

**COMPILING
AND ANNOTATING
A LEARNER CORPUS FOR
A MORPHOLOGICALLY RICH
LANGUAGE**

CZESL, A CORPUS OF
NON-NATIVE CZECH

ALEXANDR ROSEN
JIŘÍ HANA
BARBORA HLADKÁ
TOMÁŠ JELÍNEK
SVATAVA ŠKODOVÁ
BARBORA ŠTINDLOVÁ

KAROLINUM

Compiling and annotating a learner corpus for a morphologically rich language
CzeSL, a corpus of non-native Czech

Alexandr Rosen
Jiří Hana
Barbora Hladká
Tomáš Jelínek
Svatava Škodová
Barbora Štindlová

Reviewed by: Detmar Meurers, University of Tübingen, Germany
Elena Volodina, University of Gothenburg, Sweden

© Charles University, 2020
© Alexandr Rosen, Jiří Hana, Barbora Hladká, Tomáš Jelínek, Svatava Škodová,
Barbora Štindlová, 2020

Published by Charles University
Karolinum Press
Ovocný trh 560/5, 116 36 Prague 1
Prague 2020
Typeset by Jiří Hana
First edition

ISBN 978-80-246-4759-3
ISBN 978-80-246-4765-4 (online: pdf)



Univerzita Karlova
Nakladatelství Karolinum

www.karolinum.cz
ebooks@karolinum.cz

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 <i>ASK</i>	32
2.3.2 <i>CLC</i>	33
2.3.3 <i>COPLE2</i>	34
2.3.4 <i>CroLTeC</i>	34

2.3.5	<i>Falko</i>	35
2.3.6	<i>ICLE</i>	36
2.3.7	<i>MERLIN</i>	36
2.3.8	<i>RLC</i>	37
2.3.9	<i>SweLL</i>	38
2.4	Relationships of <i>CzeSL</i> with other learner corpora	39
3	Introducing the CzeSL project	41
3.1	Specifications of <i>CzeSL</i>	42
3.2	Intended usage	43
3.3	<i>AKCES</i> – 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 <i>CzeSL</i>	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 <i>CzeSL</i> corpora	155
8.1	<i>CzeSL-plain</i> – without annotation and metadata	157
8.2	<i>CzeSL-SGT</i> – with automatic annotation	158
8.3	<i>CzeSL-man</i> – with manual annotation	163
8.3.1	<i>CzeSL-man v0</i>	163
8.3.2	<i>CzeSL-man v1</i>	163
8.3.3	<i>CzeSL-man v1 downloadable</i>	164
8.3.4	<i>CzeSL-man v1 searchable</i>	165
8.3.5	<i>CzeSL-man v2</i>	167
8.4	<i>CzeSL-TH</i>	168
8.5	<i>CzeSL-MD</i>	168
8.6	<i>CzeSL-UD</i>	169
8.7	<i>CzeSL-GEC</i> and <i>AKCES-GEC</i>	169
8.8	<i>CzeSL in TEITOK</i>	170
8.9	Learner corpora of native Czech	170

9	Tools	173
9.1	Annotation tools	173
9.1.1	<i>feat</i>	173
9.1.2	<i>Speed</i>	174
9.1.3	<i>brat</i>	175
9.1.4	<i>TrEd</i>	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	<i>SeLaQ</i>	177
9.2.2	<i>Sketch Engine</i> and <i>KonText</i>	179
9.2.2.1	Token-based error annotation	180
9.2.2.2	Error annotation using structures	183
9.2.3	<i>TEITOK</i>	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 <i>Czech National Corpus</i> 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

List of abbreviations

2T	two-tiered annotation
CA	contrastive analysis
CAE	computer-aided error analysis
CALL	Computer-Assisted Language Learning
CC	Common Czech
CEFR	Common European Frame of Reference (for Languages)
CNC	Czech National Corpus
CQL	Corpus Query Language
DDL	data-driven learning
EA	error analysis
EFL	English as a foreign language
ELT	English language teaching
ESL	English as a second language
FLT	foreign language teaching
GDPR	General Data Protection Regulation
HTML	Hypertext Markup Language
IAA	inter-annotator agreement
IL	interlanguage
L1	first (native) language
L2	second (foreign) language
lit.	literally
MD	multi-dimensional annotation
NLI	natural language identification
NLP	natural language processing
NP	noun phrase
OOV	out-of-vocabulary error (non-word)
PDT	Prague Dependency Treebank

POS	part of speech
SCz	Standard Czech
SLA	second language acquisition
T0	Tier 0, the source text tier
T1	Tier 1, the tier of the intermediate target hypothesis
T2	Tier 2, the tier of the final target hypothesis
TH	target hypothesis
UD	Universal Dependencies
WSD	word sense disambiguation
XML	Extensible Markup Language

Chapter 1

Introduction

1.1 About this book

The story told in this book began more than ten years ago with the idea to collect Czech, written and spoken by native and non-native learners alike. The aim was to assist experts in teaching Czech as a foreign or native language, but also researchers in the acquisition of Czech. Learner corpora for other languages than English were still quite rare in those days, but the first release of a reference corpus and a treebank of Czech had been available at least since 2000.¹ This is why the project was concerned with how to use the methods and tools available for building and using standard corpora of native Czech. Another concern was how to adapt approaches used in learner corpora of other languages, given the specifics of Czech as a language with rich morphology and free word order. This book reflects these concerns in its focus on annotation, data formats and tools used for building and using the corpus.

Throughout the years, we have tried and used various solutions. Reports on the achievements and failures are now scattered over a number of papers. We believe it is high time to paint a more orderly picture: remedy inconsistencies, update some claims and figures, and present new, yet unpublished research.

However, cleaning up some mess is not a reason good enough to write a book. Our main aim is to introduce the project, presenting various approaches to the design of a learner corpus, including methods of collecting and transcribing texts, annotating errors and linguistic categories, and applying computational tools. It could be that some of our experience, positive or negative, may be relevant for other

¹*Czech National Corpus – SYN2000 (2000)*, *PDT – Prague Dependency Treebank 1.0 (2000)*, and Hajič et al. (2018)

non-native languages than Czech or for other types of non-standard Czech. This is why this book would not be complete without a chapter on the lessons learned.

Who could be interested? The target audience may include anyone interested in fields such as building and using learner corpora, teaching Czech as a foreign language, second language acquisition explored via corpus evidence, or natural language processing of non-standard language.

1.2 Reasons to study non-native Czech

During the last 30 years, the number of non-native speakers of Czech living in the Czech Republic has increased significantly. Currently, foreigners constitute at least 5.3% of the population (more than 16% in Prague), a significant increase from 2.5% in 2004.² Most of them come from Ukraine, Slovakia and Vietnam, followed by Russia, Poland, Germany, Bulgaria and Romania. However, in addition to tourist visitors, many EU citizens who live and work in the country, are not registered. Although the percentage of immigrants is much lower than in most other European countries, the sharp upward trend is obvious.

The Czech language has thus become a multi-ethnic communication tool, which is evident particularly in elementary and secondary schools. Many foreigners, both in schools and businesses, are learning Czech. Czech as a foreign or second language (L2)³ is also taught abroad – e. g., in the academic year 2019/2020, the Czech government supported programs at 36 universities and other institutions in 24 countries⁴ and offered courses in most of the 23 Czech Centres abroad, including courses for children.⁵

This situation brings new challenges. In comparison with a native language, L2 Czech – like any other L2 – is a fairly varied and volatile object. At each stage of learning the language, Czech of each non-native speaker has its specific

²Figures from 2018; see Boušková et al. (2019) or <https://www.czso.cz/csu/cizinci/cizinci-pocet-cizincu>. For even more up-to-date and slightly higher figures see <https://news.expats.cz/weekly-czech-news/number-of-foreigners-residing-in-the-czech-republic-is-rising/>. Note that these figures include only foreigners that registered with the Czech government, i. e., they exclude (i) illegal immigrants, (ii) EU nationals that did not register (the registration is optional), (iii) foreigners that reside in the Czech Republic but are registered in another EU country.

³In second language acquisition (SLA), foreign and second language are different terms. A foreign language is acquired in an environment where this language is not generally spoken, a second language is the language learned in a natural environment. Here, we use the two terms as synonymous. See §2.1 for more about the concept of L2 as used in this book.

⁴See <https://www.dzs.cz/cz/program-podpory-ceskeho-kulturniho-dedictvi-v-zahranici/prehled-lektoratu-a-lektoru/>.

⁵See <https://www.czechcentres.cz/en/about-us/sit-cc/>.

vocabulary and grammar structure. This is why our approach to L2 is based on the theory of interlanguage (IL) as a dynamic system, consisting of developmental stages, through which the learner passes at various stages of proficiency (Selinker 1972). The comparison of non-native and native linguistic production reveals that the utterances of learners form a distinct linguistic system (Tarone 2006). Such a system has been shown to underlie the seemingly random variety of errors made by non-native learners, adults and children alike (see, e.g., Dušková 1969; James 1998). Patterns of these errors depend on several factors including the speakers' native language, other languages they might know, the stage and ways of learning the language, etc. Investigating this system is beneficial both to the study of second language acquisition (SLA) and as a stimulus for the development of teaching methods, instructional materials and software tools for Czech as L2.

Non-native dialects of some other languages, such as English, German or French, are investigated more profoundly.⁶ Some results of research on non-native languages with richer data resources and a longer research history apply also to non-native Czech (sources of errors, native language influence etc.). However, with its rich inflectional morphology and word order reflecting information structure Czech is typologically different, and many questions about the acquisition of Czech cannot be answered by research concerning a language such as English. On the other hand, research on non-native Czech may be relevant not only to issues of SLA in similar languages (e.g., Slavic), but to SLA in general.⁷

In recent decades, research of Czech as L2 focused mainly on didactics: Czech was integrated in the Common European Framework of Reference for Languages (CEFR), the descriptions of referential levels were established,⁸ a number of new textbooks and teaching materials that reflect the CEFR levels were published, several grammar descriptions intended for non-native speakers were published (e.g., Hercíková 2009), and new university programs aimed at non-native Czech speakers were introduced.⁹ Still, issues of the acquisition of Czech and description of its L2

⁶See, e.g., Dickinson, Israel, and Lee (2010), Boyd et al. (2014), Zinsmeister, Heid, and Beck (2014), and Hirschmann et al. (2013).

⁷This can be useful, for example, for determining the acquisition order of phenomena expressing a communicative need such as request, but formed by very different means. Cf. the ease of forming the imperative mood in English vs. the relative difficulty of producing its Czech counterpart, depending on the morphological paradigm of the verb.

⁸As of 2020, the Czech Ministry of Education has published descriptions for the threshold level and the A1, A2 and B2 levels. See <https://www.msmt.cz/mezinarodni-vztahy/referencni-urovne-pro-cestinu-jako-cizi-jazyk> (in Czech). For a general description of the levels see <https://rm.coe.int/1680459f97>

⁹Such as Czech Studies for Foreigners at Charles University, Prague: <https://ubs.ff.cuni.cz/en/study/courses/bachelor-degree-course/> and <https://ubs.ff.cuni.cz/en/study/courses/>

varieties have not received enough systematic attention.¹⁰

To summarize, the increasingly stronger position of Czech as L2, the merely intuitive understanding of L2 Czech based on the teacher's experience and the persisting lack of modern didactic support for non-native speakers, including school children, make a strong case for a broadly conceived research, focused on those properties of non-native Czech which are significant, representative, identifiable, amenable to processing by formal tools, and comparable across learner texts of all types.¹¹ Therefore, research in non-native Czech is important for both theoretical and practical reasons:

1. Like every IL, non-native Czech calls for identifying *developmental patterns* and *orders of acquisition*. In addition to the research on the sequence of acquisition of L1 structures, there are studies about the acquisition of specific aspects of L2, such as morphemes, pronouns and word order (e. g., Ellis and Barkhuizen 2005) but not for an inflectional language such as Czech.
2. Analyzing IL is essential for SLA research, which helps to reveal how language works in general. IL contributes to the understanding of *linguistic universals* in SLA (White 2003). Investigating SLA of Czech is important because of its typological specifics.
3. Non-native language also offers data for studying *variability* as a key indicator of how a situation affects the learners' use of L2, either as free variations in the use of a language pattern which has not yet been completely acquired, or as systematic variations, determined by a linguistic, social or psychological context.

Studying these phenomena helps to understand the development of learners' IL while offering comparison with the acquisition of L1. The investigation of IL is worthwhile also in order to find out what types of errors learners make and what the errors say about their knowledge of target language and their ability to use it. This is important especially for didactic purposes: Czech should be described with regard to non-native speakers, and methods for teaching Czech to foreigners need to be elaborated and, i. a., translated into curricula (reflecting the relative difficulty of acquisition of individual phenomena). Analyzing IL is crucial also for language testing (individual features of IL need to be related to standard proficiency levels).

the-follow-up-master-degree-programme/.

¹⁰For studies dealing with the presentation of specific linguistic phenomena, see §10.

¹¹This applies especially to learners with a typologically distant L1, who are not acquainted with the European grammatical categories rooted in Latin.

New methodologies, based on extensive data and computational tools, help to advance this line of research. Although unsupervised methods can be used, a formal model must be based on the research of relevant aspects of IL. Its absence has also practical consequences. Many NLP tools that are taken for granted (spell checkers with suggestions, Internet search supported by morphology, machine translation, etc.) perform much worse for non-native Czech or are simply unusable, because experts developing applications for the native language cannot rely on previous research. Moreover, the study of L2 also has an intrinsic value in itself, as a study of a cultural phenomenon: L2 is part of the non-native speakers' identity.

1.3 Some properties of non-native Czech

Non-native speakers deviate from the standard language in non-arbitrary ways (see, e. g., Ellis and Barkhuizen 2005); the deviations are to a large extent systematic and predictable, but also evolving as learners receive more input and revise their hypotheses about L2 (R. Ellis 2003, 33–35). They are influenced by:

1. The speaker's native language (*interlingual* errors):

- *s matkoj* → *s matkou* 'with mother' – Russian ending *-oj* instead of Czech *-ou*
- *jsem vietnamský* → *jsem Vietnavec* 'I am Vietnamese' – adjective (as in English etc.) instead of a noun
- *žádný to ví* → *nikdo to neví* 'nobody knows it'; lit.: 'none it not-knows' – a single negative form *žádný* 'none' (as in English, German, etc.) instead of multiple negative forms *nikdo* 'nobody' and *neví* 'not-knows'

2. The general properties of the process of acquisition (*intralingual* errors):

- *v neděli spám dlouho* → *v neděli spím dlouho* 'I sleep late on Sundays' – misuse of endings from another inflectional class: *znát* 'to know' – *znám* 'I know' vs. *spát* 'to sleep' – *spím* 'I sleep', a case of "false analogy"
- *tady jsou pět stoly* → *tady je pět stolů* 'there are five tables here', lit.: 'there **is** five tables.GEN here' – a case of "overgeneralization" from simpler quantifier-free patterns *tady jsou stoly* 'there are tables here'

3. The properties of the instructional process

Many deviations of non-native Czech as compared with Standard Czech (SCz) belong to the domains of morphology and morphosyntax.¹² This is why we focus on these levels and design a system of concepts capturing the deviations in a systematic, formal and linguistically motivated way. The concepts are supported by computational models. Various contrasts in the patterns of IL can thus be made explicit and linked to parameters such as stages of acquisition and differences due to linguistic backgrounds of the speakers.

1.3.1 Morphology

In Czech, as an inflectional language, the syntactic functions of words are mostly expressed by their form, whereas word order is to a large extent constrained by information structure. Thus the domain of morphology plays a key role in Czech and due to its relative complexity represents the main source of deviations from the standard. It also deserves attention for a practical reason: many tools for computational language processing assume that methodologies and resources concerning morphology are available.

To give an example, Czech nouns have seven cases, with distinct forms for singular and plural, which means that any noun may have up to 14 different forms, although in every declension paradigm some forms are identical due to syncretism in case and/or number. For nouns, there are 14 basic paradigms, and a larger number of paradigm subtypes. The paradigm *žen|a* ‘woman’ (the most frequent for feminine nouns) has 10 forms, e. g., *žen|ě* is the form for dative and locative singular. Also adjectives, pronouns, numerals and verbs have many paradigms and inflected forms. It is nonetheless not necessary to master the entire Czech inflectional system in order to successfully communicate in Czech. It is enough to know how to use the most frequent cases and verbal forms for the common paradigms. For example, the understanding of the sentence in (1) is not disrupted by the error *Úvalách* → *Úvalech*.¹³

- (1) *Oslavil jsem Vánoce se svými příbuznými v jejich domě v*
 celebrated AUX Christmas with self’s relatives in their house in
 **Úvalách* → *Úvalech*.
Úvaly.(LOC) Úvaly.LOC

¹²For comparison, see an overview of native Czech grammar in Appendix B.

¹³A parenthesized morphosyntactic category in the gloss, such as *.(LOC)*, denotes an intended use of the category in an incorrect form. For a list of all conventions used in the examples, including identification of the source text, see Appendix A.

‘I celebrated Christmas with my relatives at their home in Úvaly.’

(KAR_MD_020 ru A2+)

The error *Úvalách* → *Úvalech* is in the form of the locative plural of the name of a Czech town *Úvaly*, a plurale tantum: the case ending *-ách* used incorrectly instead of *-ech* is an existing ending, used to express the same morphosyntactic properties of nouns of another paradigm (*Roztokách* ‘Roztoky.LOC’). As the incorrect form denotes the same nominal case of the noun, it may be noticed as unexpected by a native speaker, but it will not hinder the understanding of the whole sentence. Another type of inflection error, found in (2), can make the sentence slightly less understandable.

(2) *V životě dávám přednost *rodinu* → *rodině*.

in life give.1SG precedence family.*ACC family.DAT

‘In my life, I prefer family.’

(HRD_AS_221 ja A2)

The form *rodinu* ‘family’ is a form for accusative singular of the noun *rodina*, in a construction where the dative form *rodině* is expected. The sentence is still understandable, as it is composed of only a few words, but the use of incorrect case makes it more challenging to be understood by a native speaker. When a completely random ending is used, unrelated to paradigm or case form, the understanding is even more disrupted.

1.3.2 Syntax

Most syntactic deviations are found in morphosyntax, often when forms are lexically determined, e. g., by valency as in (3) or subject to a principle of grammar, e. g., agreement, as in (4).

(3) S: *myslím *o tobě*
think.1SG about you

T: *myslím na tebe*
think.1SG on you

‘I’m thinking about you’

(DGD_L5_143 ru A1)

(4) S: **přišli pět studentů*
came.*PL.*MA five students.GEN

T: *přišlo* *pět studentů*
 came.SG.NEUT five students.GEN
 ‘five students came’

Other deviations include non-standard word order due to an inappropriate topic-focus articulation (information structuring), or due to a misplaced clitic, such as in (5), where *jsem* ‘am’ and *se* – reflexive particle – are both 2nd position clitics and should follow the first constituent *během studování na univerzitě* ‘during university studies’ in that order.

- (5) S: *během studování na univerzitě *se seznámil *jsem s Evou*
 during studying on university REFL met AUX.1SG with Eva
 T: *během studování na univerzitě jsem se seznámil s Evou*
 during studying on university AUX.1SG REFL met with Eva
 ‘during my university studies I met Eva’ (BLAH_DZ_001 ky B2)

Similarly, reflexive pronouns are often under-used, as in (6), where the possessive *moji* ‘my’ should be replaced by the reflexive possessive *svoji*.

- (6) S: *miluji *moji práci*
 love.1SG my work
 T: *miluji svoji práci*
 love.1SG self’s work
 ‘I love my work’ (TOD_P2_247 ru A2+)

Various types of errors can occur within the same sentence, as in (7).

- (7) S: *V rodině *byli *jsme pět – táta, *mátka, *dve *sestři a já.*
 in family be.*PL.*MA AUX.1PL five dad (mother) (two) (sisters) and I
 I
 T: *V rodině nás bylo pět – táta, matka, dvě sestry a já.*
 in family we.GEN be.3SG.NEUT five dad mother two sisters and I
 ‘We were five in my family – father, mother, two sisters and me.’
 (HRD_1S_197 en A2)

In (7), the three errors in morphology and morphonology (*mátka*, *dve* and *sestři*) are combined with an error in the agreement pattern involving quantified subject,

shown in (4). The quantified subject NP, agreeing with a verb form in the 3rd person neuter singular (like *přišlo* ‘came.3SG.NEUT’ in (4)), includes the genitive form of the first person plural pronoun (*nás* – in the source sentence assumed to be nominative and thus pro-dropped). On the other hand, the 1st person plural past tense auxiliary *jsme* is dropped in the target sentence, because there is no auxiliary in the 3rd person past tense.

1.3.3 Word segmentation

Inappropriate word segmentation is not a random phenomenon either. Words are often incorrectly split after prefixes homonymous with prepositions (*do psat* ‘in write’ instead of *dopsat* ‘finish writing’). Russian speakers influenced by their native language sometimes append reflexive pronoun to the verb (*smějuse* → *směju se* ‘I laugh’) or split verb and the negative particle (*ne studuju* → *nestuduju* ‘I don’t study’). Since clitics, such as some prepositions and short pronouns, form prosodic units with their host, speakers exposed primarily to spoken Czech might spell them incorrectly as one word, as in (8).¹⁴

- (8) S: **opálímse* *ale* **musímít* *opalovací krém abych*
 sultan.1SG+REFL but must.1SG+have sunscreen so-that-AUX.1SG
 **semse* *nespálil* *moc*
 AUX.1SG+REFL burned.NEG too much

- T: *opálím* *se* *ale* *musím mít* *opalovací krém abych*
 sultan.1SG REFL but must.1SG have sunscreen so-that-AUX.1SG
se *nespálil* *moc*
 REFL burned.NEG too much

‘I will get a sun tan but I must have a sunscreen so that I would not get sunburnt too much.’
 (ss_dp_057_63 cs 11)

1.4 Learner corpus

Investigating language acquisition by non-native learners helps to understand important linguistic issues and to develop teaching methods, better suited both to the specific target language and to specific groups of learners. These tasks can now be based on empirical evidence from learner corpora.

A learner corpus consists of language produced by language learners, typically learners of a second or foreign language (L2). Such corpora may be equipped with

¹⁴Example (8) is from *SKRIPT 2015*, a corpus of young native Czech learners. The text ID is followed by the code for Czech (cs) and the age of the author.

morphological and syntactic annotation, together with the detection, correction and categorization of non-standard linguistic phenomena.

