Contents

List of abbreviations **11**

1 Introduction **13**
1.1 About this book 13
1.2 Reasons to study non-native Czech 14
1.3 Some properties of non-native Czech 17
1.3.1 Morphology 18
1.3.2 Syntax 19
1.3.3 Word segmentation 21
1.4 Learner corpus 21
1.5 Roadmap 23

2 Learner corpora **25**
2.1 Terminology 25
2.2 Various types of learner corpora 26
2.2.1 The choice of texts 26
2.2.2 Annotation 27
2.2.2.1 Textual annotation 27
2.2.2.2 Linguistic annotation 28
2.2.2.3 Error annotation – correction 28
2.2.2.4 Error annotation – categorization 29
2.2.2.5 Annotation scheme 30
2.2.3 Data access 31
2.3 Some learner corpora 32
2.3.1 *ASK* 32
2.3.2 *CLC* 33
2.3.3 *COPLE2* 34
2.3.4 *CroLTeC* 34

2.3.5 *Falko* . . . 35
2.3.6 *ICLE* . . . 36
2.3.7 *MERLIN* . . . 36
2.3.8 *RLC* . . . 37
2.3.9 *SweLL* . . . 38
2.4 Relationships of *CzeSL* with other learner corpora . . . 39

3 Introducing the CzeSL project 41
3.1 Specifications of *CzeSL* . . . 42
3.2 Intended usage . . . 43
3.3 *AKCES* – the umbrella project . . . 45

4 Procurement of texts 49
4.1 Text collection . . . 49
4.2 Transcription . . . 51
4.3 Anonymization . . . 53
4.4 Metadata . . . 54

5 Error annotation 59
5.1 Errors and learner language . . . 59
5.2 More than one way to annotate errors in *CzeSL* . . . 64
5.3 A wishlist for error annotation . . . 65
5.3.1 Interference and other types of explanation . . . 66
5.3.2 Interpretation in terms of TH . . . 66
5.3.3 Word order . . . 67
5.3.4 Style . . . 68
5.3.5 Communication goal . . . 68
5.4 The two-tier annotation scheme . . . 69
5.4.1 Annotation scheme as a compromise . . . 69
5.4.1.1 Why multiple tiers . . . 69
5.4.1.2 How many tiers . . . 71
5.4.1.3 Multiple tiers in a tabular format . . . 71
5.4.1.4 Content of the tiers . . . 72
5.4.1.5 A sample text with T1 vs. T2 corrections . . . 73
5.4.1.6 Links between tiers . . . 73
5.4.1.7 Error tags . . . 76
5.4.1.8 Morphosyntactic references . . . 76
5.4.1.9 Follow-up corrections . . . 77
5.4.1.10 Alternative target hypotheses . . . 77

5.4.2 Error tagset 78
5.4.2.1 Based on linguistic categories 78
5.4.2.2 Grammar-based vs. formal errors 80
5.4.2.3 Extent of the annotated unit 81
5.4.3 Grammar-based tags 81
5.4.3.1 Errors at T1 81
5.4.3.2 Errors at T2 83
5.4.3.3 Coarse-grained 83
5.4.3.4 An example of complex annotation 84
5.4.4 Evaluation of the manual tiered error annotation 87
5.4.4.1 Inter-annotator agreement (IAA) 88
5.4.4.2 A pilot annotation 89
5.4.4.3 IAA on all doubly-annotated texts 89
5.4.4.4 Error tags depend on target hypothesis 93
5.4.4.5 Possible causes of the annotators' disagreements . . 95
5.4.5 Formal tags 97
5.4.5.1 Automatic extension and modification of error annotation 97
5.4.5.2 Automatic detection of formal errors on T1 98
5.4.5.3 Formal orthographic errors 99
5.4.5.4 Formal errors sometimes influencing pronunciation . 100
5.4.5.5 Formal errors influencing pronunciation 101
5.4.5.6 Other types of errors 103
5.4.5.7 Automatic classification of word-boundary errors . . 105
5.5 Implicit error annotation 105
5.6 Multi-dimensional error annotation (MD) 108
5.6.1 Focus on morphology 108
5.6.2 All annotation applied to the source text 109
5.6.3 Extent of the annotated unit 109
5.6.4 Alternative error domains 110
5.6.5 Source text, target hypothesis, annotated strings 112
5.6.6 Domains and features 113

