Lemmatization 27

2.1.4 Tagging 28

2.1.5 Spelling Checking 29

2.2 Introduction to The Czech Tag System 31

2.3 The Positional Tag System for Czech 32

2.3.1 Format and Columns 32

2.3.2 Description of the Values of the Categories 34

2.3.3 Co-occurrence Tables 35

2.3.4 Part of speech (POS) 36

2.3.5 Detailed part of speech (SUBPOS) 43

2.3.6 Gender (GENDER) 66

2.3.7 Number (NUMBER) 69

2.3.8 Case (CASE) 71

2.3.9 Possessor's gender (POSSGENDER) 74

2.3.10 Possessor's number (POSSNUMBER) 76

2.3.11 Person (PERSON) 77

2.3.12 Tense (TENSE) 79

2.3.13 Degrees of Comparison (GRADE) 81

2.3.14 Negation (NEGATION) 82

2.3.15 Voice (Active/Passive) (VOICE) 83

2.3.16 Reserve position 1 (RESERVE1) 84

2.3.17 Reserve position 2 (RESERVE2) 84
2.3.18 Variant, register, style, colloquiality, etc. (VAR) 85
2.3.19 The Tag Correspondences 88
2.4 The Compact Tag System for Czech 89
2.4.1 Tag Format and "Variables" 89
2.4.2 Verbs 91
2.4.3 Nouns 92
2.4.4 Adjectives 92
2.4.5 Pronouns 93
2.4.6 Numerals 95
2.4.7 Adverbs 96
2.4.8 Prepositions 96
2.4.9 Conjunctions 96
2.4.10 Particles 96
2.4.11 Interjections 97
2.4.12 Other 97

3 The Paradigm System 99
3.1 The Paradigm Field 101
3.2 The Negation Field 102
3.3 The List of Endings 102
3.3.1 The Ending String 103
3.3.2 The Superlative Field 103
3.3.3 The List of Tags 104
3.4 Paradigm Names 104

4 The Morphological Dictionary 111
4.1 The Root Field 114
4.2 The Paradigm Field 115
4.3 The Lemma Field 116
4.4 The Tag Field 117
4.5 The Comment Field 117
4.5.1 The Derivation Info Field 118
4.5.2 The Comment Field 118
4.6 The Alternate Part of Speech Field 119
4.7 The Style Field 120
4.8 The Term Field 121

5 Analysis and Generation of Word Forms 123
5.1 Direct Analysis and Generation 124
5.2 The Data 127
5.2.1 The Ending Record 129
5.2.2 The Root Record 130
5.3 The Analysis Algorithm 132
5.3.1 Data Structures 133
5.3.2 Algorithm A1 (the main "wrapper" algorithm) 134
5.3.3 Algorithm A2 (capitalization) 136
5.3.4 Algorithm A3 (prefixing) 138
5.3.5 Algorithm A4 (token analysis proper) 140
5.4 The Generation Algorithm 143
5.4.1 Data Structures 145
5.4.2 Algorithm G1 (tag template precedence) 147
5.4.3 Algorithm G2 (tag template processing) 149
5.4.4 Algorithm G3 (word form production) 151
5.5 Other Algorithms 153

6 Morphological Disambiguation (Tagging) 155
6.1 Tagging – a step aside? 155
6.2 Orthogonality of Morphological Categories 156
6.2.1 The Morphological Analyzer Used in the Tagging Experiments 157
6.3 The Training Data 158
6.4 The Model 160
6.5 Training 164
6.5.1 Feature Weights 164
6.5.2 Feature Selection 164
6.6 Results 167
6.7 Tagging Other Languages Using Scarce Resources 171
6.7.1 Full Morphological Tagging 171
6.7.2 The Languages Used and The Training Data 172
6.7.3 The Methodology 173
6.7.4 The Results 176
6.7.5 The Differences Among Languages 182
6.7.6 Comparison to Previous Results 182
6.7.7 Dictionary vs. Training Data 183
6.7.8 Future Work 184
6.7.9 Special Acknowledgements 184

6.8 Conclusion and Further Research 185
6.9 Lemmatization and Word Sense Disambiguation 186

7 Resources & References 189

A Czech Morphological Data 197
A.1 Positional Tags – Quick Reference 198
A.1.1 Category 1: `POS` 198
A.1.2 Category 2: `SUBPOS` 198
A.1.3 Category 3: `GENDER` 202
A.1.4 Category 4: `NUMBER` 202
A.1.5 Category 5: `CASE` 202
A.1.6 Category 6: `POSSGENDER` 203
A.1.7 Category 7: `POSSNUMBER` 203
A.1.8 Category 8: `PERSON` 203
A.1.9 Category 9: `TENSE` 203
A.1.10 Category 10: `GRADE` 203
A.1.11 Category 11: `NEGATION` 204
A.1.12 Category 12: `VOICE` 204
A.1.13 Category 13: `RESERVE1` 204
A.1.14 Category 14: `RESERVE2` 204
A.1.15 Category 15: `VAR` 204
A.2 The Endings 205
A.2.1 Table Format 205
A.2.2 Summary of Signs, Abbreviations, and Headings in the Ending Table 208