Learner corpora allow to compare non-native and native speakers' language, or to compare interlanguage varieties, and can be studied on the background of standard reference corpora, which helps to track various deviations from standard usage in the language of non-native speakers, such as frequency patterns – cases of overuse or underuse – or *foreign soundingness* as compared with the language of native speakers. A range of studies have focused not only on the frequency of use of individual elements of language (e. g., Ringbom 1998), including phenomena such as negative and positive transfer, formulaic language, collocations (lexical patterns), prefabs and colligations (lexico-grammatical patterns, e. g., Nesselhauf 2005; Paquot and Granger 2012; N. C. Ellis 2017; Granger 2017; Vetchinnikova 2019), lexical analysis and phrasal use (e. g., Altenberg and Tapper 1998), but also developmental patterns, variability and the impact of the learning context (Meunier 2019; Granger, Gilquin, and Meunier 2015) and interlanguage complexity (e. g., Paquot 2019).

An error-tagged corpus can be subjected to *computer-aided error analysis* (CEA), which is not restricted to errors seen as a deficiency, but understood as a means to explore the target language and to test hypotheses about the functioning of L2 grammar. CEA also helps to observe meaningful use of non-standard structures of IL. Such studies focus on lexical errors (e. g., Leńko-Szymańska 2004), wrong use of verbal tenses (e. g., Granger 1999) or phrasal verbs (e. g., Waibel 2008).

The tasks of designing, compiling, annotating and presenting such corpora are often very much unlike those routinely applied to standard corpora. There may be no standard or obvious solutions: the approach to the tasks is often seen as an answer to a specific research goal rather than as a service to a wider community of researchers and practitioners.

The difference between a standard and a learner corpus is mainly in their annotation. Texts in a learner corpus can be annotated in two independent ways: (i) by standard linguistic categories: morphosyntactic tags, base forms, syntactic structure and functions, and (ii) by error annotation: correct version of each ill-formed part of the source text, i. e., its target hypothesis (TH), and categories specifying the nature of errors. Reasonably reliable methodologies and tools are available for linguistic annotation (i) of many languages, as long as the text is produced by native speakers. The situation is different for non-standard language of non-native learners and for error annotation (ii), where manual annotation is quite common. However, with the growing volumes of learner corpora, the need for methods and tools simplifying such tasks is increasing. Yet the annotation of learner corpora remains a challenging task, even more so for a language such as Czech, with its rich inflection, derivation, agreement, and a largely information-structure-driven

constituent order.

1.5 Roadmap

Chapter 2: Learner corpora provides some context by listing several properties which make each learner corpus different from any other. The second half of the chapter presents an overview of nine learner corpora with features relevant for the *CzeSL* corpus.

Chapter 3: Introducing the CzeSL project presents the foundations of the *CzeSL* project and outlines its main characteristics.

Chapter 4: Procurement of texts deals with the initial tasks in the compilation of the *CzeSL* corpora. It is the first in the sequence of three chapters concerned with how the texts are treated and what kind of annotation they receive. These chapters do not focus on the actual pre-processing, which is the topic of Chapter 8, but rather on the description of the principles, categories and formats.

Chapter 5: Error annotation presents the background and substance of several types of error annotation used in the *CzeSL* project. We focus on the original error annotation scheme, consisting of three parallel tiers for the source text and two tiers for its annotation. This type of error annotation is examined from several angles: we provide motivation behind this design, present the grammar-based and the “formal” error tagsets, complementing each other, and provide results of its evaluation in terms of inter-annotator agreement (IAA). The chapter follows by introducing two additional types of error annotation used in the *CzeSL* project more recently: annotation without explicit error tags, facilitating manual annotation, and a multidimensional scheme, complementing the original tiered system especially in the domain of morphology.

Chapter 6: Linguistic annotation examines the approaches adopted in *CzeSL* to the annotation of morphosyntactic categories, syntactic structure and functions. The chapter consists of three main parts: it starts with the methods analyzing the TH, proceeds to methods developed for standard language but used to annotate source learner texts, and concludes with the description of an approach to syntactic analysis designed specifically for learner Czech.

Chapter 7: Annotation process looks at the transcription, anonymization and annotation from the perspective of a step-by-step procedure, including decisions about the share of manual tasks and suitability of automatic tools. The various types of annotation described in the preceding chapters are here described in terms of input, processing and output.

Chapter 8: The *CzeSL* corpora provides an overview of searchable corpora or downloadable data sets containing the *CzeSL* texts. The various releases reflect the various approaches to the annotation, but they also differ in the choice of texts, availability of metadata and the data format, determining the search options, i. e., the choice of a suitable search tool.

Chapter 9: Tools is an overview of tools used within the *CzeSL* project for processing and annotating texts on the one hand and for searching and viewing them on the other.

Chapter 10: Using the corpus is concerned with how the *CzeSL* corpora are used in research and teaching of Czech as a foreign language, and also in NLP applications such as text scoring, text correction and natural language identification (NLI). Some of these types of use are closely related with the exploitation of standard reference corpora for the same purpose, which is why a section about the use of corpora of native Czech is also included.

Chapter 11: Lessons learned and perspectives concludes the core chapters of the book by discussing positive and negative experience from implementing various solutions throughout the project and by an outlook into the future.

Chapter 12: Acknowledgements should be seen as an important part of the book. There are many people and several funding agencies who deserve our credit for starting the project and for keeping the project alive throughout the years.

Appendix A: Notes about examples briefly summarizes the presentation of examples.

Appendix B: The Czech language presents an overview of Czech as a native language in its main features. This part may be useful especially for readers who do not speak or understand Czech.

Chapter 2

Learner corpora

Since the release of the *International Corpus of Learner English* (Granger, Dagneaux, and Meunier 2002), learner corpora have become a well-established branch of corpus linguistics. Now they are an important source of data for foreign language teaching, second language acquisition and other related disciplines (McEnery 2018; McEnery et al. 2019). The growing number of learner corpora of various kinds, formats and access options have also led to efforts aimed at making the data, methods and tools used in various projects reusable by trying to achieve some degree of conceptual and structural interoperability (Chiarcos 2012; Stemle et al. 2019).

2.1 Terminology

A learner corpus, also called interlanguage or L2 corpus, is a computerized textual database of language as produced by L2 learners (Leech 1998). A similar definition, where native language learners are excluded, is used by Granger (2008).

Although our topic is Czech as L2, we would prefer to treat as a learner corpus each corpus concerned with language acquisition, no matter whether the language represented by the corpus is L1 or L2. The reasons enumerated by Granger (2008), related to the blurring of L1 and L2 in the context of various dialects of English around the world, apply also to Czech and its varieties, such as its Romani ethnolect or dialects used by communities of heritage Czech speakers abroad. Moreover, some corpora may intentionally include texts produced by both non-native learners and native speakers of a language. Once we agree that young native speakers are also learners, such corpora, including both L1 and L2, should also be called learner corpora.

Perhaps a less controversial issue is what counts as L2. Here we use the term second language as denoting any language learned after the first/native language or mother tongue (L1). Thus we use it as a hypernym of foreign language, a second language one learns outside of the environment the language is spoken. According to this view, *CzeSL* includes data from non-native residents of the Czech Republic, including those staying for a relatively short period (e.g., 1 year), and also from students of Czech abroad. Some authors (e.g., R. Ellis 1994) use the terms second and foreign language in the same way as we do here, while others (e.g., Günther and Günther 2007) use them as complementary.

In the domain of L2 acquisition and teaching of foreign languages, the language of the learners is called *interlanguage* (Selinker 1983). Interlanguage (IL) is distinguished by its highly individual and dynamic nature. It is subject to constant changes as the learner progresses through successive stages of acquiring more competence, and can be seen as an individual and dynamic continuum between one's native and target languages. An interlanguage includes both correct and deviant forms. The possibility to examine learners' errors on the background of the correct language is the most important aspect of learner corpora (Granger 1998a).

2.2 Various types of learner corpora

Learner corpora can differ in many ways (see, e.g., Granger 2008, 260). Here we list a few distinguishing features which are relevant to the *CzeSL* project, without attempting to provide an exhaustive inventory of all items describing a learner corpus.

Similarly, we make no attempt to list all available learner corpora in the rest of the chapter.¹ Instead, we focus on corpora with some interesting properties, related in one way or another to our project.

2.2.1 The choice of texts

Corpora can be classified according to the nature of their content in several dimension:

¹For an extensive overview of learner corpora see the actively maintained list at the Centre for English Corpus Linguistics (Université catholique de Louvain): <https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html>

Medium. Learner corpora can capture written or spoken texts, the latter much harder to compile, thus less common. The texts can be born-digital or transcribed from manuscripts.²

First language (L1). The data can come from learners with the same L1 or with various L1s.

Target language (L2). Most learner corpora cover the language of learners of English as a second or foreign language (ESL or EFL). The number of learner corpora for other languages is smaller but increasing.

Proficiency in target language. Some corpora gather texts of students at the same level, other include texts of speakers at various levels. Most corpora focus on advanced students.

Cross-sectional vs. longitudinal data. Most L2 corpora are cross-sectional, gathering data from various types of learners. Only few L2 corpora are longitudinal (developmental), including data acquired over time from the same learners. Several learner corpora collect balanced data from homogeneous groups of learners at different levels of L2 knowledge and are used as quasi-longitudinal learner corpora.

2.2.2 Annotation

Some learner corpora contain only raw data, but most of them add some textual, linguistic or error annotation at least for a part of their content (Fitzpatrick and Seegmiller 2001; Granger 2003a; Abuhakema, Feldman, and Fitzpatrick 2009; Granger, Gilquin, and Meunier 2015). At least to some extent, error annotation is usually the task of human annotators.

After an overview of the types of annotation common in learner corpora we focus on the more formal aspect of annotation – the design of the annotation scheme.

2.2.2.1 Textual annotation

In a standard written corpus, textual annotation encodes the structure of a text, e. g., in terms of a markup identifying sentences, paragraphs, sections, headings or

²Manuscripts can also be OCRed, but the technology may still not be reliable enough for the task.

footnotes, and its typographical properties, such as boldface or italics. Historical, learner and other specialized corpora may require a more sophisticated textual annotation to encode some relevant aspects of handwritten documents, such as additions, deletions, corrections and other textual elements.

Instead of symbols representing the elements in a markup designed specifically for a given corpus, the common way nowadays is to use a standard such as TEI XML. Even though this approach usually requires the use of specific tools and procedures during the (often manual) transcription phase, it makes all downstream processing of the data easier and enables compatibility and data exchange with similar text resources.

2.2.2.2 Linguistic annotation

At least some parts of most learner corpora are annotated in ways similar to standard reference corpora, i. e., at least by POS, lemma and morphological categories, sometimes also by syntactic structure, syntactic function or even named entities. Such annotation is usually done by automatic tools originally developed for the analysis of the native language and trained on standard texts produced by native speakers. This is a straightforward task when the tools are applied to a normalized version of the learner texts. However, for texts produced by the learner, the result very much depends on how much the text deviates from the standard language.

Besides the practical concerns about a higher error rate there is also a conceptual issue: linguistic categories used to annotate standard native language may be ill-suited to L2. Often it is not obvious what kind of annotation an incorrect expression should receive. See §6.

2.2.2.3 Error annotation – correction

Together with error categorization, correction of erroneous text is one of the two types of error annotation. Error correction is also called normalization, reconstruction, emendation, interpretation or the assignment of target hypothesis (TH). Establishing a hypothesis about the author's intention and its expression may be far from straightforward. This is why some annotation schemes do not force the annotator (or a correcting tool) to always pick a single TH but instead provide the option of alternative THs. However, alternative THs bring additional complexity in downstream processing and data formats. Due to such practical reasons they are rather rare.

Multiple THs may be desirable also when a single word form or phrase is incorrect for a number of reasons: spelling, mismatch between the word root or stem and

the derivation or inflection suffix, wrong word class or wrong inflection in a given syntactic context, wrong choice of a lexeme, inappropriate construction or inadequate word order. Instead of providing a single correction for all possible types of erring, the annotation scheme may provide space for successive corrections along a sequence of error types according to levels defined by the grammar. In a scheme designed consistently in this way, each correction can be paired with a corresponding error category.

Correction can also be used as the sole component of error annotation. Error categorization can then be viewed as implicit in the target hypothesis, especially when the scheme allows for successive corrections. The advantage of this approach is the absence of an error classification scheme – the annotator does not need to learn any classification rules, which speeds up the annotation task and avoids misclassification. See §5.5 for more on implicit error annotation.

2.2.2.4 Error annotation – categorization

Error categorization involves annotation of errors with categories from a predefined error taxonomy. While every error taxonomy reflects its theoretical background, categorization can be very useful for searching and statistical investigations. Error-tagged corpora may use one or more of the following types of taxonomies to classify the type of error:

- Linguistically-based taxonomies, with a varying degree of detail, ranging from general categories (morphology, lexicon, syntax) to specific labels (auxiliary, passive, negation).
- Taxonomies based on a formal classification of surface alternations of the source text, such as missing, redundant, faulty or incorrectly ordered element.
- A combination of several taxonomies, e. g., a multi-dimensional scheme consisting of an error domain (formal, grammar, lexicon, style), an error category (agglutination, diacritics, inflection, derivation, gender, mode), and word class (POS).

Despite the time-consuming manual effort involved, the number of error-annotated learner corpora is growing. However, the level, extent and concept of error annotation differ.

2.2.2.5 Annotation scheme

If we compare learner and standard corpora in terms of how they are annotated, error annotation of at least some learner corpora represents the major difference. Obviously, there are many more resources for textual and linguistic annotation than for error annotation, and guidelines such those developed by TEI are available even for highly complex textual and linguistic phenomena.

Error annotation is not the only neglected domain. Standard approaches to corpus annotation may fail also when linguistic annotation is combined with other than very basic textual annotation. E. g., few corpus search tools support both textual and linguistic annotation. The problem is due to the fact that the inline or stand-off XML-based format often used for textual annotation is at odds with the tabular format assumed by most search tools and also by tools providing linguistic annotation (taggers and parsers). Whereas the tabular (also called vertical) format is oriented towards annotating individual tokens rather than potentially discontinuous sequences of tokens or embedded structures, an XML scheme can accommodate any, even overlapping annotation (the latter in a stand-off manner). As a result, the textually annotated version of data is often detached from the linguistically annotated version.

Similar problems may occur even in some corners of linguistic annotation, including tokenization: the contraction *isn't* as a single orthographic word vs. its interpretation as two syntactic words *is* and *not*.

Yet it is typically easier to build a standard corpus than a learner corpus with error annotation. The problem is in the error annotation alone and also in its combination with the other annotation types. Here we list the possible choices facing someone designing a learner corpus, depending on her requirements for the error annotation.

- No error annotation in the presence of any other type of annotation
 ⇒ the same annotation scheme as in a standard corpus
- Token-based error annotation, i. e., error tags and/or corrections restricted as annotation of individual tokens, including multiple corrections of a single token (successive or alternative)
 ⇒ inline or tabular format; both can be used even if linguistic annotation is provided for the incorrect as well as corrected forms. The tabular token-based format is supported by the standard tools such as *Corpus Workbench – CWB* (Evert and Hardie 2011), *Corpuscle* (Meurer 2012), *Korp* (Borin, Forsberg, and Roxendal 2012), *Sketch Engine* or *KonText* (see §9.2.2) and is used in the *CzeSL* project for *CzeSL-SGT* (see §8.2) and *CzeSL-man v1* (see §8.3.2).

- Error annotation spanning contiguous strings of tokens, as in incorrectly split or joined word forms, including multiple corrections
 ⇒ inline XML – *SKRIPT 2015* (see §8.9) and *CzeSL in TEITOK* (see §8.8), or tabular format with structures used for error annotation – *CzeSL-man v2* (see §8.3.5), or multi-tier tabular format – *Falko* (see §2.3.5) and *MERLIN* (see §2.3.7),³ supported by tools such as *EXMARaLDA* (Schmidt 2009; Schmidt et al. 2011).
- Error annotation of discontinuous strings, including multiple corrections
 ⇒ multiple parallel tiers – *CzeSL-man v0* (see §8.3.1), or stand-off XML – *CzeSL in TEITOK* (see §8.8). The multi-tier format is supported by the *PAULA* format,⁴ the *SVALA* tool/format,⁵ and by tools around the PML format such as *feat*, *SeLaQ* and *PML-TQ*.⁶
- Error annotation of segments shorter than a word, i. e., of morphs or characters, even across word boundaries; also when several morphs annotated as a whole form are not adjacent within a word or across word boundaries
 ⇒ *CzeSL-MD* (see §8.5), annotated in *brat* (see §9.1.3), searchable and editable in *TEITOK* (see §9.2.3).

2.2.3 Data access

Some learner corpora are available under an open license for on-line searching or even for download as full data sets, other are accessible with some restrictions, from the condition of academic use or the payment of a license fee, to the exclusive right of access to the staff or collaborators of a publisher (in the case of proprietary corpora). Sometimes such a corpus is available in part and/or with impoverished annotation.

Especially when the corpus is freely accessible, approval of the authors and/or the teaching or testing institution is important, and any sensitive information should be removed. Names, addresses, phone numbers are anonymized, i. e., replaced by codes, or pseudonymized, i. e., replaced by other word forms, usually taken from a list of words with similar usage properties and morphological paradigms. Pseudonymization may be preferable e. g., when a specific type of error in the original word form (such as missing capitalization) should be preserved in the published

³In *Falko* and *MERLIN*, word-order corrections are also done in the multi-tier tabular format, however, then the correspondences between tokens across the tiers may be lost.

⁴See <https://www.sfb632.uni-potsdam.de/en/paula.html>; Zeldes, Zipser, and Neumann (2013).

⁵See §2.3.9; <https://github.com/spraakbanken/swell-editor>; Wirén et al. (2019).

⁶See §9; <https://ufal.mff.cuni.cz/pmltq>

version. In some projects, handwritten texts or audio recordings are not disclosed to every corpus user to protect the learner’s identity.

To some extent, the choice of query interface is determined by the data and annotation format. However, for many users the search tool is the only window to the corpus, so the options of querying and visualization offered by the search tool are crucial. For example, some users may prefer a tool such as *TEITOK*, which is able to display the text together with its annotation and properties of the handwritten source at the same time. Other users need an interface with a rich menu of statistical functions such as *KonText* or a tool combining the search and annotation environments (*Korp* and *SVALA*, *TEITOK*).

2.3 Some learner corpora

This is a very partial overview of some currently available learner corpora. They were selected because some of their features are related in one way or another to the *CzeSL* project.⁷

2.3.1 *ASK – Norsk andrespråkskorpus*⁸

The corpus of Norwegian as second language, developed in 2006–2014, consists of transcripts of essays, hand-written by learners who had passed the higher level test in Norwegian for adult immigrants. The texts (1,936 items, 770 thousand words, 1,130 thousand tokens) were selected to achieve typological diversity in L1s: German, Dutch, English, Spanish, Russian, Polish, Bosnian-Croatian-Serbian, Albanian, Vietnamese and Somali. The corpus also contains texts from native Norwegians as control data.

The texts and metadata are marked up in XML according to the TEI Guidelines.⁹ The texts were typed in and validated using a standard XML editor (*Oxygen*).

For error annotation, the TEI guidelines are extended by the attributes *corr* (corrections) and *sic* (errors), *type* (error category) and *desc* (subcategory). The

⁷For a more exhaustive overview of learner corpora see, e. g., Pravec (2002), Nesselhauf (2005), Štindlová (2011, 2013), and Xiao (2008), or more up-to-date lists at <https://www.uclouvain.be/en-cecl-lcworld.html>, and <https://www.clarin.eu/resource-families/L2-corpora>.

⁸<https://clarino.uib.no/ask/>, Tenfjord, Meurer, and Hofland (2006) and Tenfjord, Hagen, and Johansen (2009). The methodology and infrastructure of *ASK* were also used to build a pilot learner corpus for Slovene (*PiKUST*, Stritar 2009).

⁹In this respect, *ASK* preceded the learner corpora available in *TEITOK* (see §9.2.3), including *CzeSL in TEITOK*.

sic tags can be used recursively to mark up more than one error in a word or phrase.

The error tagset is rather small in order to avoid inconsistencies in the error coding and redundancy due to the presence of POS tags.¹⁰ The tags are of the following seven types: lexical (lexeme, spelling, foreign word, word boundary, capitalization, derivation), morphology (category, paradigm), syntax (missing or redundant word or phrase, word order: inversion, adverbial), punctuation, uninterpretable, follow-up.

Error annotation of the source texts is complemented by linguistic annotation using a tagger for standard Norwegian, with a facility for manual tag correction. The same tagger is applied also to the corrected texts. The source texts and their corrected versions are aligned as a parallel corpus. They can be searched and corresponding sentences displayed in parallel using *Corpuscle*, a corpus query engine and web-based corpus management system (Meurer 2012). The corpus is available under the CLARIN Res (Priv) license.¹¹

2.3.2 CLC – Cambridge Learner Corpus¹²

CLC is an English learner corpus built and used by Cambridge University Press as a proprietary resource and a part of the *Cambridge English Corpus*.¹³ The texts are collected from learners taking one of the various types of Cambridge English Language Assessment exams in English: general, academic, business, legal, finance, or life skills.

The whole corpus consists of 55 million words. The corpus is tagged and lemmatized, and about one third is error-annotated with a tagset of nearly 90 tags.¹⁴ Authorized users can search the corpus using *Sketch Engine* (see §9.2.2) with all its functionalities, including Word Sketches. Error annotation is implemented as pairs of XML structural elements `err` and `corr`, representing an incorrect form and its correction (see §9.2.2.2). E. g., to find all errors in incorrect verb tense associated with past participles the user should use the following query:

¹⁰Both reasons were also behind the decision to use a relatively small tagset in the manual tiered annotation of the *CzeSL* corpus. This *CzeSL* tagset consists of 26 tags.

¹¹For details of the license see <http://urn.fi/urn:nbn:fi:lb-2019071729>.

¹²<https://www.cambridge.org/sketch/help/>; Nicholls (2003).

¹³The other part is the *Cambridge Reference Corpus*, consisting of 2 billion words of native English, both written and spoken.

¹⁴For a list of the CLC error tags see https://www.cambridge.org/sketch/error_codes_grouped.html.

[tag="VVN"] within <err type="#TV"/>.¹⁵ Alternatively, a simple error query interface can be used to search for the source word forms, error tags and corrections. The corpus metadata include L1, nationality, exam, CEFR level, year, educational level, age, years of English study, gender, pass or fail.