6 Linguistic annotation **119**
6.1 Annotation with tools for Standard Czech 120
6.1.1 Annotation of target hypothesis 120
6.1.2 Annotation of T1 121
6.1.3 Annotation of source texts 121
6.2 Annotation of interlanguage in UD 122

6.2.1 Tokenization 124
6.2.2 Part-of-speech and morphology 124
6.2.3 Lemmata 126
6.2.4 Syntactic Structure 128
6.2.5 Evaluation 130

7 Annotation process 131
7.1 Overview of the annotation process 131
7.2 Transcription and anonymization of manuscripts 132
7.3 Tiered error annotation 133
7.3.1 Manual error annotation 134
7.3.2 Automatic annotation checking 135
7.3.3 Data format for the tiered annotation scheme 136
7.4 Automatic error tagging 136
7.5 Automatic correction 139
7.6 Multi-dimensional error annotation 140
7.6.1 Morphemic analysis 141
7.6.2 Automatic error annotation 143
7.6.3 Experiments with automatic identification of errors in inflection 144
7.6.4 Manual error annotation 147
7.6.5 Post-processing of manually annotated texts 149
7.7 Implicit annotation 150
7.8 Universal Dependencies 152

8 The *CzeSL* corpora 155
8.1 *CzeSL-plain* – without annotation and metadata 157
8.2 *CzeSL-SGT* – with automatic annotation 158
8.3 *CzeSL-man* – with manual annotation 163
8.3.1 *CzeSL-man v0* 163
8.3.2 *CzeSL-man v1* 163
8.3.3 *CzeSL-man v1 downloadable* 164
8.3.4 *CzeSL-man v1 searchable* 165
8.3.5 *CzeSL-man v2* 167
8.4 *CzeSL-TH* 168
8.5 *CzeSL-MD* 168
8.6 *CzeSL-UD* 169
8.7 *CzeSL-GEC* and *AKCES-GEC* 169
8.8 *CzeSL in TEITOK* 170
8.9 Learner corpora of native Czech 170

9 Tools **173**
- 9.1 Annotation tools . . . 173
  - 9.1.1 *feat* . . . 173
  - 9.1.2 *Speed* . . . 174
  - 9.1.3 *brat* . . . 175
  - 9.1.4 *TrEd* . . . 175
  - 9.1.5 Error annotation tools . . . 175
    - 9.1.5.1 Automatic error tagging in 2T . . . 175
    - 9.1.5.2 Automatic error detection and tagging in MD . . . 176
  - 9.1.6 Conversion tools . . . 176
- 9.2 Search tools . . . 177
  - 9.2.1 *SeLaQ* . . . 177
  - 9.2.2 *Sketch Engine* and *KonText* . . . 179
    - 9.2.2.1 Token-based error annotation . . . 180
    - 9.2.2.2 Error annotation using structures . . . 183
  - 9.2.3 *TEITOK* . . . 189

10 Using the corpus **201**
- 10.1 Learner corpora from the perspective of language teachers . . . 202
- 10.2 The use of corpora in language research and teaching . . . 203
  - 10.2.1 Benefits of learner corpora . . . 204
  - 10.2.2 Limitations of learner corpora . . . 205
- 10.3 Corpus-based research and teaching of Czech as a foreign language . 207
  - 10.3.1 The *Czech National Corpus* in the service of Czech as a foreign language . . . 207
  - 10.3.2 Analyses based on learner corpus data . . . 209
- 10.4 Applications in natural language processing . . . 211
  - 10.4.1 Text scoring . . . 211
  - 10.4.2 Text correction . . . 212
  - 10.4.3 Natural language identification . . . 214

11 Lessons learned and perspectives **217**
- 11.1 What we would do the same way again . . . 217
- 11.2 Blind alleys and second thoughts . . . 220
- 11.3 Outlook . . . 225

12 Acknowledgements **227**

A Notes about examples **231**

B The Czech language 233
B.1 Morphology 234
B.1.1 Nouns 234
B.1.2 Adjectives 234
B.1.3 Pronouns 236
B.1.4 Numerals 236
B.1.5 Verbs 237
B.2 Syntax 240
B.2.1 Agreement 240
B.2.1.1 Subject-predicate agreement 240
B.2.1.2 Agreement within the NP 242
B.2.2 Numeral expressions 243
B.2.3 Negation 244
B.3 Word order and clitics 244
B.4 Romani ethnolect of Czech 245

Bibliography 247

Index of Authors 271

Index of Corpora 275

Index of Tools 277

Index 279