A part of the corpus is accessible without error annotation as one of the *Sketch Engine* corpora under the name *Open Cambridge Learner Corpus (Uncoded)*.¹⁶ This corpus consists of 11.5 thousand texts consisting of 3 mil. words from learners with 7 different L1s.

Apart from its use in the publishing house for creating methodologies, textbooks and other English Language Teaching (ELT) materials, the corpus has also been used for creating the English Vocabulary Profile.¹⁷

2.3.3 *COPLE2 – Corpus de Português Língua Estrangeira / Língua Segunda*¹⁸

COPLE2 is a corpus of written and spoken texts produced by students of Portuguese as L2 and by applicants for exams in Portuguese, built since 2013. The corpus contains about 1,100 texts (230 thousand words) from learners with 15 different L1s and proficiency levels from A1 to C1, and covers different topics and tasks.

The corpus is in the TEI format, built, maintained and searchable in the *TEITOK* environment (see §9.2.3). Together with *CroLTeC*, this corpus served as a model for *CzeSL in TEITOK*.

The metadata include the L1, CEFR level, months of studying Portuguese, nationality, knowledge of other foreign languages, text genre, topic and text type. Manuscripts or oral productions are also available. The transcripts encode modifications by the student and the teacher. The corpus is annotated for POS, lemma, TH and error type.

2.3.4 *CroLTeC – CROatian Learner TExt Corpus*¹⁹

CroLTeC consists of essays written in weekly intervals as a part of a course in Croatian, collected since 2016 from 755 non-native learners of Croatian at all levels

¹⁵The error annotation based on the XML structural elements combined with the tabular (vertical) taken-based format has been adopted also in one of the *CzeSL* corpora (see §8.3.5).

¹⁶<https://www.sketchengine.eu/cambridge-learner-corpus/>

¹⁷<http://vocabulary.englishprofile.org>

¹⁸<http://teitok.clul.ul.pt/learnercorpus/>; Mendes et al. (2016), Rio et al. (2016), and Rio and Mendes (2019).

¹⁹<http://teitok.clul.ul.pt/croltec/>; Preradović, Berać, and Boras (2015).

of proficiency with 36 different L1s. The size of the corpus is 1 million words. About 3.5 thousand texts were hand-written and transcribed, 1.2 thousand texts were digitally born.

Like *COPLE2*, *CroLTeC* uses the *TEITOK* environment (see §9.2.3), which means that the corpus can be extended, modified, annotated and otherwise improved while being available for online searching at the same time.

The transcripts encode corrections made by learners themselves (deletions, insertions and word order changes). The texts are POS tagged and lemmatized, hand-corrected and assigned error tags.

Metadata include gender, age, nationality, mother tongue, bilingual and multi-lingual competence, parents' language proficiency, required linguistic competence for the task, genre, scope, time limit, size limit and the task circumstances (home-work, part of an exam, field work, etc.).

2.3.5 *Falko – Ein fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache*²⁰

Falko, built since 2004, contains 641 texts (about 380 thousand words) written by non-native learners of German, complemented by 152 texts (about 92 thousand words) in its comparative native German section.

The L2 part alone comprises several sub-corpora: text summaries, essays written by advanced learners, and a longitudinal corpus from learners with different proficiency levels. The comparative part includes texts for each of the non-native sub-corpora.

All annotation is strictly stand-off, each type in a separate tier. The tiers for POS and lemmas are available for all texts. Error annotation, consisting of TH and error tags, is available only in some sub-corpora. Additional tiers can be added at any time, which means that alternative THs are possible in addition to successive THs. For the essay subcorpus, alternative THs are available, tagged for POS and lemma: “minimal” – grammatically correct and “maximal” – approaching the standard native language. Error tags, annotating differences between a TH and the source text, are represented as separate tiers.

Unlike the concept of parallel tiers in *CzeSL* (see §5.4), which allows for any reordering of words at the neighboring tiers while preserving the cross-tier links between corresponding (even non-contiguous sequences of) words, the tiers in *Falko* can be represented as rows in a table with columns standing for the cross-tier links.

²⁰<https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko>; Reznicek et al. (2012).

For any word order corrections, the cells for the word order region must be merged (horizontally) at the TH tier, which means that the cross-tier links between the individual words are lost. In an extreme case of a region afflicted by the need to correct an error in word order spanning an entire sentence, the whole sentence may end up as a single column. The tabular format is due mainly to the annotation tools²¹ rather than to the format or the search tool.

The corpus is available under the CC BY 3.0 license and can be searched using the powerful *ANNIS* tool.²²

2.3.6 *ICLE – The International Corpus of Learner English*²³

The *ICLE* project, launched in 1990, includes essays written by university students of English mainly in their second or third year. In 2002 the corpus was released as a CD-ROM accompanied with a handbook. *ICLE* was the first academic learner corpus of a considerable size and is still seen as the paradigm of a methodologically mature approach to the design of the content of a learner corpus.

ICLE v3, the latest, web-based and on-line searchable version, published in 2020, includes over 9 thousand essays (5 million words, the length of each between 500 and 1,000 words), written by learners from 26 mother tongue backgrounds. The corpus is balanced in terms of the share of various L1s. There are 14 metadata items about the learner and 7 items about the task. The texts are tagged and lemmatized, but they are without error annotation.

Besides a trial version with some restrictions, the full version allows the download of entire texts.²⁴

2.3.7 *MERLIN – Multilingual Platform for European Reference Levels: Interlanguage Exploration in Context*²⁵

MERLIN, built in 2012–2014, consists of 2,286 texts (340 thousand words) from learners of three languages: German (1,033 texts), Italian (813 texts) and Czech (442 texts, 64.5 thousand words). The texts come from written exams of acknowledged test institutions, aiming to test knowledge across the CEFR levels A1–C1. The corpus is tagged, lemmatized, parsed and on-line searchable using a custom

²¹ *Falko* add-in for *Microsoft Excel* or *EXMARaLDA* <https://exmaralda.org/en/>

²² <https://korpling.german.hu-berlin.de/falko-suche/>

²³ <https://corpora.uclouvain.be/cecl/icle/>; Granger (1998b, 2003b).

²⁴ The license is available for a fee or to institutions that are members of the eduGAIN interoperation (<https://edugain.org/>), using the Shibboleth log-in system.

²⁵ <https://www.merlin-platform.eu/>; Wisniewski et al. (2014) and Boyd et al. (2014).

platform, based on *ANNIS*, with a detailed error taxonomy and the option of two target hypotheses (minimal and extended, similar to *Falko*).

There is a specific motivation behind the project. *MERLIN* is meant to provide examples of authentic texts for the individual CEFR levels in order to highlight the distinctions on the basis of comprehensive empirical characteristics. Moreover, some of the error tags are designed to check whether the CEFR descriptors, concerning a language and a CEFR level, correspond to the way learners actually use the language.

In addition to the CEFR specifications, the design of the error tagset is based on issues in SLA research, features reported by experts in teaching, analyses of textbooks, language tests and learner texts. In fact, each of the tags is labeled for its source.

Throughout the corpus creation process, the strategy was to reuse existing methodologies, formats and tools, resulting in a combination of a number of tools, many of them adopted from the *Falko* project.²⁶

The corpus is available under an open license (CC BY-SA 4.0). In addition, the project approach and computational architecture is designed to be adaptable to other languages for which CEFR level illustration is needed.

2.3.8 *RLC – The Russian Learner Corpus*²⁷

As of 2016, *RLC* was a collection of 2,000 texts produced by learners of Russian as L2 and 1,500 texts by speakers of heritage Russian with various dominant languages, altogether 730 thousand tokens. The texts include academic writings, movie and picture descriptions, book summaries, expository essays and others. A part of the corpus are speech transcripts. Some texts constitute a longitudinal subcorpus of academic writing.

The corpus is annotated by morphological tags and lemmas, and includes two tiers of error annotation, based on deviations from Standard Russian: formal corrections (spelling, case forms, gender/number agreement, tense and aspect) and lexical/constructional violations.²⁸ There are 59 error tags²⁹ for errors in spelling (6), morphology (6), syntax (2), constructions (1), lexicon (5), and 7 supplementary tags (combined with the above tags). There are 10 metadata items for each text.

²⁶See Stemle et al. (2019) for an overview of the *MERLIN* strategy.

²⁷<http://web-corpora.net/RLC>; Rakhilina et al. (2016). The corpus can be searched from <http://web-corpora.net/RussianLearnerCorpus/search/>.

²⁸The two annotation tiers in *RLC* resemble the two-tiered error annotation scheme of *CzeSL*. However, the range of errors annotated at Tier 1 is larger in *RLC*.

²⁹See <http://www.web-corpora.net/RLC/help>.

2.3.9 *SweLL – research infrastructure for Swedish as a second language*³⁰

The aim of the *SweLL* project (2017–2020) is to provide methods and tools for processing learner texts and to build a corpus of L2 Swedish consisting of about 600 texts. Some of the texts are transcribed from manuscripts and some are digitally born. The results include a portal for data collection via file import and online exercises. Handling of sensitive data is a priority – all texts are anonymized or pseudonymized according to precise rules.

Error annotation is done using *SVALA*, an annotation editor developed within the *SweLL* project (Volodina, Matsson, et al. 2019). In a way, the editor is similar to *feat* (see §9.1.1): there is a tier for the source text and another parallel tier for its corrected version (the TH) with links across the tiers connecting corresponding tokens. Error tags label links with a correction. The display with a sequence of vertical links connecting tokens on the two tiers is called spaghetti mode. In a text where the source and the target tokens correspond 1:1, the aligned spaghetti are straight, uncooked. Like in *feat*, there may be more than one or even no corresponding token on either of the two tiers, and a link may cross other links when the word order changes, resulting in a cooked spaghetti (curly and/or split). It is the task of the annotator to correct the TH tier, edit the alignment links and add error tags. The source and target substrings, corrected within a single form, are highlighted.

There are altogether 36 error tags of five main types: orthographic (3), lexical (4), morphological (8), punctuation-related (4), and syntactic (11). The “Other” type (6) includes a tag for follow-up (“consistency”) and unidentified corrections, intelligible and foreign strings, and comments (internal and for the corpus user). Although the tagset is not too large, it specifies some error types with respect to a more detailed grammatical category: e. g., morphological errors include tags for errors in case, definiteness, gender and number. Annotating a single error by a combination of tags is allowed, e. g., for an error in orthography and morphology, lexicon and syntax, or morphology and syntax.

In addition to POS and lemmas, linguistic annotation includes syntactic parse and word-sense disambiguation. The plans include experimental linguistic annotation of the source texts to obtain a parallel treebank. The corpus can be searched using the general *CWB*-based *Korp* tool.³¹ To see the annotation in the spaghetti mode, a click takes the user to the *SVALA* editor with the text including the concordance line.

³⁰<https://spraakbanken.gu.se/en/projects/swell>; Volodina et al. (2016) and Volodina, Granstedt, et al. (2019).

³¹<https://spraakbanken.gu.se/en/tools/korp>

The corpus annotation will be available under the CLARIN RES (Priv) licence.

2.4 Relationships of *CzeSL* with other learner corpora

Each of the learner corpora briefly described above was selected for a reason. Some of the projects, such as *ICLE* or *Falko*, were crucial by providing inspiration for the design and development of *CzeSL*, while other projects are noteworthy because *CzeSL* shares some important features with them.

The concept of parallel annotation tiers in *Falko*, supporting alternative and successive THs and implemented in the stand-off fashion, was at the origin of the tiered annotation scheme of *CzeSL* (see §5.4). The main difference is in the flexibility of the cross-tier links: instead of the spreadsheet-like tabular format of the annotation editor used in *Falko*, the annotation editor used in *CzeSL* retains links between corresponding tokens at different tiers even though the tokens are moved to remedy incorrect word order. Another difference is that unlike *Falko* and like *RLC*, *CzeSL* allows only for two annotation tiers.

Also, unlike *Falko*, we did not adopt *ANNIS*, a general-purpose search tool supporting stand-off annotation. We explored several other directions instead: a search tool built to fit the annotation scheme (see §9.2.1) and conversion strategies into several other formats: the standard token-based tabular format (see §8.3.4), the *Sketch Engine* format used in the *CLC* corpus (see §8.3.5) and the TEI XML format used in *COPLE2* and *CroLTeC* (see §8.8).

Together with *CroLTeC* and *RLC*, *MERLIN* is included as another learner corpus of a Slavic language, actually of Czech as one of its three languages. *MERLIN* is also interesting for its strategy to reuse existing tools and formats and for one of its goals: to discover language-specific features pointing to the individual proficiency levels.

We find several meeting points with the two Scandinavian projects. Like one of the more recent *CzeSL* releases, *ASK* uses the TEI format, and like the *CzeSL* tiered annotation, *ASK* also uses a restricted error tagset to avoid inconsistency and redundancy in the presence of linguistic annotation (see §5.4.2). Probably a more common feature is the use of the same tagger for the source and the target text, as in an automatically annotated *CzeSL* release (see §8.2). From our perspective, the most interesting part of the *SweLL* project is *SVALA*, the annotation editor of two parallel texts: the source and the target, with links between the corresponding tokens, reminiscent of the annotation editor used in *CzeSL* for the tiered annotation (see §9.1.1). However, there are other interesting parts: the project's policy con-

cerning sensitive personal data, the search tool combining standard concordances with the parallel text view and the plan to turn the corpus into a parallel treebank.

A more detailed picture of *CzeSL* will emerge from the following chapter.

Chapter 3

Introducing the project of Czech as a Second Language

Czech as a Second Language is the name of a long-term project and its results – a series of learner corpora. After a historical note this chapter presents some context of the project and an overview of principles and properties embodied in the results. At first we list the main properties of the *CzeSL* corpora (see §3.1): the scope of L1s and CEFR levels, their size, annotation and available metadata. We follow by outlining the intended use (see §3.2) and an overview of *AKCES*, a larger project of which *CzeSL* is a part, which includes additional corpora of Czech as L1, spoken and written mostly by schoolchildren (see §3.3).

In many ways, building a learner corpus of Czech as a second/foreign language has been a unique enterprise. To the best of our knowledge, *CzeSL* was one of the first learner corpus ever built for a highly inflectional language.¹ *CzeSL* texts have also been used in a number of studies related to FLT or SLA and in NLP applications (see §10). A case study of the *CzeSL* error annotation scheme appeared in *The Cambridge Handbook of Learner Corpus Research* (Meurers 2015).

CzeSL has been advancing since 2009 in the volume and types of texts, in the extent and quality of annotation, and in the access options. Throughout the time, new methods and tools have been tested and implemented. *CzeSL* is still not a closed and finished project. It is extendable by additional annotation and more data, including longitudinal, spoken, comparative L1 texts.

¹There was one learner corpus for a Slavic language available at the time *CzeSL* was released, namely *PiKUST* (Stritar 2009), including 35,000 words with error annotation adopted from the Norwegian project *ASK* (see §2.3.1) and one of the few using multi-layer annotation.

3.1 Specifications of *CzeSL*

Most of the texts were collected in 2009–2012. Other texts are being collected from non-native learners of Czech attending various language courses both in the Czech Republic and abroad. They are processed, annotated and published together with the old texts in new releases of the corpus.

The texts are elicited during all types of situations throughout the language-learning process. There are texts produced during the class, as homework and in an examination. A large portion consists of short essays and school exams, collected as manuscripts, scanned and transcribed into an electronic form. The rest are academic texts, Bachelors', Masters' and doctoral theses, written in Czech by non-native students and obtained from the authors already in an electronic form.

CzeSL is focused on four main groups of non-native learners of Czech:

- Speakers of related Slavic languages, represented mainly by Russian, other Eastern Slavic languages and Polish; other Slavic languages are covered marginally
- Speakers of distant non-Indo-European languages, with a majority of Chinese, followed by Vietnamese and Arabic
- Speakers of other Indo-European languages, with a slight majority of German, followed closely by French, English, Spanish and other languages
- In some releases of *CzeSL*, also speakers of the Romani ethnolect of Czech²

The corpus is based on texts covering all CEFR levels, from real beginners (A1 level) to advanced learners (level B2 and higher). In the original collection of texts (see [Table 8.5](#) on page 161) levels A1 and A2 prevail with the higher proficiency levels under-represented. More recently, some efforts aiming at a more balanced mix of levels and L1s have been made, both in terms of collected texts and in the share of texts manually error-annotated, but the result is still far from a well-balanced corpus (see [§8.7](#) and [§8.8](#)).

The largest released *CzeSL* corpus – *CzeSL-plain* – consists of nearly 2.5 million tokens (see [Table 8.1](#) on page 156). *CzeSL-plain* (see [§8.1](#)) includes also a substantial part of Romani ethnolect. Short essays written by non-native learners of Czech and students speaking the Romani ethnolect of Czech account for 1.3 mil. and 0.4 mil. tokens, respectively, while theses written in Czech by foreign students

²It is not clear whether Czech is L1 or L2 of such speakers. For more about the Romani ethnolect of Czech see [§B.4](#).

account for 0.7 mil. tokens. A part of the hand-written essays, including about 0.3 mil. tokens, is error-annotated manually, from which 0.2 mil. tokens are doubly annotated. However, most of the manually annotated texts represent the Romani ethnolect (59%).

The information about L1, CEFR level and other characteristics of the learner, the text and the situation where the text was written, is available as metadata for an overwhelming majority of *CzeSL* texts. There are 15 items that relate to the learner, while other 15 items specify the character of the text and circumstances of its production – see §4.4 for details.

Many texts were collected at regular intervals from learners attending long-term language courses. By using the author’s ID, the evolution of the author’s interlanguage can be analyzed. Some parts of *CzeSL* can thus be used for longitudinal research.

Incorrect forms in some transcripts are manually corrected (normalized, emended, reconstructed, assigned a target hypothesis) and labeled by error categories. Most texts are also tagged by tools trained on native Czech in a way similar to standard corpora, i. e., by lemmas, morphosyntactic categories, in some releases of the corpus also by syntactic functions and structure. Some error annotation tasks are done automatically: the assignment of formal error labels and even the correction step.

There is more than one approach to error annotation of *CzeSL*. A part of the texts is annotated and represented in various ways. See §5 for more about error annotation and §6 for more about linguistic annotation. Annotation as a process is described in §7.

Most *CzeSL* texts are searchable and downloadable in various formats under the Creative Commons license.³ The searchable corpora are hosted by the *Czech National Corpus (CNC)*,⁴ accessible via the *KonText* corpus search interface.⁵ Most of the downloadable corpora are available via the *LINDAT* repository.⁶ For privacy reasons, the texts are anonymized and scans of handwritten text are not publicly accessible.

3.2 Intended usage

Texts produced by learners of a second or foreign language are a precious source of linguistic evidence for experts in language acquisition, teachers, authors of didactic

³See §8 and <http://utkl.ff.cuni.cz/learncorp/> for links and more details.

⁴<https://www.korpus.cz>

⁵<https://kontext.korpus.cz/>

⁶<https://lindat.mff.cuni.cz>

tools, and students themselves. A corpus of such texts can be used to compare different varieties of non-native language, or non-native and native language on the background of traditional native language corpora. An error-tagged corpus can also be subjected to computer-aided error analysis as a means to explore the target language and to test hypotheses about the functioning of L2 grammar, e. g., in the domains of verbal tenses (Granger 1999), lexical errors (Leńko-Szymańska 2004) or phrasal verbs (Waibel 2008).

Building a resource such as *CzeSL* is expensive, so it does not make much sense to tailor its design according to the needs of a specific task or a group of users. Instead, *CzeSL* was designed to meet at least some expectations of as many users as possible, although the main intended use was pedagogical. Thus the corpus is intended for:

- Education of teachers of Czech as a foreign language: the corpus can be used to train future teachers to identify, describe and explain particular error types. From the very beginning of the project, the language data are used in language analysis in seminars on Czech as a second language at the Technical University of Liberec and at Charles University in Prague.
- Research of Czech as a second language, the Czech interlanguage and second language acquisition in general
- Compilation of teaching materials and optimization of the learning process, to provide data (specific examples or entire texts) for the analysis of non-native speakers' competence in Czech. Such analysis can serve as a basis for improving the teaching process through a focus on actual problems students make and can be used in the production of teaching materials, to tailor instructions and teaching materials to specific groups of learners (e. g., groups with different native languages or groups of different ages), and in the instruction of future teachers of Czech as a second language.
- Language testing
- NLP applications, such as CALL tools (Computer-Assisted Language Learning), spell/grammar checkers, writing assistants, including tools intended primarily for native speakers

For more about the use of *CzeSL* see §10.

3.3 *AKCES* – the umbrella project

CzeSL is built as a part of an umbrella project, the Acquisition Corpora of Czech (*AKCES*), a research program pursued at Charles University in Prague since 2005 (Šebesta 2010). *AKCES* is designed as a collection of acquisition corpora capturing written and spoken Czech of various categories of speakers:

- Preschool children
- Children and young people aged 5 to about 24 years
- Non-native speakers of Czech (foreigners learning Czech)
- Socio-culturally or otherwise disadvantaged groups, especially Roma pupils from communities at risk of social exclusion
- Czech diaspora in Romania (to be completed soon), Argentina (under construction), Bosnia and other countries (in preparation)

AKCES is also supposed to cover:

- Czech in the educational context (corpora of recordings and transcripts of classes, corpus of Czech language textbooks)
- Czech of people suffering from language impairments
- Foreign languages as spoken by Czech youth (allowing to study the influence of Czech as the first language on the acquisition of the target language)

This spectrum of various types of texts is unique in the context of other learner corpora. Apart from the *CzeSL* corpora described in this book, *AKCES* includes the following corpora, currently available or under construction:

1. Primary and secondary school classes

SCHOLA 2010 – orthographic transcripts of recordings of dialogues between teachers and pupils during standard 45-minutes' classes, collected in 2005–2008 from various Czech regions; 204 transcripts of 143.5 hours of recordings, 2,410 speakers, 61 thousand speaker turns, 1 million tokens, 793 thousand words; with metadata about the region, school, class

and speaker, without linguistic annotation; searchable in the *KonText* tool at the site of the *Czech National Corpus* (hence *CNC KonText*)⁷

AKCES 2 v2 – texts from *SCHOLA 2010*, without metadata; downloadable from the *LINDAT* repository⁸

2. Essays by primary and secondary school pupils

SKRIPT 2012 – 1,694 texts, 709 thousand tokens, 588 thousand words (347 per text), with metadata about the task, school, pupil and teacher, tagged and lemmatized; searchable from *CNC KonText*,⁹ see §8.9 for more details

AKCES 1 – texts from *SKRIPT 2012*, with self-corrections and metadata, without tags and lemmas; downloadable from *LINDAT*¹⁰

3. Longitudinal collections

CzeFL-LONG – a longitudinal corpus of *Czech as the First Language*; written (\approx 100 thousand words) and spoken (\approx 35 hours); based on samples from identical native learners of Czech within a four-year period; under construction

CzeSL-LONG – a longitudinal corpus of *Czech as the Second Language*; written (\approx 80 thousand words) and spoken (\approx 17 hours; based on samples from identical non-native learners of Czech within a period of 2–4 years; under construction

4. Transcripts of texts written and spoken in the Romani ethnolect of Czech

SKRIPT 2015 – a balanced mix of essays extracted from *AKCES 4* and *SKRIPT 2012*; searchable from *LINDAT KonText* and *TEITOK*; see §8.9 for details

AKCES 4 – a complete set of transcripts of Romani ethnolect collected for the *CzeSL* project; downloadable from *LINDAT*; see §8.9 for details

⁷Description: <https://wiki.korpus.cz/doku.php/en:cnk:schola2010>; search interface: https://kontext.korpus.cz/first_form?corpname=schola2010

⁸<https://hdl.handle.net/11858/00-097C-0000-0023-3FBB-3>

⁹Description: <https://wiki.korpus.cz/doku.php/cnk:skript2012>; search interface: https://kontext.korpus.cz/first_form?corpname=skript2012

¹⁰<https://hdl.handle.net/11234/1-1741>; Šebesta et al. (2016)

ROMi 1.0 – speech and transcripts; 50 recordings obtained in schools, during leisure activities and at home from pupils aged 13 to 24 years; 120 thousand words; 15 thousand dialogue turns and 142 speakers; searchable and available from *LINDAT*¹¹

5. Pre-school children

EARLYFAMILY 2018 – *Longitudinal Corpus of Early Language Development*:¹² spoken dialogues with a caregiver, 6 participants aged 1–4 years, transcripts, 175 recordings, about 58 hours

6. Story telling in native and non-native Czech

Frog, where are you? – the *Czech Frog Story Corpus*,¹³ in preparation

Methods and tools used for collecting, transcribing, annotating, managing and searching the written texts are the same at least for some of the *AKCES* corpora. This represents a significant synergic effect, allowing for comparative analyses of native and non-native language in corpora built on identical principles.

After an outline of the basics we are now ready to explore various aspects of designing, building, annotating and using a learner corpus, employing the Czech language and the *CzeSL* project as an example setup for showing solutions we have found or adopted, implemented and tested. We start with a series of four chapters concerned with the process of compiling the corpus: from the procurement of texts, including transcription, provision of metadata and protection of sensitive personal information (see §4), through an extensive discussion of the conceptual aspects of error annotation (see §5) and linguistic annotation (see §6), to annotation as a process (see §7).

¹¹ Accessible at <https://lindat.mff.cuni.cz/services/dialogy.org/#> after Demo Login using an institutional Shibboleth account.

¹² <https://childes.talkbank.org/access/Slavic/Czech/Chroma.html>, DOI: 10.21415/3ZNE-HX03

¹³ <https://childes.talkbank.org/access/Frogs/>

Chapter 4

Procurement of texts

Each corpus starts with collecting its content and related tasks. As it happens with most steps related to learner corpora, there are more issues specific to learner texts than to standard texts produced by native speakers. Using *CzeSL* as an example, we show how learner texts can be obtained, transcribed, equipped with metadata and protected from potential infringement of personal rights.

4.1 Text collection

The texts included in the *CzeSL* corpus in the first round (i. e., until 2012) were collected mainly from learners attending an educational institution in the Czech Republic. Most of the learners were adults (18 or older), but there were also some younger learners (15–17). A substantial share of responsibility was with the text collectors, often teachers of the class. In detailed guidelines and during extensive schooling, the collectors were instructed about the choice of learners and text topics, the handling of texts, the acquisition of metadata and the learner’s consent about the use of the text. However, the guidelines were not always fully observed and some texts eventually included in the corpus did not follow the rules.

According to the rules, 3 to 4 texts were collected from a learner within a single time interval (a school term). The task specification, included as a part of the metadata of each text, were defined as follows:

1. A text on an assigned topic, depending on the proficiency level, such as “My Family”.

2. A text on a topic selected by the learner from a list of up to 16 suggestions in the guidelines, adaptable to the learner's age and proficiency level.
3. One or two texts on a free topic. However, this did not always mean a really free choice. It was often the case that a topic picked as appropriate by the teacher/collector was assigned.

The texts were written in regular classes, as a part of final exams, but also as homework. The decision to include homeworks was mainly due to the fact that texts produced in class are rather short and not too many. This is because teachers prefer to use some of the classroom time to provide students with resources and backgrounds needed for the homework rather than spending the time on writing. Indeed, there is a risk that the use of various aids or applications in an environment beyond the teacher's control may give a distorted picture about the learner's vocabulary and grammar-related competence. However, students are allowed to use various aids even in the classroom and thus the difference between home and school does not matter that much.

We also considered the opposite approach of collecting texts from test situations, when the use of aids is controlled. However, in testing learners tend demonstrate only a part of their real competence, choosing means they are sure of to avoid failing the test. A homework on a topic of their interest is a very different task. The learners are much more ready to experiment while trying to express even complex thoughts. The comparison with texts on the same topics included in the *MERLIN* corpus is striking: the *MERLIN* texts, written during the CEFR exams, are rather stereotypical and uncreative.

The collectors also had an important role in the protection of personal rights. In addition to the task assignment, text handling, and the acquisition of metadata (see §4.4), the collectors had to negotiate the consent for making the anonymized learner texts public (see §4.3 about anonymization). In the days before GDPR, when the texts were collected, the legal demands were less strict, but the procedure was still taken seriously. For adult learners, the collector signed a solemn declaration that all participating learners agreed to the use of their texts in the project and in the corpus. For juvenile learners, the collector had to obtain the headmaster's approval. For juvenile learners attending the preparatory language courses at Charles University in Prague, a person authorized to represent the parents had to agree.

The authors of more recent additions are only adult learners and each of them signs a legally conformant statement of consent. For the previously obtained consents we assume the prohibition of legal retroactivity.

4.2 Transcription

Like most texts currently written by students in educational contexts, the materials we collected for *CzeSL* were mostly hand-written. This is usually the only available option, given that their most common source are language courses and exams.¹ The avoidance of an electronic format is also due to the concern about the use of automatic text-editing tools by the students, which may significantly distort the authentic interlanguage. Therefore, many texts have to be transcribed.²

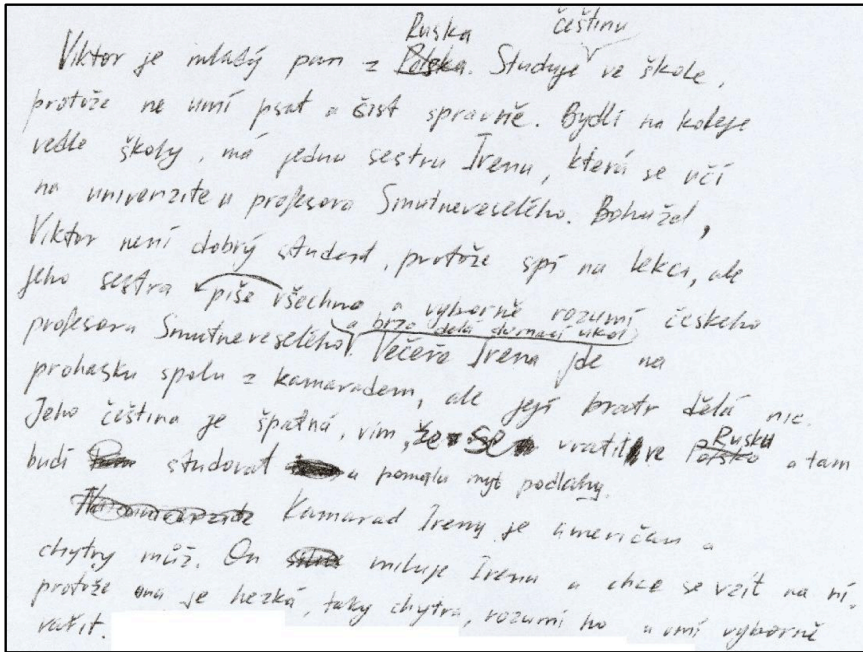
The manuscript properties are recorded in order to support the research of handwriting, especially of students with a different native writing system. Also captured are corrections made by the student (insertions, deletions, etc.), useful for investigating the process of language acquisition. While we strive to capture only the information present in the original hand-written text, often some interpretation is unavoidable. For example, the transcribers have to take into account specifics of hand-writing of particular groups of students and even of each individual student (the same glyph may be interpreted as *i* in the hand-writing of one student, *e* of another, and *a* of yet another).

Parts of some texts may be completely illegible and are marked as such. Sometimes the text allows multiple interpretation, e. g., the case of initial letters or word boundaries are often unclear. When the transcriber is not able to provide a single interpretation, two or even more variants can be used. Their order is assumed to signify preference of the first variant as the most likely interpretation. Some of the downstream processing steps which do not accept variants take advantage of this order by accepting the first interpretation and discarding the rest of them.

While deciphering unclear handwriting, transcribers sometimes have to rely on context and their best guess. However, they are not instructed explicitly to apply the “principle of positive assumption” of Volodina, Granstedt, et al. (2019): “Whenever one of the alternatives involves better intelligibility or closer adherence to standard norms, that is the alternative which should be chosen.” Unlike this principle, the approach of encoding variant interpretations is more focused on details of the learner’s handwriting. In retrospect, a single interpretation guided by the principle would have prevented some processing issues downstream at a bearable cost.

¹Electronic texts (BA, MA and Ph.D. theses) represent a minority. While these texts were not written in a class or with the aim to be included in a corpus, their final form may have been affected by an automatic spellchecker. More recently, learner texts typed in an electronic format have become more common additions to *CzeSL*.

²For transcription and anonymization from the perspective of annotation as process see §7.2.



Viktor je mladý pan z ~~Polska~~ Ruska. Studuje {čestinu}<in> ve škole, protože ne umí psát a číst správně. Bydlí na koleje vedle školy, má jednu sestru Irenu, která se učí na univerzite u profesora Smutneveselého. Bohužel, Viktor není dobrý student, protože spí na lekci, ale jeho sestra {piše všechno -> všechno piše} a vyborně rozumí českého profesora Smutneveselého {a brzo dělá domácí ukol}<in>. Večere Irena jde na prohasku spolu z kamaradem, ale její bratr dělá nic. Jeho čestina je špatná, vím, že se vrátit ve ~~Polsko~~ Rusko a tam bude studovat u pomalu myt podlahy. Kamarad Ireny je {A}meričan a chytrý muž. On miluje Irenu a chce se vzít na ní. protože ona je hezká, taky chytrá, rozumí ho a umí vyborně vařit.

Viktor je mladý pan z Polsko<add>Rusko</add>. Studuje <add>čestinu</add> ve škole, protože ne umí psát a číst správně. Bydlí na koleje vedle školy, má jednu sestru Irenu, která se učí na univerzite u profesora Smutneveselého. Bohužel, Viktor není dobrý student, protože spí na lekci, ale jeho sestra <subst>piše všechno<add>všechno piše</add></subst> a vyborně rozumí českého profesora Smutneveselého <add>a brzo dělá domácí ukol</add>. Večere Irena jde na prohasku spolu z kamaradem, ale její bratr dělá nic. Jeho čestina je špatná, vím, že se vrátit ve Polsko<add>Rusko</add> a tam bude studovat u pomalu myt podlahy. Kamarad Ireny je Američan a chytrý muž. On miluje Irenu a chce se vzít na ní. protože ona je hezká, taky chytrá, rozumí ho a umí vyborně vařit.

Figure 4.1: A sample hand-written document with its transcription in the plain and the XML-based format (NEM_GD_008 ru B2)

An example of a manuscript and its transcription can be seen in [Figure 4.1](#).³ The text is transcribed in two different ways, which differ in how some relevant features of the handwriting, mostly self-corrections, are encoded. For example, the author replaced the form *Polska* ‘Poland’ by *Ruska* ‘Russia’, inserted a word *češtinu* ‘the Czech language.ACC’ and changed the word order by moving the word *všechno* ‘everything’ leftwards. The codes are set on gray background.

At first, the hand-written texts were transcribed using off-the-shelf editors supporting HTML (e.g., Microsoft Word or Open Office Writer). As in the first transcript, a set of codes is used to capture variants, illegible strings, self-corrections and emoticons.⁴ Deletions are transcribed as **strikeout text** (**Polska**), insertions use transcription codes in angle brackets following a string in braces (**{češtinu}<in>**), word order changes are annotated using an infix arrow-like notation (**{piše všechno -> všechno piše}**). This format also supports alternative interpretations (**{A|a}meričan**), where the first option is the preferred reading. For example, the string ... represents omission (...), **&img;** indicates the place, where there was a picture in the manuscript, **&unclear;** stands for an unrecognized word or passage, **&rdot;** is a string indicating the character with a dot above etc. Unreadable characters or words were transcribed as **XXX**.

The original transcription method was prone to unchecked typos in the markup and was replaced later (in texts transcribed since 2018) by a different setup, based on an editor checking for inconsistencies in an XML-based format, including XML codes for the transcription markup.⁵

The second transcript uses such XML codes, e.g., `Polska` for deletion and `<add>češtinu</add>` for insertion. Alternatives are not supported in this format. The transcription and anonymization codes follow the TEI guidelines wherever possible.⁶

4.3 Anonymization

In most cases, the author’s identity cannot be revealed in the text or through meta-data. The hand-written texts are anonymized during the transcription: personal information is replaced either by generic names (e.g., for names of persons and

³This is the text presented with glosses and target hypotheses in [Table 5.1](#) on page 74.

⁴For details, see Štindlová (2011, 106; in Czech), or an abbreviated transcription guide <http://utkl.ff.cuni.cz/~rosen/public/transcription-reference.pdf> (in English).

⁵See <http://utkl.ff.cuni.cz/~rosen/public/TranscriptionGuideXML-cs.pdf> for a transcription guide and <http://utkl.ff.cuni.cz/~rosen/public/TranscriptionMarkupXML-cs.pdf> for a list of codes.

⁶See <https://tei-c.org/release/doc/tei-p5-doc/en/html/CC.html>.

towns) or special codes (e. g., for telephone numbers). The use of substitute names is sometimes called pseudonymization. In pseudonymization we strive to preserve agreement features (by matching the name's gender, number and case) and some of the possible errors (e. g., capitalization and some errors in declension).

We use different substitutes for declinable and non-declinable names, but we do not attempt to match the declension class – e. g., all declined female given names are substituted by an appropriate form of the name *Eva*, even if the original name (such as *Lucie*) has a different set of declension endings. Unsurprisingly, male given names are replaced with a form of *Adam*. According to the original guidelines, names of smaller places (streets, villages, small towns) and other potentially sensitive data were replaced by `QQQ`. Later, all such names, together with email addresses, phone numbers, zip codes and other information potentially revealing the author's identity, were coded as `&priv;`, e. g., `{ulice}<&priv;>` for street (*ulice*).

The XML-based anonymization codes use a single element `anon` with a `type` attribute, which identifies the type of the anonymized item, e. g., `<anon type="female FirstName">Eva</anon>`. The substitute names can be used according to similar rules also for place names and institutions, e. g., `<anon type="street">Dlouhá</anon>`. Substitutes need not be used where they cannot reflect any linguistically relevant irregularities in the original forms, e. g., `<anon type="phone"/>`.⁷

4.4 Metadata

In a learner corpus, metadata about the author of the text are at least as important as all other types of annotation.⁸ The same set of metadata items is available in most *CzeSL* corpora for nearly all texts. There are 15 items about the author of the text and 15 items about the text itself.

The sociological and linguistic data about the learner include age, gender, first language, proficiency level in Czech according to CEFR, knowledge of other (non-native) languages, bilingual competence, country of birth and residence, duration and conditions of the acquisition of Czech, including an indication of the institution, duration, or location (whether abroad or in the Czech Republic), textbooks used in learning Czech, and whether a family member has been a speaker of Czech. Specifications of the character of the text and circumstances of its production include

⁷For more details about both types of transcription and anonymization, including the codes, see the *CzeSL* site <http://utkl.ff.cuni.cz/learncorp/> – *CzeSL-man* (for the HTML-based transcription), or *CzeSL in TEITOK* (for the XML-based transcription).

⁸The role of metadata has been emphasized by many authors (e. g., Granger 2003a, 2008; Tono 2003).

the availability of language reference tools, the extent and type of elicitation, and the temporal and size restrictions.

The content of the individual items is listed in [Table 4.1](#) and [Table 4.2](#). Identifications of the items in the first column are used as XML attributes in the text headers of several downloadable and searchable *CzeSL* corpora.⁹

<code>s_id</code>	Identification of the learner: e. g., TOU_H305
<code>s_sex</code>	Sex: m or f
<code>s_age</code>	Age: e. g., 17
<code>s_age_cat</code>	Age category: 6-11, 12-15, or 16-
<code>s_L1</code>	First language: an ISO 639-1 code, e. g., sq (Albanian) ¹⁰
<code>s_L1_group</code>	Language group of the first language: IE (Indo-European non-Slavic), nIE (non-Indo-European), or S (Slavic)
<code>s_other_langs</code>	Knowledge of other languages: one or more ISO 639-1 codes
<code>s_cz_CEF</code>	Proficiency in Czech at the time of writing: A1, A1+, A2, A2+, B1, B2, C1, or C2
<code>s_cz_in_family</code>	Knowledge of Czech in the family; one or more values: mother, father, partner, sibling, 3 (3 family members), other, nobody
<code>s_years_in_CzR</code>	Years in Czechia: -1, 1, -2, or 2-
<code>s_study_cz</code>	Past or present study; one or more values: 1to1 (individual tutoring), paid, TY (self-study), university, foreign, primary-secondary, other
<code>s_study_cz_months</code>	Months of studying Czech: -3, 3-6, 6-12, 12-24, 24-36, 36-48, 48-60, or 60-
<code>s_study_cz_hrs_week</code>	Hours of studying Czech per week: -3, 5-15, or 15-
<code>s_textbook</code>	Textbook used by the learner; one or more values: BC (<i>Basic Czech</i>), CC (<i>Communicative Czech</i>), CE (<i>Čeština pro ekonomy</i>), CMC (<i>Chcete mluvit česky?</i>), CpC (<i>Čeština pro cizince</i>), ECE (<i>Easy Czech Elementary</i>), NCSS (<i>New Czech Step by Step</i>), other
<code>s_bilingual</code>	Bilingual: yes or no

Table 4.1: Metadata about the learner

⁹In the on-line searchable version of the *CzeSL-SGT* corpus the metadata items are identified by Czech labels. For a list of English and Czech metadata identifiers see http://utkl.ff.cuni.cz/~rosen/public/meta_attr_vals.html.

¹⁰See https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes. If necessary, the three-character code ISO 639-3 is used, e. g., xal (Kalmyk), see https://en.wikipedia.org/wiki/ISO_639-3.

<code>t_id</code>	Identification of the text: e.g., TOU_H305_442
<code>t_date</code>	Date of the text collection: YYYY-MM-DD
<code>t_medium</code>	Medium of the text: <code>manuscript</code> or <code>pc</code>
<code>t_limit_minutes</code>	Time limit in minutes: 10, 15, 20, 30, 40, 45, 60, <code>other</code> , or <code>none</code>
<code>t_aid</code>	Permitted resources; one or more values: <code>yes</code> , <code>dictionary</code> , <code>textbook</code> , <code>other</code> , <code>none</code>
<code>t_exam</code>	Was the text part of exam?; one or more values: <code>yes</code> , <code>interim</code> , <code>final</code> , <code>n/a</code>
<code>t_limit_words</code>	Assigned size limit in words: e.g., 150
<code>t_title</code>	Title of the essay; one or more values: e.g., <code>Událost, která změnila můj život</code>
<code>t_topic_type</code>	Type of the topic: <code>general</code> or <code>specific</code>
<code>t_activity</code>	Activity before writing the text: <code>exercise</code> , <code>discussion</code> , <code>visual</code> , <code>vocabulary</code> , <code>other</code> , or <code>none</code>
<code>t_topic_assigned</code>	Assigned topic: <code>multiple choice</code> , <code>specified</code> , <code>free</code> , or <code>other</code>
<code>t_genre_assigned</code>	Assigned genre: <code>free</code> or <code>specified</code>
<code>t_genre_predominant</code>	Genre predominant in the resulting text: <code>informative</code> , <code>descriptive</code> , <code>argumentative</code> , or <code>narrative</code>
<code>t_words_count</code>	Actual number of words: integer
<code>t_words_range</code>	Range of the actual number of words: -50, 50-99, 100-149, 150-199, or 200-

Table 4.2: Metadata about the text

In most texts, the metadata were specified by the text collector. For some items, such as the proficiency level, rather than applying a set of objectively defined criteria, collectors had to estimate the level by combining their impression of the learner with instructions received during training sessions and included in the collectors' guidelines. As a result, this metadata item should be taken with a grain of salt. For more about the issue of inaccurate CEFR levels in the *CzeSL* corpus see §11, page 220.

The representation of metadata is not the same in all *CzeSL* corpora. In the tiered format generated by the *feat* tool and used in *CzeSL-man v1 downloadable* (see §8.3.3), metadata concerning a specific text are stored in a separate file, together with files corresponding to the individual tiers and following the same naming convention. For example, a text identified as `KAR_MI_005` has its metadata in a file named `KAR_MI_005.meta.xml`. The metadata items are represented as XML elements, see Figure 7.3, page 138.

Releases of *CzeSL* corpora searchable in *KonText* have their metadata encoded

as XML attributes in the text headers, see [Figure 8.1](#) on page 162. This applies to *CzeSL-SGT* (including its downloadable version), *CzeSL-man v1 searchable*, and *CzeSL-man v2* (including its downloadable version). Yet, the metadata for the searchable and downloadable releases of *CzeSL-SGT* also differ: the metadata items are named in Czech in the searchable version.¹¹

In the *CzeSL in TEITOK* corpus, metadata are represented in a way conformant with the TEI guidelines, as long as they are available for the specific *CzeSL* items. They are part of the text XML header, for an example see [Figure 8.3](#) on page 171. Metadata can be displayed in a user-friendly way and in a preferred language, depending on the available localization and the setting of the corpus tool for the specific corpus. Like the text itself, metadata can also be edited in the user interface by an authorized user.

¹¹See http://utkl.ff.cuni.cz/~rosen/public/meta_attr_vals.html for a bilingual list of the attributes and values.

Chapter 5

Error annotation

5.1 Errors and learner language

Corpus-based research of learner language inherited its two main methodologies from the field of second language acquisition (SLA): contrastive analysis (CA) and Error Analysis (EA). The main focus of CA is the comparison of the language of the learner with the native language, typically resulting in data about the underuse or overuse of specific linguistic phenomena. For CA, learner errors in the corpus need not be annotated, it is enough to identify comparable exponents of the researched features in the native and the learner language. On the other hand, it is hard to imagine a corpus-based EA study without an explicit identification of such errors in the corpus. This is the reason why some error annotation is available for most learner corpora, and also why learner corpora are typical by including error annotation as their specific feature.¹

To design and implement error annotation is not an easy task. Despite a rich pool of literature devoted to the subject, only few solutions seem to be reused in new projects. Yet a thorough research of available options is advisable. Discussions of learner language predating the boom of learner corpora often concern issues recurring in contemporary efforts to design and build resources optimally suited to the language and the goals of the project.²

There are many points where the *CzeSL* project touches upon such issues. Staying with the topic of this chapter, one of the key points is the status of error. At least

¹See, e.g., Díaz-Negrillo and Fernández-Domínguez (2006).

²For overview and references see, e.g., Lüdeling and Hirschmann (2015) and Stemle et al. (2019).

since Corder (1967), errors are treated as a necessary part of language acquisition and an important indication of the discrepancy between the learner's competence in L2 and its native counterpart. Once the hypothesis of learner's internal grammar (IL) is accepted (Selinker 1972), errors can be treated as an important source of knowledge about the hidden system. However, only errors reflecting the system (competence errors) rather than random errors (performance errors or mistakes) are relevant for the study of IL (Corder 1967, 166). The problem is that the two error types can be more or less reliably distinguished only by analysing all data available from a specific source. It is impossible to decide separately for each individual instance, thus errors and mistakes cannot be told apart during annotation. To remedy that, *CzeSL*, like most other learner corpora, offers metadata about the learner, usable to decide about errors occurring in texts of a specific learner or a group of learners.

A more profound problem related to errors in relation to IL was pointed out by Bley-Vroman (1983): errors are the result of comparison of a language as L2 with the same language as L1. Errors are therefore a concept seen from the perspective of a native speaker. On the other hand, the learner's internal grammar (IL) and its evolution can be based on categories and principles very different from those of the target language. Bley-Vroman (1983, 4) sees the reliance on errors in the effort to describe IL as “a departure from the original spirit of the interlanguage hypothesis as advocated in Selinker (1972)”, as something actually preventing to see the systemacity of the learner language:

Language systems are to be considered in their own right, on the basis of their own “internal logic.” The structuralist linguistic rejection of what were perceived to be “Latin-based” grammars must be seen in this light. Languages which seem absurd and illogical when viewed from the standpoint of Latin turn out to be reasonable and (as we would say) systematic when allowed to stand on their own. In the same way that we would not judge the systematicity of Nootka by comparing it with Latin (or even with Kwakiutl) we do not appropriately measure the internal systematicity of an interlanguage by comparing it with another (albeit related) language, the target language. (Bley-Vroman 1983, 15)

Yet we agree that “many of the properties of learner language can only be understood if learner language is compared to target language structures” (Lüdeling and Hirschmann 2015, 155) and see error annotation as crucial for studying learner language. Rather than obscuring the access to IL, appropriately used error annotation can be used as a platform for studying IL from various angles.

There is also a more practical issue, namely the issue of what counts as an error. While some errors are definable as violating a grammar rule (in terms of R. Ellis 1994, 701, “overt errors”), native speakers often do not agree about cases of inappropriate but still grammatically correct use (“covert errors”). The latter tend to occur especially in the language of advanced learners and are perceived as features typical for a non-native speaker rather than as errors as such. Acknowledging the grey zone, Lennon (1991, 182) defines an error vaguely in contrast to what a native speaker would produce in the same context and under similar conditions.

We tried to be less vague in *CzeSL* by using SCz as the yardstick wherever the codified norm offered any support, in the grammar or in the lexicon, or by discouraging annotators from correcting/annotating stylistically inappropriate expressions unless they impede understanding or sound very unnatural (see §5.3). The strategy to keep the annotators’ intervention with the text at a minimum is probably motivated by similar concerns and can have the same effect as the “principle of positive assumption,” intended to prompt the preference for interpreting the text to the learner’s advantage (Volodina et al. 2016; Volodina, Granstedt, et al. 2019).³

Some errors can be detected and corrected within a very restricted span of text, e. g., a word, a morph or even a single character. Other errors may require a much larger context (a constituent, clause, sentence, paragraph or even the entire text) or even a glimpse into the extralinguistic situation. Lennon (1991, 191) conceptualizes the range of context needed for the error to become apparent as “error domain”, and the span of text which needs to be repaired as “error extent”. There is some correlation between the two notions on the one hand and the grammar-based type of error. E. g., in English, an error in the choice of a preposition could have a domain spanning a clause, while its extent would be limited to the preposition. In case-marking languages the extent could include the NP following the preposition.

In *CzeSL*, annotators are instructed to find a TH with regard to the context and the probable intention of the author, while at the same time making sure that it is as close as possible to the original, i. e., to minimize the changes of word forms, word order, lexical setting, number of words, etc. In this sense, the principle of contextual interpretation is applied in parallel with the principle of minimal intervention.

However, error domain and error extent are not primitive concepts in any of the *CzeSL* error annotation schemes. They are only implied in the tiered (2T) error annotation (see §5.4). For cases when the domain and the range coincide, the annotation scheme uses links connecting the source word forms with their TH coun-

³Strictly speaking, the principle of positive assumption formulated in Volodina et al. (2016) and Volodina, Granstedt, et al. (2019) concerns the strategy of resolving uncertainty about the interpretation of written text, but it seems to be also applicable to the decision whether a borderline case should be counted as an error.

terparts. The domain of an error is specified by linking potentially multiple (even non-contiguous) source forms and expanding the joined links again as potentially multiple links targeting the correct forms, representing the error extent. This mechanism is used for cases such as joining and splitting forms, changing word order or correcting multi-word units. For cases when the domain is wider than the extent, the scheme uses pointers to forms motivating the correction, e. g., to a subject as the agreement source for a predicate form supposed but failing to agree. For cases when a wider context must be repaired, the notion of a follow-up or secondary error is used (see §5.4.1.9), as in the above example involving the replacement of a preposition. If the correct preposition assigns a different case to its NP, the corrected NP is tagged as subjected to a follow-up correction.⁴

A single form can be diagnosed as erroneous for several reasons. Lennon (1991, 193) points out the special case of embedded errors, as in *he seems to be drunken* → *he seems to be drowned*. The morphologically incorrect form *drunken* → *drunk* must still be corrected as *drunken* → *drowned*. Lennon suggests to ignore the embedded (intermediate) error for the purposes of error counting (and, we might add, error annotation). His main reason is not the impossibility of a different solution, but concern about consistency of the error analysis. In *CzeSL*, we tried both ways: embedded errors are annotated in the 2T (see §5.4) and implicit schemes (see §5.5) and ignored in the MD (see §5.6) and UD schemes (see §6.2). The former extracts more information, the latter simplifies the process.

Discussion of error analysis and error annotation cannot avoid the role of TH. Is an explicit formulation of TH needed for annotating (error-tagging) each error? Are multiple successive or even alternative THs needed? Can errors be identified only by THs?

According to Lüdeling and Hirschmann (2015, 141), “Errors cannot be found and analysed without an implicit or explicit target hypothesis – it is impossible not to interpret the data.” In *Falko* (as in the tiered error annotation of *CzeSL*), TH is a necessary component of error annotation. In fact, a single TH is often not enough. Typically, there are two THs in *Falko*, one for strictly grammatical errors, one for stylistic issues, idioms etc. One of the arguments for multiple THs is also valid for successive THs in *CzeSL* (see §5.4.1.1):⁵

⁴Unlike in *CzeSL*, where the implied domain and extent are specified in terms of word forms (tokens), Lennon (1991) delimits their range by linguistic units: morphemes, words, constituents, sentences. This can be the reason why in *CzeSL* the extent of some errors appears to be wider than their domain, as in some follow-up corrections, contrary to Lennon (1991, 192): “for any given error, domain will be at a higher rank than or equal rank to extent, but never at a lower rank”.

⁵However, the tiered annotation scheme of *CzeSL* does not support alternative THs (see

an error-annotated corpus which does not provide target hypotheses hides an essential step of the analysis – this could lead to mistakenly assuming that the error annotation which is present in a corpus is the ‘truth’ or ‘correct analysis’ instead of just one among many interpretations (142)

However, in a sentence *The girls was laughing* the error can be tagged as an error in agreement without deciding about the grammatical number and thus without a TH.⁶ There is also a similar German example (144, ex. 7) *Jeder werden davon profitieren* → *Jeder wird davon profitieren / Alle werden davon profitieren*. Yet in such cases the omission or underspecification of alternative THs does not eliminate the notion of TH from error annotation.

It is another example (145, ex. 10) that could raise doubts about the existence of a TH in all cases. One of the three possible THs for *it sleeps inside everyone from the start of being* actually includes an error tag instead of a correction: *it sleeps inside everyone UNIDIOMATIC*. This is still supposed to be a case of an implicit TH – the annotator just could not think about an appropriate idiom. Here we are not convinced that a TH is present, in any case it has a very ephemeral status.

On the other hand, most errors usually require less expert knowledge to correct than to classify – see **Implicit error annotation** (§5.5). In cases where the TH is obvious from the incorrect word form and/or the context, the annotator need not speculate about the learner’s intention. In fact, Lüdeling and Hirschmann (2015, 141) claim that the learner’s intention is not involved in the process of finding a TH in general:

It is important to note that the construction of a target hypothesis makes no assumptions about what a learner wanted to say or should have said. The analyser cannot know the intentions of the learner. The ‘correct’ version against which a learner utterance is evaluated is simply a necessary methodological step in identifying an error.

Based on our experience with all sorts of Czech texts from learners of various proficiency and L1 background we cannot agree. What else is there to guide the annotator towards a TH than a guess about the learner’s intentions?

§5.4.1.10).

⁶This would only be possible if the annotation scheme allowed for the omission of TH and for specifying an error in agreement as a relation between the two disagreeing words. Technically, the tiered error annotation scheme of *CzeSL* can be adapted to accommodate such a solution, although it would not be in accordance with the annotation guidelines.

For a discussion about the practical issue whether TH and error tags are better annotated in one go or separately see §7.3.1.

5.2 More than one way to annotate errors in *CzeSL*

Error annotation is a crucial component of most *CzeSL* corpora. Our approach to error annotation is one of the aspects of the corpus design reflecting the aim to serve many types of users. Rather than focusing on a narrow domain of learner language as the annotation target (such as spelling or lexical errors), the corpus is intended as open to as many research goals as possible. This is one of the main reasons why the target hypothesis is aimed at SCz and the error taxonomy is based on the standard linguistic concepts (spelling, morphology, syntax, semantics, agreement, valency), rather than on categories rooted in the concepts of interlanguage, communication strategy or specific research goals.

Following the same approach of not aiming at any specific group of users, we have designed, tested and used two complementary error annotation schemes, and tested and used another one. Our first proposal, referred to as the *two-tier annotation scheme* – 2T, is based on parallel tiers, representing the source text and supporting successive corrections in two stages: corrections of spelling and all other corrections (see §5.4). The two annotation tiers were introduced as a compromise between several theoretically motivated levels and practical concerns about the process of annotation. They enable the annotators to register anomalies in isolated forms separately from the annotation of context-based phenomena but saves them from difficult theoretical dilemmas.

To determine the target hypotheses and to apply the grammar-based error categorization (see §5.4.2) was the task of human annotators. Some of the texts were annotated independently by two annotators and evaluated (see §5.4.4).

The tagset used by the annotators, slightly biased towards morphosyntax, and less detailed than most other error tagsets, was meant to be complemented by other annotation. The absence of POS distinctions in the error tags is a way to avoid redundancy in a corpus which is also annotated by POS tags (see §6). The lack of a detailed analysis of errors in spelling and morphonology is to some extent remedied by an automatically applied tagset identifying formal distinctions between the source forms and their corrections (see §5.4.5).

Our second proposal, the *multidimensional* annotation scheme – MD, was developed to complement the 2T scheme by filling the gaps in the categorization of errors in spelling, morphonology and morphology, while allowing for alternative in-

terpretations of a single error, e. g., as an error which could be explained as an issue of spelling, morphonology, morphology or morphosyntax (see §5.6).

The motivation for using *implicit annotation* (see §5.5), i. e., corrections without error tags, is twofold. Firstly, the full-fledged manual 2T or MD annotation requires a well-trained annotator and more time for the same amount of text. Eliminating the error tagging task makes the perspective to hand-annotate all currently available and new *CzeSL* texts realistic, while leaving open the option to assign error tags later. Secondly, corrections can be assigned to specific error interpretation levels, corresponding to the tiers in the 2T scheme, or to a more sophisticated system of linguistic domains, as in the MD scheme. Thus, the three error annotation schemes are compatible and complementary. In fact, the three schemes can be implemented in a single corpus.

5.3 A wishlist for error annotation

Designing an error annotation scheme for non-native Czech is a challenging task. Czech, at least in comparison to most languages of the existing annotated learner corpora, has a more complex morphology and a less rigid word order, which opens annotation issues that had not been addressed before the error annotation of *CzeSL* started. As can be expected, the language of a learner of Czech may deviate from the standard in a number of aspects: spelling, morphology, morphosyntax, semantics, pragmatics or style. To cope with the multi-level options of erring in Czech and to satisfy the goals of the project, the annotation scheme should:

1. Properly handle Czech as an inflectional and free-word-order language, e. g., support successive corrections and annotation of errors in discontinuous expressions
2. Be detailed and informative but manageable for the annotators, e. g., preserve the original text alongside with its corrected version and represent syntactic relations for errors in agreement, valency, pronominal reference
3. Be open to future extensions, allowing for alternative/more detailed taxonomy to be added later
4. Provide solutions for issues of interference (see §5.3.1), interpretation (see §5.3.2), word order (see §5.3.3) and style (see §5.3.4)

The resulting annotation scheme and the error typology is a compromise between the limitations of the annotation process and the demands of research into learner corpora.

5.3.1 Interference and other types of explanation

Interference figures prominently among the candidates for the most relevant explanation of an error. Interference (also called positive or negative language transfer, or crosslinguistic influence) involves an inappropriate use of linguistic features from another language known to the learner, usually their native tongue, or the inappropriate avoidance of such features.

A sentence such as *Tokio je pěkný hrad* ‘Tokio is a nice castle’ is grammatically correct, but its author, a native speaker of Russian, was misled by “false friends”, assuming *hrad* ‘castle’ as the Czech equivalent of Russian *gorod* ‘town, city’. Similarly in *Je tam hodně sklepů* ‘There are many cellars’. The formally correct sentence may strike the reader as implausible in the context. The interpretation becomes clear only with the knowledge that *sklep* in Polish means ‘shop’, not ‘cellar’ (i. e., *sklep* in Czech).

However, to identify and correct the error without more or less thorough knowledge of the other language is impossible. In practical terms, the identification of all types of interference in a corpus with many L1s is very hard. Most of our annotators were no experts in Czech as a foreign language or in L2 learning and acquisition, and unaware of possible interferences between languages the learner knows. Thus they would have very likely failed to recognize an interferential error.

Interference is just one of many types of error diagnostics which is different from grammar-based annotation or other relatively straightforward categorization. The perspective subsuming interference is concerned with the discovery of causes or explanations. Apart from the practical issue of annotating such properties without researching other resources, such as additional texts from the same learner, perhaps at different stages of the acquisition of L2, there is also a theoretical reason why the explanation of errors should be kept separate from the more down-to-earth types of linguistic annotation. Even though all annotation is interpretation, interpretation in terms of grammar-based categories or even stylistic appropriateness is governed by instructions, linguistic rules and/or relations to L1, while finding an explanation for an error can hardly be guided by guidelines.

For such reasons, instead of its explicit annotation, interference and error explanations of other types are assumed to be identified by the corpus user in the process of interpreting the corpus data, other types of annotation and the metadata.

5.3.2 Interpretation in terms of TH

For some types of errors, the problem is to define the limits of interpretation in terms of TH. Example (9) shows two possible interpretations (TH1 and TH2) of a

grammatically incorrect clause (S), corresponding to the concepts of “minimal” and “maximal” TH in the *Falko* corpus (see §2.3.5). The clause is roughly understandable as its TH1 version, but it can also be rewritten as TH2, which is further from the source clause. The TH1 version is less natural but closer to the original. However, to provide annotation in terms of TH2 the task of the annotator is interpretation rather than correction.

- (9) S: *kdyby *cítila na tebe *zlobna*
 TH1: *kdyby se cítila na tebe rozzlobená*
 if REFL felt at you angry
 ‘if she felt angry at you’
 TH2: *kdyby se na tebe zlobila*
 if REFL at you was-angry
 ‘if she was angry at you’

Without the option to provide both THs, as in *Falko*, it is difficult to provide clear guidelines. In the manual annotation of *CzeSL*, the TH is not supposed to aim at perfect Czech. Instead, the source text is corrected conservatively to arrive at a coherent and well-formed result, without any ambition to produce a stylistically optimal solution, refraining from too loose interpretation. In this sense, the annotator is instructed to minimize interpretation. In general, the ultimate TH in *CzeSL* is closer to *Falko*’s TH1 rather than TH2, unless the grammatically correct version is hard to understand or very unnatural.⁷ Where a part of the input is not comprehensible, it is marked as such and left without correction.

5.3.3 Word order

Czech constituent order reflects information structure (see §B.3) and it is sometimes difficult to decide (even in a context) whether an error is present.⁸ The sentence *rádio je taky na skříní* ‘a radio is also on the wardrobe’ suggests that there are at least two radios in the room, although the more likely interpretation is that among other things which happen to sit on the wardrobe, there is also a radio. The latter interpretation requires a different word order: *na skříní je taky rádio*.

In accordance with the preference of conservative target hypotheses (see §5.3.2), word order should be corrected only when it is perceived as ungrammatical. Misplaced 2nd position clitics are a typical example, as in *rozhodli se jsme* → *rozhodli*

⁷For a related discussion about what counts as an error in L2 see §5.1.

⁸See §5.1 for more about “covert errors”.

jsme se ‘we have decided’. However, in cases when word order (i) makes a difference in meaning, as in the switched order of the two NPs above (‘the radio’ and ‘the wardrobe’), and (ii) the context makes it clear which meaning is appropriate, word order should be corrected even though it is grammatical. In this sense, word order and lexical corrections share the same approach: correction is due whenever an item or pattern does not fit the meaning of the context.

5.3.4 Style

The phenomenon of Czech diglossia (see Appendix §B) is reflected in the problem of annotating non-standard language, usually individual forms with colloquial morphological endings. Because learners may not be aware of the status of these forms and/or an appropriate context for their use, Colloquial Czech (CCz) is corrected under the rationale that the authors expect the register of their text to be perceived as unmarked.

To give a prototypical example, one of the most frequent problems in learner texts is the absence of appropriate diacritics. At the same time, a missing acute accents on some verbal endings in written text is perceived as colloquial, because it is supposed to reflect the colloquial pronunciation of these forms: *znam* → *znám* ‘I know’, *nosim* → *nosím* ‘I wear’. There are nearly 2.5 thousand instances of 180 different apparently colloquial verbs in the 1st person singular in the 1 million *CzeSL-SGT* corpus. Cases like this are treated as errors (in spelling or morphonology), but they are also labeled as colloquial style, suggesting that the learner could have used a colloquial instead of an incorrect form.⁹ It is up to the user of the corpus to interpret the annotation according to a wider context or the learner’s profile in the metadata.

5.3.5 Communication goal

Other features of the learner language may also be considered as candidates for annotation, such as a measure estimating to what extent the learner’s communication goal is achieved. In fact, there is hardly anything that matters more in practice and could be reflected even at the level of individual utterances.

⁹The colloquial marker is in fact a category from the domain of linguistic rather than error annotation. However, it is used in the manual error annotation to make the point that some forms annotated as incorrect can also be interpreted as colloquial forms. The colloquial marker can be confirmed in the annotation provided by a tagger applied to the source text, although the automatic linguistic annotation of such forms may be less reliable than of their SCz counterparts.

On the other hand, not every aspect of the learner language must be explicitly annotated. It could even be a more proper move to leave some of the trickier phenomena such as interference or exhaustive interpretation for the corpus user while providing a reliable annotation of errors where a safer ground is available in linguistic theory, established categories and the annotators' competence. Such annotation, provided ideally by the combination of the three annotation schemes, can help the user to interpret the search results or statistical findings in ways not previewed in the annotation.

5.4 The two-tier annotation scheme

The two-tier scheme, including its error tagset, was designed to suit the specifics of learner Czech. In this respect, the scheme proved to be adequately expressive and practically useful.¹⁰ As the most sophisticated of the three schemes used in the *CzeSL* project, it deserves to be presented and examined from multiple angles, together with its merits and drawbacks.

At first, we focus on foundations of the scheme, namely on why there are several parallel tiers, why exactly two tiers representing up to two successive THs, how the words represented at those tiers are related and how the errors can be tagged (see §5.4.1). The rest of the section is concerned with the error tagset (see §5.4.2), its evaluation (see §5.4.4), and a complementary “formal” tagset, used in rules comparing source forms and their corrections, without human intervention (see §5.4.5).

5.4.1 Annotation scheme as a compromise

5.4.1.1 Why multiple tiers

After a careful examination of available options, we have arrived at a two-stage annotation design, consisting of three parallel tiers: a tier of the source text and two annotation tiers. Between the two opposite options of a flat inline annotation and a scheme consisting of more parallel tiers (see §2.2.2.5), we decided for a compromise solution.

The choice of a multi-tier annotation scheme with a specific number of tiers calls for some justification. The optimal error annotation strategy is determined both by the goals and resources of the project and by the type of the language. A simple

¹⁰One of the two case studies in Meurers (2015) presents the scheme as a showcase example of a “state-of-the-art learner corpus annotation project integrating insights and tools from NLP”.

flat scheme with all annotation inline could be used for a specific narrowly defined purpose, such as investigation of morphological properties of the learner language, or for a language without an elaborate inflection system. Such a scheme can be appealing if corrections concern individual word forms or contiguous sequences of forms and successive or alternative corrections are not required.

A scheme with a tier for the original text and a single parallel annotation tier would be appropriate if we were interested only in the original text and in the annotation at some specific level (fully emended sentences, or some intermediate stage, such as corrected word forms). This design could be used even if we insisted on registering some intermediate stages of the passage from the original to a fully emended text, and decided to store such information with the word-form nodes. However, such information might get lost in the case of significant changes involving deletions or additions. For example, in Czech as a pro-drop language, the annotator may decide that a misspelled personal pronoun in the subject position should be deleted. Then the information about the spelling error would disappear.

Given the goals of the project and the properties of Czech, either of the two solutions – the inline annotation and a single annotation tier – was problematic. There were at least three reasons:

1. The corpus should be open to multiple research goals. Thus, it would not do to accommodate the analysis of a restricted set of linguistic phenomena within the inline annotation or a single tier.
2. Due to the fairly rich morphology and the relatively free word order of Czech, it is necessary to provide space for successive corrections. At the same time, it is important to maintain links between the original and the corrected forms even when the word order changes or when words are dropped or added. Otherwise it would be difficult to find the ultimate target hypothesis for a faulty expression or to find a corresponding expression in the original text given its target hypothesis.
3. Learner texts include word-boundary errors, i. e., incorrectly split or joined forms, or errors spanning multiple forms, even in discontinuous positions. The most natural way to annotate errors of this type with a target hypothesis and error label is in a multi-tier annotation scheme.

Actually, the decision to use a multi-tier design was mainly due to our interest in annotating errors in single forms as well as those spanning (potentially discontinuous) strings of words.

5.4.1.2 How many tiers

Once we have a scheme of multiple tiers available, we can provide them with theoretical significance and assign a linguistic interpretation to each of them. In a world of unlimited resources of annotators' time and experience, this would be the optimal solution. Annotators could be free to use an arbitrary number of tiers to suit the needs of successive emendations. They could choose from a set of linguistically motivated tiers or introducing annotation tiers ad hoc. The first annotation tier would be concerned only with errors in graphemics, followed by tiers dedicated to morphemics, morphosyntax, syntax, lexical phenomena, semantics and pragmatics. More realistically, there could be a tier for errors in graphemics and morphemics, another for errors in morphosyntax (agreement, government) and one more for everything else, including word order and phraseology.

On the other hand, annotators should not be burdened with theoretical dilemmas and the result should be as consistent as possible, which somewhat disqualifies a scheme using a flexible number of tiers. This is why we adopted a compromise solution with two tiers of annotation, distinguished by formal but linguistically founded criteria to make the annotator's decisions easy. It is a compromise between an inline or single-tier annotation and an open multi-layer format, but a compromise preserving links between split, joined and re-ordered tokens, corrected in two stages, something not obviously supported in the multi-layered tabular format described below in §5.4.1.3.

Each of the choices made in the design of the annotation scheme is a compromise between its feasibility in a practical large-scale annotation process and the requirement of a detailed and complex analysis. The restriction in the number of annotation tiers has proved its feasibility while still being useful and linguistically relevant.

5.4.1.3 Multiple tiers in a tabular format

Many corpora use simple inline error annotation, denoting the scope, correction and categorization of an error. A few corpora such as *Falko* (see §2.3.5) adopt multi-tier annotation in a tabular format, with the option of specifying multiple corrections and several error types for single word tokens or strings thereof at several linguistically motivated tiers: orthography, morphology, syntax, lexicon, pragmatics, intelligibility. The tabular format is also used in *MERLIN* (see §2.3.7), one of the two currently available corpora including Czech. The format and the corresponding tools were considered also for the manual two-tier annotation of the *CzeSL* texts.

Originally, the multi-tier tabular format and related tools were designed for an-

notating speech. The environment allows for an arbitrary segmentation of the input and multi-tier annotation of segments (Schmidt 2009). Typically, the annotator edits a table with columns corresponding to words and rows corresponding to tiers. A cell can be split or more cells merged horizontally to allow for annotating smaller or larger segments. This way, phenomena such as agreement or word order can be emended and tagged (Lüdeling et al. 2005).

However, the tabular format is not quite suitable for languages with free word order and rich inflection, where a single form may be incorrect in several domains at once: typography, orthography, morphosyntax, lexicon, word order. In the tabular format, vertical correspondences between the original word form and its corrected equivalents or annotations at other tiers may be lost. It is difficult to keep track of links between forms merged into a single cell, spanning multiple columns, and the annotations of a form at other tiers (rows). This may be a problem for successive corrections involving a single form, starting from a typo up to an ungrammatical word order, but also for morphosyntactic tags assigned to forms, whenever a form is involved in a multi-word annotation, and its equivalent or tag is no longer present in the column of the source form.

5.4.1.4 Content of the tiers

As a compromise between corpus users' expected demands and limitations due to the annotators' time and experience, the two-stage annotation design reflects the distinction roughly between errors in orthography and morphemics on the one hand and all other error types on the other.

The scheme consists of three interconnected tiers – see Figure 5.1 for an annotated example glossed in (10).¹¹ Annotation tiers are represented as a graph consisting of a set of interlinked parallel paths where a path is a sequence of word forms corresponding to a sentence at a given level. Each word in the input text is represented at every level, unless it is split, joined (as *kdy by* in Figure 5.1), deleted or added by the annotator. Whenever a word form is corrected, the type of error can label the link connecting the incorrect form with its corrected version (such as *incorInfl* or *incorBase* for morphological errors in inflectional endings and stems, *stylColl* as a stylistic marker, *wbdOther* as a word boundary error, and *agr* as an error in agreement).

Tier 0 (T0) – Anonymized transcript of the hand-written original string of graphemes, with some properties of the manuscript preserved in the transcription mark-up (self-corrections, variants, illegible strings).

¹¹Figure 5.1 is a screenshot of the annotation editor *feat* (see §9.1.1).

Tier 1 (T1) – The tier of orthographic and morphological normalization. As a rule of thumb, this is where forms incorrect in isolation are corrected. The result is a string consisting of correct Czech forms, even though the sentence may not be correct as a whole. The rule of “correct forms only” has a few exceptions: a faulty form is retained if no correct form could be used in the context, or if the annotator cannot decipher the author’s intention. On the other hand, a correct form may be replaced by another correct form if the author clearly misspelled the latter, creating an unintended homograph with another form. A formally correct form *weak* in a sentence such as *I’ll see you in a weak* would be corrected since the author clearly misspelled the form she intended to use, creating an unintended homograph. On the other hand, the form *week* in *I’ll see you in two week* is an error in morphosyntax and will be corrected at T2.

Tier 2 (T2) – Handles all other deviations, resulting in a grammatically correct sentence. This includes errors in syntax (agreement, government), lexicon, word order, usage, style, reference, negation, or overuse/underuse.

- (10) T0: **Myslím* že ****kdy*** ****by*** *byl* *se* **svím* **dítem* ...
 think.(SG1) that if would.*3RD WAS.MASC with self’s child
- T2: *Myslím*, že *kdybych* *byl* *se* *svým* *dítětem*, ...
 think.SG1 that if+would.1SG WAS.MASC with self’s child
 ‘I think that if I were with my child, ...’ (KKOL_AV_007 ru B1)

A more complex example is presented below in §5.4.1.5.

5.4.1.5 A sample text with T1 vs. T2 corrections

To exemplify various types of deviations of L2 Czech from the standard, the sample text in Table 5.1 highlights errors according to the tier they are corrected. Forms wrong in any context due to an error in spelling or morphology, corrected at T1, are set in boldface, while forms wrong due to a morphosyntactic or lexical anomaly, corrected at T2, are underlined. Some forms may be faulty for both reasons; these are in bold and underlined.

5.4.1.6 Links between tiers

While in the tabular format the correspondences between elements at various tiers are captured implicitly, in our annotation scheme these correspondences are explicitly encoded. The format supports the option of preserving correspondences across

T0	T2
<p><i>Viktor je mladý pan z Ruska.</i> Viktor is a young <u>Mr.</u> from Russia.</p>	<p><i>Viktor je mladý pán z Ruska.</i> Viktor is a young <u>man</u> from Russia.</p>
<p><i>Studuje češtinu ve škole, protože ne umí psát a číst správně.</i> He studies Czech at school, because he can not write and read correctly.</p>	<p><i>Studuje češtinu ve škole, protože neumí psát a číst správně.</i> He studies Czech at school, because he cannot write and read correctly.</p>
<p><i>Bydlí na koleje vedle školy, má jednu sestru Irenu, která se učí na univerzite u profesora Smutněveselého.</i> He lives at <u>residence halls</u>.<u>GEN</u> next to the school, has one sister Irena, who is a student of professor Smutněveselý at the university.</p>	<p><i>Bydlí na koleji vedle školy, má jednu sestru Irenu, která se učí na univerzitě u profesora Smutněveselého.</i> He lives at <u>residence halls</u>.<u>LOC</u> next to the school, has one sister Irena, who is a student of professor Smutněveselý at the university.</p>
<p><i>Bohužel, Viktor není dobrý student, protože spí na lekci, ale jeho sestra všechno píše a výborně rozumí českého profesora Smutněveselého a brzo dělá domácí úkol.</i> Unfortunately, Viktor is not a good student, because he sleeps in the class, but his sister writes everything and perfectly understands the Czech professor Smutněveselý and does her homework soon.</p>	<p><i>Bohužel, Viktor není dobrý student, protože spí na lekci, ale jeho sestra všechno píše a výborně rozumí českému profesorovi Smutněveselému a brzo dělá domácí úkoly.</i> Unfortunately, Viktor is not a good student, because he sleeps in the class, but his sister writes everything and perfectly understands the Czech professor Smutněveselý and does her homework soon.</p>
<p><i>Večere Irena jde na prohasku spolu z kamaradem, ale její bratr dělá nic.</i> Dinner Irena goes for a walk with her friend, but her brother does nothing.</p>	<p><i>Večer Irena jde na procházku spolu s kamarádem, ale její bratr nedělá nic.</i> In the evening Irena goes for a walk with her friend, but her brother doesn't do anything.</p>
<p><i>Jeho čeština je špatná, vím, že se vrátit ve Rusku a tam budí studovat u pomalu myt podlahy.</i> His Czech is poor, I know that he will return to Russia and there he wakes study at slowly wash floors.</p>	<p><i>Jeho čeština je špatná, vím, že se vrátí do Ruska a tam bude studovat a pomalu mýt podlahy.</i> His Czech is poor, I know that he will return to Russia and there he will study and slowly wash floors.</p>
<p><i>Kamarád Ireny je Američan a chytrý muž.</i> Irena's boyfriend is an American and a smart guy.</p>	<p><i>Kamarád Ireny je Američan a chytrý muž.</i> Irena's boyfriend is an American and a smart guy.</p>
<p><i>On miluje Irenu a chce se vzít na ní, protože ona je hezká, taky chytra, rozumí ho a umí výborně vařit.</i> He loves Irena and wants to marry on her, because she is pretty, also smart, she understands him.<u>GEN</u> and is an excellent cook.</p>	<p><i>On miluje Irenu a chce si ji vzít, protože ona je hezká, taky chytrá, rozumí mu a umí výborně vařit.</i> He loves Irena and wants to marry her, because she is pretty, also smart, she understands him.<u>DAT</u> and is an excellent cook.</p>

Table 5.1: A sample text with English glosses, original (T0) and correction (T2). Errors are marked in the same way at both tiers: those corrected at T1 are in bold, errors corrected at T2 are underlined. (NEM_GD_008 ru B2)

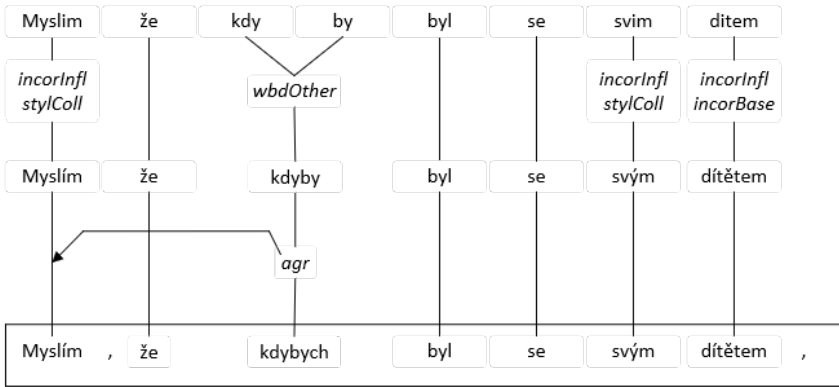


Figure 5.1: Example of the two-tier error annotation scheme

tiers, both between individual word forms and their annotations, while allowing for arbitrary joining and splitting of any number of non-contiguous segments.

In general, these labeled relations can link an arbitrary number of elements at one tier with an arbitrary number of elements at a neighboring tier. The elements at one tier participating in this relation need not form a contiguous sequence. Multiple words at any tier are thus identified as a single segment, which is related to a segment at a neighboring tier, while any of the participating word forms can retain their 1:1 links with their counterparts at other tiers. This is useful for splitting and joining word forms, for changing word order, and for any other corrections involving multiple words. Nodes can also be added or omitted at any tier to correct missing or odd punctuation signs or syntactic constituents.

The links are used for connecting tokens in the source transcript with their counterparts at the two successive tiers. The links can be labeled with the type of error. In this way, correspondences between successively emended forms are made explicit. Nodes at neighboring tiers are usually linked 1:1, but words can be joined (*kdy by* → *kdyby* as in Figure 5.1) or split, deleted or added. These relations can interlink any number of potentially non-contiguous words across the neighboring tiers. Multiple words can thus be identified as a single unit, while any of the participating word forms can retain their 1:1 links with their counterparts at other

tiers.

5.4.1.7 Error tags

Whenever a word form is corrected, the type of error can be specified as a label at the link connecting the incorrect form with its correction at the neighboring tier (such as `incorInfl` or `incorBase` for morphological errors in inflectional endings and stems, `stylColl` as a style marker, `wbdOther` as a word boundary error, and `agr` as an error in agreement).

We use two types of error tags: (i) “formal” tags describing the formal nature of the error (a letter is missing, extra diacritics, etc., see §5.4.5), and (ii) “grammar-based” tags attempting to capture the related grammatical phenomena (error in inflection, style, agreement, etc, see §5.4.2.)

In practice, error tags can be assigned manually or automatically. In the 2T annotation of *CzeSL*, most grammar-based tags are assigned manually in *feat* (see §7.3.1), while automatic procedures are used to assign the formal tags (see §5.4.5.2).

5.4.1.8 Morphosyntactic references

Some error types, such as a form incorrect due to violated rules of agreement or valency, may be complemented by simple syntactic annotation, linking the error label with a different form determining the correct version and further explaining the reason of the error. E. g., the subject or another form exhibiting the same agreement categories is the target of this type of link in case of a faulty finite verb form – in Figure 5.1, the link goes from the corrected form *kdyby* → *kdybych* ‘when.1sg’ to the form showing the agreement categories of 1st person singular, i. e., *myslím* ‘think.1sg’.

In this case, there is typically one correct form involved, e. g., the subject in subject-predicate agreement, the noun in adjective-noun agreement, the verb assigning case to a complement, the antecedent in pronominal reference, or – in the absence of an explicit agreement source (e. g., pro-drop subject in Figure 5.1) – a form sharing the same morphosyntactic categories. The incorrect form is corrected using a 1:1 link with an option to refer to a correct form at the same tier. More precisely, such “morphosyntactic references” lead from an error label at the cross-tier “error” link (e. g., the `agr` label) to another cross-tier link (not necessarily with an error label). They do not link the forms themselves to enable possible references from a multi-word unit (or) to another multi-word unit, represented as tokens linked by edges branching from points in between the tiers. For the last two words *pro mně*

→ *mi* ‘for me’ in Figure 5.2, such a reference is represented by the link originating in the error label **dep**.

Morphosyntactic references can be very useful in a text without syntactic annotation. In a parsed text, syntactic structure, i. e., one of the types linguistic annotation, can provide the same information.

5.4.1.9 Follow-up corrections

Corrections of morphosyntactic errors often result in follow-up (secondary) errors, as in (11), where a single error may result in multiple incorrect forms. The adjective *americkéěm* ‘American’ correctly agrees with the head noun in the locative case, but when the noun’s case is corrected to accusative, assigned by the preposition *na*, the case of the adjective must be corrected as well. Then multiple references are made: to the verb (or the preposition) as the case assigner for the noun, and to the noun as the source of agreement for the adjective, while the error of the form of the adjective is the result of the follow-up correction and it is marked as such.

- (11) T0: *Dívá se na *americkéěm *filmu.*
 looks REFL at American.*_{LOC} film.*_{LOC}
- T2: *Dívá se na americký́ film.*
 looks REFL at American._{ACC} film._{ACC}
- “She/He is watching an American movie.”

5.4.1.10 Alternative target hypotheses

It can be difficult to decipher the learner’s intention in a text which is partially incomprehensible or ambiguous. Even in a broader context, the meaning of a sentence may be entirely opaque or the sentence may allow for several interpretations. Moreover, an error can often be identified only in relation to a target hypothesis, while more than one such hypothesis may be available.

Annotation using multiple target hypotheses exists as a theoretical possibility, because the annotation format supports alternatives. On the other hand, the annotation tool does not support local disjunctions and the downstream processing of alternative annotation all the way until its accommodation in a corpus search tool, would be faced with considerably higher requirements due to the possible occurrence of multiple versions of 2T structures for a sentence or its parts. For such reasons, we refrained from allowing the option of alternative corrections.

In the transcription guidelines the policy is different. In fact, alternatives in the transcripts are used fairly often, even in cases where additional interpretations,

e.g., of a hand-written glyph, are unlikely. According to the adopted solution, alternatives are retained in the transcripts following a transcription markup, but only one of the alternatives (the first one) is treated as part of the annotated text.

In addition to alternatives in target hypotheses for a single error or text unit, there could be alternative interpretations of a single error even with a single target hypothesis. However, the manual annotation guidelines for the 2T scheme do not support alternatives in error categorization either. See §5.4.2.2 for more details.

5.4.2 Error tagset

5.4.2.1 Based on linguistic categories

The annotation scheme, including the error tagset design, answers the following requirements, based on the typological properties of native Czech, features of non-native Czech and goals of the project:

1. Preservation of the original text alongside target hypotheses
2. Successive corrections, allowing to identify various types of errors in a single form, such as *kdy by* ‘if would.3SG/PL’ → *kdyby* ‘if+would.3SG/PL’ → *kdybych* ‘if+would.1SG’ in Figure 5.1
3. Ability to capture errors in multi-word discontinuous expressions
4. Annotation of syntactic relations for some error types: agreement, valency, pronominal reference
5. Automatic assignment of errors wherever possible

The design of the taxonomy was preceded by research of frequent error types and reflects hypotheses about the acquisition of an inflectional language, implicit in contemporary teaching methods or explicitly described in textbooks of Czech as a foreign language, teachers’ experience presented in various papers and conference talks, and SLA research.

The resulting taxonomy anticipates errors attested in the texts and explicated in the teaching process, extended to reflect the grammatical system of Czech, so that deviations possible with respect to the system can also be captured. Conceptually, the taxonomy is based on standard linguistic categories, complemented by a classification of superficial alternations of the source text in the target hypothesis, such as the indication of a missing, redundant, faulty or incorrectly ordered element.¹²

¹²For the detailed annotation guide (in Czech) see Štindlová and Rosen (2012).

Given the linguistically tractable categorization of errors actually made by learners of Czech, anchoring the error taxonomy in standard linguistic categories is a natural step. In fact, this is how error taxonomies for learner corpora are often designed.¹³ Moreover, due to its more formal and objective status, a taxonomy defined in terms of standard linguistic categories has a better chance of being applied consistently throughout the annotation.

While a taxonomy based on linguistic categories has its merit in an established and independently motivated theoretical background, some categorial distinctions are still not accepted by all annotators as appropriate and/or well defined. Indeed, such categories are perceived as being concerned more with the analysis of standard language rather than with the analysis of deviations. If this proves to be the case, such a piece of error annotation can be replaced by linguistic annotation, assigned by an automatic tool, e. g., by substituting morphosyntactic references by a regular syntactic parse.

Overall, we are convinced that annotation guided by well-defined linguistic criteria is useful at least as a base for comparison with native speakers' language, for automatic (error) annotation, and for annotating additional aspects of the texts, such as communicative adequacy or style. As a further step towards a common ground for the comparison and as guidance for the annotators, grammatical and lexical aspects of the learner language are corrected and tagged to conform to the rules of SCz.

A doubly annotated pilot sample (about 10 thousand tokens) was evaluated for inter-annotator agreement to verify that the annotation scheme and taxonomy are sufficiently robust to be used in the corpus. Higher agreement was found for formally well-defined error categories, with satisfactory results even for categories requiring subjective judgment (see §5.4.4.2). The evaluation was later extended to all doubly annotated texts, i. e., to 175 thousand tokens (see §5.4.4.3).

In the resulting granularity of error taxonomy, we have anticipated errors in phenomena that are explicated to the learners of Czech in the teaching process and that we knew were occurring; in parallel with these anticipated phenomena, we systematically extended the taxonomy with respect to the grammatical system of Czech, so that we could capture deviations hypothetically possible with respect to the system.

¹³For some taxonomies used in previous projects see, e.g., Granger (2003a), Nicholls (2003), Izumi, Uchimoto, and Isahara (2005), Díaz-Negrillo and Fernández-Domínguez (2006), Lüdeling (2008), and Lüdeling and Hirschmann (2015).

5.4.2.2 Grammar-based vs. formal errors

For practical reasons we have abandoned the idea of alternative target hypotheses (see §5.4.1.10). However, this does not exclude the option to use alternative error categories for a single error with a single target hypothesis.

For example, some errors in spelling can also be interpreted as errors in morphemics and morphosyntax. It is hard to decide which of the interpretations is correct without some more research into the individual learner's competence in Czech. In the 2T scheme, the rule of thumb for choosing the appropriate interpretation is to prefer the “more sophisticated” error type, i.e., morphosyntax rather than spelling. As a result, the 2T scheme does not support alternatives in error categorization either. To compensate for this rather strict restriction, the grammar-based tags, for the most part manually assigned, are complemented by “formal” errors, assigned automatically. For the MD scheme, systematically supporting multiple interpretations, see §5.6.

A single incorrect form is cross-classified as belonging to one or more types in each of the following two classes:

- Grammar-based error types – a taxonomy classifying errors from a grammatical perspective: errors in spelling, morphology, word boundary, agreement, government, lexical issue, style, punctuation; these error types are similar to the “linguistic category classification” (James 1998, 104–113) or “linguistically based errors” (Lüdeling and Hirschmann 2015, 146)
- Formal error types – error types capturing the formal nature of an error without referring to possible underlying grammatical reasons: diacritics, capitalization, metathesis, missing element; also called “target modification taxonomy” (James 1998, 104–113) or “edit-distance based errors” (Lüdeling and Hirschmann 2015, 146)

Unlike the grammar-based types, the formal errors are detected by automatic tools by comparing the source forms with their corrections and the manually assigned tags. Yet the tools can also detect some of the grammar-based error types. Thus, errors can be identified in the following ways:

- manually
- automatically, by comparing the faulty and the corrected forms
- automatically, by subdividing certain manually assigned error tags, often on the basis of the relevant word forms, their morphological tags or lemmas

5.4.2.3 Extent of the annotated unit

In the 2T scheme, the minimal annotated text units are tokens. There are three exceptions:

1. For some error types the locus of the error is identified. To signal that a non-word is ill-formed, we manually distinguish an error in the stem (`incorBase`) from an error in the inflectional ending (`incorInfl`).
2. Word boundary errors are annotated by joining multiple tokens or splitting a single token.
3. The formal error tags are typically concerned with phenomena at the level of morphs or characters, but without specifying the exact location and only to the extent the error can be identified by an algorithm.

The exact locus of the error within a word is identified in the MD annotation scheme (see §5.6).

In the following, we first describe the grammar-based tags (§5.4.3). Then the grammar-based tagset is evaluated (§5.4.4). The formal error tags are discussed in §5.4.5.

5.4.3 Grammar-based tags

Below, we describe the grammar-based tagsets for each tier separately. The tagset consists of 22 error tags, 8 for T1, 11 for T2, and 3 that can be used at both tiers. After a brief discussion of the granularity of the tagset, we show a commented example of a sentence annotated with the tagset.

5.4.3.1 Errors at T1

Errors in individual word forms, treated at T1, include misspellings (also diacritics and capitalization), misplaced word boundaries but also errors in inflectional and derivational morphology and unknown stems – fabricated or foreign words. Except for misspellings, all these errors are annotated manually. The result at T1 is the closest correct form, which can be further modified at T2 according to context, e. g., due to an error in agreement or semantic incompatibility of the lexeme.

Table 5.2 lists the errors manually annotated at T1. Some error types (`stylColl`, `stylOther` and `problem`) are used also at T2. Some error categories, such as `incor`, have subtypes. While some of these subtypes are tagged manually (`incorBase` and

`incorInfl`), other tags are added automatically: in the absence of any other tag at T1, a corrected form satisfying some conditions is tagged as `incorOther`. The column with the heading “A” identifies whether the tag is assigned Manually or Automatically. The \uparrow column indicates the number of edges going upwards from the error label, located in between the tiers, in the source text direction, i. e., towards T0 from a T1 error or towards T1 from a T2 error. The \downarrow column indicates the number of edges going downwards from the error label towards the target hypothesis, i. e., towards T1 from a T1 error or towards T2 from a T2 error. The \leftrightarrow column identifies how many outgoing morphosyntactic references are allowed for a given error type.

Error type	Error subtype	Description	Example	A	\uparrow	\downarrow	\leftrightarrow
<code>incor</code>		incorrect form					
	<code>incorInfl</code>	incorrect inflection	<i>pracovají</i> → <i>pracují v továrně</i> ; <i>bydlím s matkoj</i> → <i>matkou</i>	M	1	1	0
	<code>incorBase</code>	incorrect word base	<i>lidé jsou moc měrní</i> → <i>mírní</i> ; <i>musíš to posvětlit</i> → <i>vysvětlit</i>	M	1	1	0
	<code>incorOther</code>	other incorrect forms	<i>rád pivuju</i> → <i>piju pivo</i>	A	1-n	1-n	0
<code>fw</code>		foreign, coined, unidentified word					
	<code>fwFab</code>	made-up, coined word	<i>pokud nechceš slyšet smášky</i> → <i>posměšky</i>	M	1	1	0
	<code>fwNc</code>	foreign word	<i>váza je na Tisch</i> → <i>stole</i> ; <i>jsem ve truong</i> → <i>škole</i>	M	1	1	0
<code>flex</code>		used with <code>fw..</code> to mark inflection	<i>jdu do shopa</i> → <i>obchodu</i>	M	1	1	0
<code>wbd</code>		wrong word boundary					
	<code>wbdPre</code>	prefix separated by a space or preposition without space	<i>musím to při pravit</i> → <i>přípravit</i> ; <i>veškole</i> → <i>ve škole</i>	M	1-2	1-2	0
	<code>wbdComp</code>	wrongly separated compound	<i>český anglický</i> → <i>česko-anglický slovník</i>	M	2-n	1	0
	<code>wbdOther</code>	other word boundary error	<i>mocdobře</i> → <i>moc dobře</i> ; <i>atak</i> → <i>a tak</i> ; <i>kdy kolí</i> → <i>kdykoli</i>	M	1-n	1-n	0
<code>styl</code>		colloquial, bookish, dialect					
	<code>stylColl</code>	colloquial form	<i>dobrej</i> → <i>dobrý film</i>	M	0-n	0-n	0-n
	<code>stylOther</code>	bookish, dialectal, slang, hyper-correct	<i>holka s hnědými očimi</i> → <i>hnědýma očima</i>	M	0-n	0-n	0-n
<code>problem</code>		problematic cases		M	0-n	0-n	0-n

Table 5.2: Grammar-based error tags at T1

5.4.3.2 Errors at T2

Corrections at T2 concern errors in agreement, valency, analytical forms, pronominal reference, negative concord, the choice of aspect, tense, lexical item or idiom, and also in word order. For the agreement, valency, analytical forms, pronominal reference and negative concord cases, there is usually a correct form, which determines some properties (morphological categories) of the faulty form. [Table 5.3](#) shows a list of error types manually annotated at T2. The automatically identified errors include word order errors and subtypes of the analytical forms error `vbx`.

5.4.3.3 Coarse-grained

In comparison to some other error tagsets, the taxonomy is relatively coarse-grained. There are several reasons:

- We assume that error annotation can be combined with linguistic annotation, both in queries or in statistical analysis. For example, an error in agreement need not specify further that the incorrect form is an adjective because that information is available from the POS tag. Linguistic annotation for the source and corrected forms is provided by automatic tools.
- Whenever the type of error can be determined from the way the incorrect form is corrected, the type is supplied by an automatic post-processing step (see [§5.4.5](#)).
- Ever since the first design of the two-tier annotation scheme, we expected to provide a more detailed classification of errors or a classification of errors from other perspectives later. So far, two additional error annotation schemes – the multidimensional (MD) scheme (see [§5.6](#)) and the implicit annotation scheme (see [§5.5](#)) – and one linguistic annotation scheme, based on the Universal Dependencies guidelines (see [§6.2](#)), have been designed and tested on a smaller corpus sample.
- Too detailed tags could be hard to apply consistently even by a single annotator. On the other hand, the usability of a tagset can be measured by the IAA score and depends on more factors than its granularity, such as well-defined denotation (see [§6.2.5](#)). Moreover, tags cross-classifying errors by using a clearly defined POS or grammar domain component can be easily applicable despite their large number in the tagset. Still, we see the relatively low number of error tags as both a practical and theoretical advantage.

Error type	Error subtype	Description	Example	A	↑	↓	↔
agr		violated agreement rules	<i>to jsou hezké→hezcí chlapani; Jana čtu→čte</i>	M	1	1	0-n
dep		error in valency	<i>bojí se pes→psa; otázka čas→času</i>	M	0-1	0-1	0-n
ref		error in pronominal reference	<i>dal jsem to jemu i jejím→jeho bratrovi</i>	M	1	1	0-1
vbx		error in analytical verb form or compound predicate		M	1-n	1-n	0-1
	cvf	analytical verb	<i>kluci jsou→∅ běhali</i>	M	1-n	1-n	0-1
	mod	modal or phase verb	<i>musíš přijdeš→přijít</i>	M	1-n	1-n	0-1
	vpn	compound predicates	<i>Petr má→je unavený</i>	M	1-n	1-n	0-1
rflx		error in reflexive expression	<i>dívá ∅→se na televizi; Pavel si→se raduje</i>	M	0-n	0-n	0-n
neg		error in negation	<i>nikdo to ví→neví; půjdu ne → nepůjdu do školy</i>	M	1-n	1-n	0-1
odd		redundant item	<i>Petr dělá→∅ čte</i>	M	1	0	0
miss		missing item	<i>není ∅→to tak dávno</i>	M	0	1	0
wo		wrong word order	<i>mají hezký velmi dům → mají velmi hezký dům</i>	A	1-n	1-n	0
lex		error in lexicon or phraseology	<i>jsem ruská→Ruska; dopadlo to přírodně→přirozeně</i>	M	0-n	0-n	0-1
use		error in the use of a grammar category	<i>pošta je nejvíc blízko → nejbliž</i>	M	1-n	1-n	0-1
sec		secondary error (supplementary flag)	<i>stará se o našich rybičkách → naše rybičky</i>	M	1-n	1-n	0-n
styl		colloquial, bookish, dialect					
	stylColl	colloquial expression	<i>viděli jsme hezký→hezké holky</i>	M	0-n	0-n	0-n
	stylOther	bookish, dialectal, slang, hyper-correct expression	<i>zvedl se mi kufr→žaludek</i>	M	0-n	0-n	0-n
	stylMark	redundant discourse marker	<i>no; teda; jo → ∅</i>	M	1	0	0
disr		intelligible or disrupted construction	<i>kratka jakost vyborné ženy → ?</i>	M	n	n	0
problem		supplementary label for problematic cases		M	0-n	0-n	0-n

Table 5.3: Grammar-based error tags at T2

5.4.3.4 An example of complex annotation

Splitting, joining and reordering words, together with the morphosyntactic references, may result in a complex network of both labeled and unlabeled links, as in

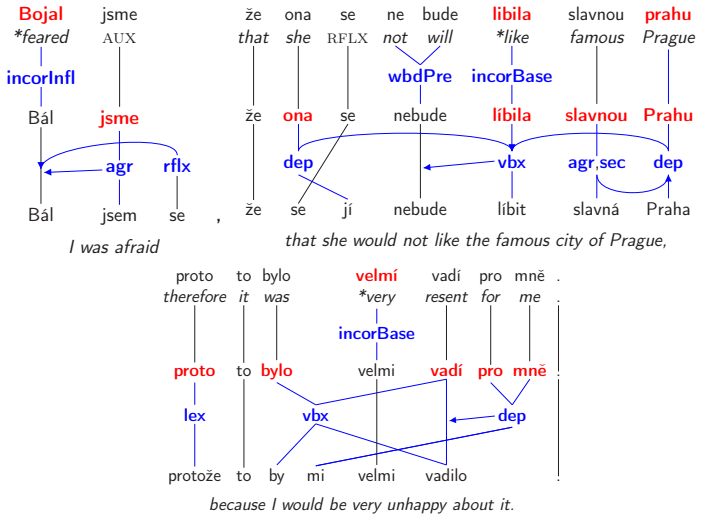


Figure 5.2: Two-level manual annotation of a sentence in *CzeSL*, the English glosses are added (VOB_KA_049 kk A1)

Figure 5.2 (re-drawn from the *feat* representation for clarity and to save space). It is an authentic sentence, split in two parts for space reasons.

- As in Figure 5.1, the three parallel strings of word forms represent the three tiers: the tier of transcribed input and the two annotation tiers. The tiers are parallel strings of word forms with links for corresponding forms. The asterisked forms in the English glosses below mark forms that are incorrect in any context, but they may be comprehensible – as is the case with all such forms in this example.
- Correct words are linked directly with their copies at T1, for corrected words most links are labeled with an error type. The first line is T0, imported from the transcribed original.
- T0 is followed by the level of orthographic and morphemic corrections (T1), where only forms incorrect in any context are treated. Errors at T1 are mainly non-word (OOV) errors while those at T2 are real-word and grammatical errors. However, a faulty form that happens to be spelled as a form which

would be correct in a different context, is still corrected at T1. Thus the result at T1 is a string consisting of correct Czech forms, even though the sentence may not be correct as a whole.

- All other types of errors are corrected at T2, representing a grammatically correct, though stylistically not necessarily optimal target hypothesis. A syntactic error label may be linked by a pointer to a word token, specifying an agreement, valency or referential relation.
- In the first (upper) part of the sentence, a form is labeled at T1 as an error in inflectional ending (**incorInfl**) *bojal* → *bál*, and another form as an error in the word stem (**incorStem**) *libila* → *l**í**bila*. The rest of the T1 errors are purely orthographic: according to the rules of Czech spelling, the negative particle *ne* is joined with the verb using the error label **wbdPre**, and the initial character of the place name is capitalized (the error label is assigned automatically).
- Staying with the first part of the sentence, at T2 another form is emended as an error in agreement (*jsme* → *jsem* ‘am’) with reference to a form exhibiting the correct morphological category of singular number (*bál*).
- The missing reflexive particle *se* is inserted with reference to the inherently reflexive verb and the comma is inserted without any label, because this type of error is identified automatically.
- The second reflexive particle *se*, a second position clitic, is misplaced in the source text and should be reordered (the **wo** label for a word-order error is assigned automatically).
- The pronoun *ona* ‘she’ in the nominative case is governed by the form *l**í**bit se* and should be assigned the dative case: *j**í*** ‘her’, with reference to the head verb.
- The head verb has changed its finite form *libila* into the infinitive, because it is now a part of the analytical future tense, identified by the error type **vbx** and a link to the future auxiliary.
- The accusative case of *Praha* in the source is changed into nominative, again with a reference to the governing verb. The form of the adjective *slavnou* must be modified accordingly with an additional label **sec** as a secondary (follow-up) error.

- The result could still be improved by positioning *Praha* after the clitics and before the finite verb *nebude*, resulting in a word order more in line with the underlying information structure of the sentence, but our policy is to refrain from more subtle phenomena and produce a grammatical rather than a perfect result.
- In the second (lower) part of the sentence, there is only one T1 `incorBase` error in diacritics (*velmí* → *velmi* ‘very’), but quite a few errors at T2.
- *Proto* ‘therefore’ is changed to *protože* ‘because’ as a lexical error (`lex`).
- The main issue in the second part of the sentence are the two finite verbs *bylo* ‘was’ and *vadí* ‘resents’. The most likely intention of the author is best expressed by the conditional mood. The two non-contiguous forms are replaced by the conditional auxiliary and the content verb participle in one step using a 2:2 relation. The intermediate node is labeled by `vbx` for complex verb forms.
- The prepositional phrase (PP) *pro mně* ‘for me’ is another complex issue. Its proper form is *pro mě* (homonymous with *pro mně*, but with ‘me’ bearing accusative instead of dative), or *pro mne*. The accusative case is required by the preposition *pro*. However, the head verb requires that this complement bears bare dative – *mi*. Additionally, this form is a second position clitic, following the conditional auxiliary (also a clitic) in the clitic cluster. The change from PP to the bare dative pronoun and the reordering are both properly represented, including the pointer to the head verb.
- What is missing is an explicit annotation of the faulty case of the prepositional complement, which is lost during the transition from T1 to T2. This is the price for a simpler annotation scheme with fewer levels. It might be possible to amend the PP at T1, but it would go against the rule that only forms wrong in isolation are corrected at T1.¹⁴

5.4.4 Evaluation of the manual tiered error annotation

To evaluate the consistency of annotation of learner corpora, texts are annotated independently by two or more annotators and the results are compared. This is an approach commonly used for many other types of manual annotation.

¹⁴The implicit annotation scheme (see §5.5) can accommodate a case like this due to a more flexible system of successive corrections. In the setup of correction levels for *CzeSL in TEITOK* (see §8.8), *mně* → *mě* would be corrected as an orthographic correction (`ort`) while *pro mě* → *mi* as a morphosyntactic correction (`gram`).

However, this was not always the standard practice in learner corpus research. The issue of singly annotated learner texts, used as application training data, was raised for the first time by Tetreault and Chodorow (2008), who investigated native-speakers’ classification of prepositions usage. They concluded that two native annotators performing the task of tagging errors in prepositions on the same text reach at best an agreement level on the border between moderate and substantial (their kappa value was $\kappa = 0.63$ – the metric is explained in §5.4.4.1 below). Rozovskaya and Roth (2010) also report low inter-annotator agreement ($\kappa = 0.16 - 0.40$) for the task of classifying sentences written by ESL learners. Meurers (2009) discusses the issue of verification of error annotation validity, viewing the lack of studies investigating inter-annotator agreement in the manual annotation of non-native speakers texts as a serious barrier for the development of annotation tools.

5.4.4.1 Inter-annotator agreement (IAA)

The manual annotation of *CzeSL* was evaluated using the metric κ (kappa, Cohen 1960), the standard measure of inter-annotator agreement, especially for tagged corpora. It is calculated as:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the observed agreement among the annotators, and $P(E)$ is the expected agreement, i. e., $P(E)$ is the probability that the coders agree by chance. The values of κ are within the interval $[-1, 1]$, where $\kappa = 1$ means perfect agreement, $\kappa = 0$ agreement equal to chance, and $\kappa = -1$ “perfect” disagreement.

The problem is to determine which error tags in one annotation correspond to which error tags in the other and how their scopes align. T0, the source text, is shared by both annotations. However, annotators might use a different target hypothesis, and thus the higher tiers can differ. Moreover, they often differ not only in the shape of tokens but also in their number. Because of this, we project error tags to T0 tokens and then calculate differences relative to that tier. When there are multiple tokens on T0 corresponding to a token on the relevant tier, we project the tag on the first T0 token only.

5.4.4.2 A pilot annotation

Early in the project, we calculated IAA on a pilot sample.¹⁵ It consisted of 67 texts totalling 9,848 tokens, most of them written by native speakers of Russian; the texts are classified according to the CEFR scale as A2 or B1. The sample was corrected and assigned error tags according to the error taxonomy presented above in §5.4.2 by 14 annotators. They were split into two groups. Each group annotated the whole sample independently. On average each annotator processed 1,475 words in 11 texts.

Table 5.4 summarizes the distribution of selected error tags for the pilot sample and for all doubly annotated texts available at the time of the evaluation. The first column gives the error tag; some tags (marked with an asterisk) are used only in the evaluation as a more general error category.¹⁶ The column headed by ‘avg tags’ gives the number of times the tag was used by an average annotator (calculated simply as the total for the two annotators divided by two).

For the comparison we only considered deviant (error-annotated) forms. Table 5.4 shows differences in error tags, ignoring potentially different THs. However, different THs are an obvious reason for disagreements in the error tags. For example, the THs were different in 54% of cases when the annotators did not agree on the use of the *agr* tag ($\kappa = 0.54$). The differences in THs were either on T1 (15%) or on T2 (39%). In a more extensive evaluation described below the impact of TH on the IAA in error tags was explored in more detail.

5.4.4.3 IAA on all doubly-annotated texts

Using the feedback gained from the pilot experiment we modified the definition of some tags, refined the annotation guidelines and improved the training of annotators. In a few cases we also slightly modified the error taxonomy. A substantially larger subset of the transcribed texts was annotated by 31 annotators in three groups specializing on Slavic, non-Slavic and Roma learners. The evaluation was extended

¹⁵For complete results see Štindlová et al. (2012). Note that when there are multiple tokens on T0 corresponding to a token on the relevant tier, the tag was projected to all such tokens rather than to the first T0 token. Also note that the numbers for *incorInf1* and *incorStem* are switched by mistake in the reported results.

¹⁶As described in §5.4.2, the error taxonomy is hierarchical: the error types are partitioned into domains, which are further divided into more specific subcategories, tagged manually or automatically. For example, the domain of complex verb form errors on T2 can be further specified as errors in analytical verb forms (*cvf*), modal verbs (*mod*), verbo-nominal predicates, passive or resultative form (*vnp*).

Tag	Type of error	Pilot sample		All annotated texts	
		κ	avg tags	κ	avg tags
incor*	incorBase+incorInfl	0.84	1,038	0.88	14,380
incorBase	Incorrect stem	0.75	723	0.82	10,780
incorInfl	Incorrect inflection	0.61	398	0.71	4,679
wbd*	wbdPre+wbdOther+wbdComp	0.21	37	0.56	840
wbdPre	Incorrect word boundary (prefix/prepos.)	0.18	11	0.75	484
wbdOther	Incorrect word boundary	–	0	0.69	842
wbdComp	Incorrect word boundary (compound)	0.15	13	0.22	58
fw*	fw+fwFab+fwNc	0.47	38	0.36	423
fwNc	Foreign/unidentified form	0.24	12	0.30	298
fwFab	Made-up/unidentified form	0.14	20	0.09	125
stylColl	Colloquial style at T1	0.25	8	0.44	1,396
agr	Agreement violation	0.54	199	0.69	2,622
dep	Syntactic dependency errors	0.44	194	0.58	3,064
rflx	Incorrect reflexive expression	0.26	11	0.42	141
lex	Lexical or phraseology error	0.37	189	0.32	1,815
neg	Incorrectly expressed negation	0.48	10	0.23	48
ref	Pronominal reference error	0.16	18	0.16	115
sec	Secondary (consequent) error	0.12	33	0.26	415
stylColl	Colloquial style at T2	0.42	24	0.39	633
use	Tense, aspect etc. error	0.22	84	0.39	696
vbv	Complex verb form error	0.13	15	0.17	233

Table 5.4: Inter-annotator agreement on selected tags

to all usable texts doubly-annotated so far, i.e., to 1,396 texts totalling 175,234 words.¹⁷

As a result, the reliability of the annotation has generally improved – see IAA for the whole doubly-annotated part of the corpus in Table 5.4. At the same time we are aware that if two annotations differ, it does not necessarily mean one of them is wrong. Language, especially the language of non-native learners, is fuzzy and ambiguous and we do not intend to cover up this fact by providing instructions aimed solely at high IAA just for the sake of it.

For example, one annotator might perceive the word *checkni* in *checkni moje stránky* ‘check my site’ to be a clearly non-Czech word (annotating it as **fwNc**), while another would consider it a colloquial form (annotating it as **stylColl**). In such cases, one might instruct the annotators to prefer a certain tag. However, even though this would lead to a higher IAA, it would conceal the fact that these

¹⁷For more details see Rosen et al. (2014).

expressions are perceived differently by different native speakers.

The table shows that on T1 the annotators tend to agree in the domain categories **incor*** and **wbd***, i. e., for incorrect morphology and for improper word boundaries ($\kappa > 0.8$ and $\kappa > 0.6$, respectively). IAA was lower ($\kappa < 0.4$) for categories with a fuzzy interpretation, where a target hypothesis is difficult to establish, such as **fw*** and its subcategories, used to tag attempts to coin a new Czech lexeme (**fwFab**), or foreign/unidentified strings of words (**fwNc**). Even the choice between the two subcategories was problematic, as can be seen from [Table 5.5](#).

At T2, the annotators agree in agreement errors (**agr**, $\kappa > 0.6$) and errors in expressing syntactic dependency (**dep**, $\kappa \sim 0.6$), and also in the well-defined category of errors in reflexive expressions (**rflx**, $\kappa \sim 0.4$). However, pronominal references (**ref**), secondary (follow-up) errors (**sec**) and – surprisingly – also errors in analytical verb forms / complex predicates (**vbz**) and negation (**neg**) show a very low level of IAA, even though they are identifiable by formal linguistic criteria. In all these four cases, the distribution of tags and the annotators’ feedback suggest that the annotation manual fails to provide enough guidance and formal criteria in distinguishing between the error types **ref** vs. **agr** and **ref** vs. **dep** (in either case the disagreement represents 19% of all the inconsistent uses of the tag **ref**).

IAA in the distribution of tags for usage and especially lexical errors is lower ($\kappa < 0.4$). The application of these tags is highly dependent on the annotator’s judgment, and the results are low as expected. An analysis has revealed that the tag **lex** has a systematic distribution: if the original lexeme and its ‘ideal’ correction differ in their meaning distinctly, the annotators agree in their corrections in most of the cases. On the other hand, if the lexemes show semantic proximity, the annotators highly disagree in the correction and therefore also in the consequent annotation.

Example (12) illustrates the former situation. The learner confused *housky* ‘buns’ with *housenky* ‘caterpillars’. However, the context leaves no doubt about the target hypothesis, as attested by the agreement of both annotators.

- (12) T0: *v pekařství kupuju *housenky*
 ‘I buy **caterpillars** in the baker’s shop’
 T2: A1: ... *housky*._{LEX} ‘buns’ ...
 A2: ... *housky*._{LEX} ‘buns’ ...

On the other hand in (13), the choice of the most appropriate lexeme for the correction is less obvious. There are several factors at work here: (i) even though *druhé* ‘other/second’ is ambiguous, *jiné* ‘other’ collocates with *kultury* ‘cultures’ more often than *druhé* ‘other/second’; (ii) maybe the author really meant to use the phrase

in the sense ‘second cultures’ – the context does not help here; (iii) there is probably no candidate word for a correction with a meaning vague enough to cover both ambiguous readings of the source lexeme. While both annotators tagged the error in agreement ending (**agr**), allowing for its interpretation as a proper ending in colloquial Czech (**stylColl**), Annotator 1 (A1), probably following the rule of minimal intervention, decided to stick with the source lexeme, while Annotator 2 (A2) picked the stylistically more appropriate lexeme. The additional **lex** tag reflects this difference between the two annotators.

- (13) T0: **kdýž se *divá na *druhý kultury*
 ‘when one looks at **other/second** cultures’
 T2: A1: ... *druhé*.AGR+STYLCOLL ‘second/other’ ...
 A2: ... *jiné*.LEX+AGR+STYLCOLL ‘other’ ... (KKOL_A3_008 ru B1)

Tables 5.5 and 5.6 present a confusion matrix for T1 and T2 error tags, respectively. The ‘?’ column/row covers cases when there were multiple tags provided by either annotator and they did not include the relevant tag (so we know that the annotators disagreed, but we cannot say which tags correspond to which). Note that the totals might be larger than the sums of the respective row or column as the table shows counts for selected tags only. Thus we can see, for example, that in 8,989 cases the annotators agreed on the **incorBase** tag, but in 400 cases Annotator 2 used the **incorInfl** tag instead. Far less common were cases when Annotator 2 assumed the error to be one of the **fw*** tags. Finally in 574 cases Annotator 2 used multiple tags, but none of them was **incorBase** (so we cannot say which one of those corresponds to the **incorBase** tag of Annotator 1).

From these tables, we can see that the annotators most commonly confused the following tags:

- **incorBase** (error in stem) for **incorInfl** (error in inflectional affix)

Most of these mismatches are cases where it is debatable whether the error occurred in the stem or in the inflection. The annotation manual often chooses one possibility, but either the annotators were not careful enough or some space for different opinions still remained. For example, all errors in root vowel changes should be considered **incorBase** errors, the logic being that most of the time, this is no longer a productive process (e.g., *práce* ‘work.SG.NOM’ but *prací* ‘work.PL.GEN’ – a remnant of the Indo-European ablaut). Some annotators marked such cases as errors in inflection.

- **fwNc** (foreign word) for **incorBase** (error in stem)

The word may look foreign to an annotator who knows the foreign language, but may seem to include a plain mistake to an annotator who does not know the language or just does not realize the foreign influence.

- **agr** (agreement error) for **dep** (valency error), less frequently for **lex** (lexicon error) or **vbx** (compound verb form error)

There are robust rules for **agr/dep/vbx** and to some extent even for **lex**, but they may not be easy for the annotator to apply. For example, annotators often mistagged quantifier errors. A quantifier shares its morphological case with the quantified NP whenever the quantifier is not a numeral five and above or an expression such as *mnoho* ‘many’ **and** the quantifier is not in a position assigning nominative or accusative. Otherwise the quantified NP is assigned the genitive case.

	incorBase	incorInfl	wbdPre	wbdOther	wbdComp	fwNc	fwFab	stylColl	?	Total
incorBase	8,989	400	5	4	0	21	19	3	574	10,866
incorInfl	450	3,379	1	4	0	4	6	2	555	4,797
wbdPre	2	0	363	23	3	0	0	0	39	488
wbdOther	7	1	16	580	6	2	1	0	98	855
wbdComp	3	0	3	5	13	0	0	0	8	58
fwNc	52	2	0	5	0	89	13	0	69	296
fwFab	15	2	0	1	0	17	11	0	44	119
stylColl	4	3	0	0	0	0	0	617	718	1,353
?	496	514	26	95	6	68	64	803	0	2,246
Total	10,694	4,561	481	830	58	300	131	1,439	2,254	

Table 5.5: Confusion matrix on T1 for all data (selected tags)

5.4.4.4 Error tags depend on target hypothesis

Analysis of the tagged data (see Table 5.7) shows that the disagreement in using error tags is not necessarily caused by an annotator’s fault, but could rather be dependent on the choice of the TH. For example, while **agr** has an overall agreement of 0.69, it is 0.82 for identical corrections, but only 0.24 if the target (T2) hypotheses are different.

The situation of other tags is similar. See (14) for an example. Besides two errors in spelling (*kdyz* → *když* ‘when’ and *stratil* → *ztratil* ‘lose.PAST’), the main

	agr	dep	rflx	lex	neg	ref	sec	stylColl	use	vbx	?	Total
agr	1,825	118	2	20	0	7	1	4	20	9	181	2,571
dep	180	1,790	9	105	0	10	0	0	60	8	289	3,130
rflx	3	6	59	4	0	4	0	0	0	3	13	130
lex	34	135	4	590	4	4	0	0	42	7	329	1,927
neg	0	0	0	3	11	0	0	0	1	2	16	54
ref	15	10	11	7	0	18	0	0	5	0	31	131
sec	0	0	0	0	0	0	108	0	0	0	330	440
stylColl	3	4	0	2	0	0	0	248	0	0	354	693
use	17	42	1	33	0	5	0	0	273	10	71	683
vbx	30	4	3	5	0	0	0	0	23	41	45	248
?	191	234	10	332	10	28	274	303	72	27	0	1,578
Total	2,674	2,998	152	1,704	42	100	390	573	708	218	1,715	

Table 5.6: Confusion matrix on T2 for all data (selected tags)

problem in the source version of the clause is the use of the verb. If *manžel* ‘husband’ is meant to be its subject, a reflexive particle is missing, which is what A2 assumes. If, on the other hand, ‘husband’ is the verb’s object, then the case ending of the object must be different (*manžela* ‘husband.ACC’), and the verb should agree with a pro subject in feminine gender (*ztratila* ‘LOSE.PAST.FEM’). Actually, A1 chose the present tense which does not show agreement gender (*ztratí* ‘loses’), but anyway, the two target hypotheses are not too far apart in their meaning, especially in a wider context, yet their structure and error annotation are very different.

(14) T0: *a *kdyz *stratil *manzel*

T2: A1: *a když ztratí.AGR manžela.DEP*
‘and when she loses her husband’

A2: *a když se.RFLX ztratil manžel*
‘and when the husband got lost’

(KKOL_AV_007 ru B1)

However, sometimes annotators arrived at identical target hypotheses, but still interpreted the original text differently, thus labeling it with different error tags. In some cases, this is manifested by different corrections on the lower tier, i. e., T1. For example, consider the expression in (15): both annotators corrected the non-existent word *tezki* to *těžké* ‘difficult’, but they differ in their interpretation of the original word. A1 interpreted it as an incorrectly spelled colloquial form *těžký* ‘difficult’ (*y* and *i* have the same pronunciation in Czech), correcting it to the official *těžké* on the next tier. A2 interpreted *tezky* simply as an incorrect form, and

corrected *tezky* directly to *těžké*. Both approaches make sense, and it is difficult to choose between them without knowing more about the language of the speaker.

- (15) T0: **tezki* *období*
 ‘difficult period’
- A1: T1: *těžký*.INCORSTEM+INCORINFL *období*
 T2: *těžké*.AGR+STYLCOLL *období*
- A2: T1: *těžké*.INCORSTEM+INCORINFL *období*
 T2: *těžké* *období* (KKOL_AV_001 ru B1)

In all these cases, tagging is correct vis-à-vis the selected target hypothesis. After an investigation of the impact of corrections on error annotation at the individual tiers, we can support the requirement of explicit target interpretation in the annotation scheme (Lüdeling 2008). The scheme can thus be verified by the calculation of IAA in the distribution of the tags, depending on the final hypothesis (cf., i.a., Meurers 2009).

5.4.4.5 Possible causes of the annotators’ disagreements

To summarize, we can identify the following causes of the annotators’ disagreements:

1. Invalid or imprecise annotation scheme – the annotators’ disagreement can be caused by the annotation scheme itself if it includes poorly defined or redundant tags or misses some needed tags. The pilot showed that it was problematic in several points, such as:
 - (i) poorly distinguished subtypes of word boundary error (**wbd**),
 - (ii) a fuzzy definition of the error in pronominal reference (**ref**), which also leaves some room for a confusion with the **agr** and **dep** types, and
 - (iii) an imprecise boundary between the error due to a wrong choice of verbal tense (**use**) and the error in the analytical verb form (**vbx**).
2. Insufficient screening and training of the annotators. The level of screening and training process has a significant effect on the IAA rate. Higher IAA was demonstrated for annotators exposed to extensive and detailed pre-annotation training. It would be interesting to test what kind of impact the annotators’ exposure to Czech as a foreign language has on the consistency of their annotation.

	tag	total		same emendations		different emendations	
		κ	avg. tags	κ	avg. tags	κ	avg. tags
T1	incor*	0.88	14,380	0.95	12,376	0.48	2,004
	incorBase	0.82	10,780	0.89	9,323	0.44	1,456
	incorInfl	0.71	4,679	0.79	3,887	0.36	791
	wbd*	0.56	840	0.71	525	0.33	315
	wbdPre	0.75	484	0.90	336	0.40	148
	wbdOther	0.69	842	0.90	479	0.41	363
	wbdComp	0.22	58	0.38	23	0.12	34
	fw*	0.36	423	0.45	235	0.24	187
	fwNc	0.30	298	0.31	165	0.28	132
	fwFab	0.09	125	0.13	70	0.04	55
	stylColl	0.44	1,396	0.51	1,088	0.20	307
T2	agr	0.69	2,622	0.82	2,050	0.24	572
	dep	0.58	3,064	0.71	2,303	0.19	760
	rflx	0.42	141	0.58	98	0.05	43
	lex	0.32	1,815	0.53	847	0.14	968
	neg	0.23	48	0.62	16	0.03	32
	ref	0.16	115	0.13	70	0.20	45
	sec	0.26	415	0.43	224	0.06	191
	stylColl	0.39	633	0.53	403	0.14	230
	use	0.39	696	0.61	399	0.10	296
	vbx	0.17	233	0.25	135	0.07	98

Table 5.7: IAA depends on emendation agreement

3. Different target hypotheses (see §5.4.4.4). Some annotations require a considerable amount of interpretation, while each annotator can have her/his own interpretation because of age, gender, education, etc. Moreover, in the case of multi-tier annotation, annotators can differ also on intermediate tiers, even though their target hypothesis might be identical. However, the annotation scheme of *CzeSL*, supporting corrections on both tiers, makes reasons for possible disagreements explicit.

5.4.5 Formal tags

5.4.5.1 Automatic extension and modification of error annotation

From the first designs of the 2T *CzeSL* error annotation, it was assumed that, following the manual annotation, some types of errors would be annotated automatically. The automatic annotation would add a different point of view on the errors, based on a simple comparison of the source and corrected words. Manual annotation of errors on T1 is relatively simple, and the automatic annotation of errors can give the user much more detailed information about the error and sometimes even the most likely cause of the error. For example, the word form *hřipku*, used instead of the correct form *chřipku* ‘flu.SG.ACC’, is probably not a case of a character omission, but an error in voicing (**formVcd1**) – the character *h* is prototypically used for the voiced phoneme /ɦ/, while the digraph *ch* for the voiceless phoneme /x/ ([h] and [x] form a voicing pair in Czech). The manual annotation of this error only assigns a simple distinction that the word contains an unspecified error in the stem (flective base of the word): **incorBase**.

The automatic error annotation in *CzeSL* on T1 addresses errors of a formal nature, i. e., (broadly) orthographic errors, such as incorrect capitalization, wrong use of diacritics or wrong choice of the characters $i \leftrightarrow y$, and errors reflecting wrong pronunciation, such as voicing or (so called) hard and soft consonants, e. g., $d \leftrightarrow d'$: the character d' (d with a caron) marks in Czech the phoneme /j/ (voiced palatal plosive). Apart from formal error detection, the automatic error annotation refines the error tagging of errors in word boundaries (incorrect division or joining of word forms).

The T1-formal errors make no distinction whether the error occurs in the stem or in the inflectional affix; this distinction is assumed to come from the manual annotation: **incorBase** or **incorInfl**. The reason is that morpheme boundaries are often blurred in Czech (as in many other inflective languages) and it is not trivial to distinguish errors in inflection from errors in the stem automatically.

Some T2 errors are annotated automatically to some extent – the annotators mark errors with a more general tag and an automatic procedure automatically assigns a detailed tag.

The automatic annotation of T2 uses only a limited amount of information about morphological tags and lemmas. We do not attempt to identify the cause of the error. A form may be incorrect due to an incorrectly applied morphological paradigm or due to an incorrect syntax structure of the sentence. Instead, we only further specify or complement manually assigned tags. For example, an error in a periphrastic verb form is manually corrected and labeled with a general error tag,

the automatic annotation then adds a more detailed tag distinguishing past-tense errors and modal-verb errors.

5.4.5.2 Automatic detection of formal errors on T1

Automatic extension of the error annotation on T1 is performed for those T0 forms that are corrected on T1. It is based on a comparison of the source T0 form and the corrected T1 form. The result is the assignment of formal error tags: broadly orthographic errors, pronunciation-related errors. All such tags are prefixed by the **form...** string. Moreover, manually annotated errors in word boundaries (**wbd**) are specified as words incorrectly joined or split. For a list of all formal error tags on T1 see [Table 5.8](#) and [Table 5.9](#) (page 106–107).

The algorithm identifies individual differences between the corresponding T0 and T1 forms (*delamé* → *děláme* ‘make.1PL’ contains three individual differences: *e* → *ě*, *a* → *á*, *é* → *e*) and assigns an error tag to each difference. The 2T *CzeSL* annotation scheme does not track the exact location of error, therefore error tags are assigned to the whole word. If there are multiple errors, the word is assigned multiple tags (*delamé* → *děláme*: **formCaron0|formQuant0|formQuant1**). A single difference can also be assigned multiple error tags (the difference *i* → *ý* in *úteri* → *úterý* ‘Tuesday’ is assigned **formY0|formQuant0** to mark the confusion of *i* → *y* and a missing accent *y* → *ý*).

For the formal tags, we use the following convention: they end with 0 for incorrectly missing phenomena and 1 for incorrectly realized phenomena. For example, incorrect spelling of words due to voicing assimilation can be marked either with **formVcd0** or with **formVcd1**: *pohátková* → *pohádková* ‘fairytale.ADJ’ uses voiceless *t* instead of the correct voiced *d* and is therefore marked with **formVcd0**, while *svadba* → *svatba* ‘wedding’ uses voiced *d* instead of the correct voiceless *t* and is therefore marked with **formVcd1**.

For expository reasons, we classify automatically assigned formal tags into two groups: (broadly) orthographic errors, and formal errors affecting pronunciation. Orthographic errors concern misspellings that do not affect pronunciation of the misspelled form by native speakers. They are often the types of errors that even native speakers commonly make in their texts. Formal errors affecting pronunciation include errors in which there would be a noticeable difference in pronunciation between the original and the corrected form. The most common errors of this type include missing diacritics indicating the quantity of vocals or softening of consonants. We list only errors that actually occurred in student corpora; other errors are handled by the algorithm as well, but they are not present in the examined texts.

5.4.5.3 Formal orthographic errors

Automatically identified orthographic errors include errors in capitalization. Czech uses capital letters to mark sentence beginnings and proper names. Rarely, they are also used for emphasis of certain words. The error tag **formCap1** marks words that should be written with a lower-case letter, but are capitalized (*ona Rozumí* → *rozumí trochu český* ‘she Understands → understands little Czech’). Missing capitalization is labeled with **formCap0** (*Libí se mi praha* → *Praha* ‘I like prague → Prague’).

Another group of orthographic errors are errors in the spelling of vowel groups with *ě*. In Czech, some phoneme sequences can be spelled in two ways: /bje/ as *bě* or *bje*, /pje/ as *pě* or *pje*, /vje/ as *vě* or *vje*, /mje/ as *mě* or *mně* (the spelling depends on the origin of the word, for example, *bje*, *pje*, *vje* are used across a morpheme boundary). We distinguish formal errors in the spelling of the phonemes /je/ (*rozbjehl* → *rozběhl* ‘started to run’: **formJe1**) and errors in the spelling of /mɛ/ (*vzpoměla* → *vzpomněla* ‘remembered’ **formMne0**, *mněla* → *měla* ‘had’ **formMne1**).

A similar phenomenon applies to the orthographic notation of phonetic groups with palatal plosives /c/ (*t*), /ʃ/ (*d*) and palatal nasale /ɲ/ (*ň*). These phonemes are spelled with a diacritic mark caron (wedge, hacek), but in a combination with the vowel *e*, they are spelled as *tě*, *dě*, *ně* and in a combination with the vowel *i*, *í* as *ti*, *di*, *ni*, and *tí*, *dí*, *ní*, respectively. The spelling with a caron on the consonant (*kuchyně* → *kuchyně* ‘kitchen’) is wrong and is labeled with the error tag **formDtn**.

Another error concerns the marking of length of the vowel *u*. In Czech, two variants with the same pronunciation are used: *ú* (*u* with an acute diacritic) and *ů* (*u* with a ring diacritic). Simplifying somewhat, *ú* is used word and morpheme initially, and *ů* otherwise. Mistakes of this nature (*dúm* → *dům* ‘house’, *úkol* → *úkol* ‘task’) are labeled with **formDiaU**.

Non-native speakers sometimes make mistakes in spelling of the vowel *i* followed by a syllable boundary and another vowel. In original Czech words, the vowels are always separated by *j*, both in spelling and pronunciation. Borrowed words are often spelled without *j*, even though the glide is present in a correct pronunciation *j* (*piano* ‘piano’ is orthoepically pronounced as *píjáno*). Non-native speakers sometimes confuse this rule: we label with **formEpentJ0** an incorrect omission of *j* (*přiela* → *přijela* ‘arrived’, *žiou* → *žijou* ‘live.3PL’, *pieme* → *pijeme* ‘drink.1PL’), and with **formEpentJ1** a superfluous *j* (*fotografijemi* → *fotografiemi* ‘photos.INST’, *studijum* → *studium* ‘study’, *dijamant* → *diamant* ‘diamant’).

5.4.5.4 Formal errors sometimes influencing pronunciation

Several automatically identified types of formal errors are at the boundary between orthographic errors and errors affecting pronunciation. In some contexts, the error may affect pronunciation, in others it is purely orthographic, but we did not see any benefit in introducing another group of errors.

The confusion of the homophonous letters $i \leftrightarrow y$, which are both used to spell the vowel [ɪ], is labeled with **formYO** when replacing y with an incorrect i (*kdiž* → *když* ‘when’) or y with $í$ (*svími* → *svými* ‘self’s’), or with **formY1** when i is replaced with an incorrect y (*hystorek* → *historek* ‘stories.GEN’), or $í$ with $ý$ (*ostatným* → *ostatním* ‘others’). This error affects pronunciation in combination with the characters t , d and n , in the other cases it is a purely orthographic error. In the words of Czech origin, y never follows $č$, $ř$, $š$, $ž$, j , and i never follows h , ch , k , r . After some characters (b , f , l , m , p , s , v , z), both i and y are commonly used in Czech; in some cases, the use of the character distinguishes the meaning of homophones (*bíl* ‘beated’ vs. *byl* ‘was’).

The pronunciation of the word-initial character j preceding certain consonants (s , m , d) is optional (the prescribed pronunciation of *jsem* ‘am’ is [sɛm]). In colloquial use, such words are often written without the initial j , even though it often distinguishes meaning (*jsem* ‘am’ vs. *sem* ‘here’). Such words in learners texts are given the error tag **formProtJ0**. Similarly, words with an incorrectly added initial j (*jsi* → *si* ‘REFL.DAT’) are labeled with **formProtJ1**. In these cases (significantly predominant in terms of frequency), this error is orthographic, however, the error tag is used for any missing or redundant j before any consonant at the beginning of a word. Therefore, it is also used for errors such as *jšla* → *šla* ‘went.FEM’ where the change of pronunciation is possible.

Errors at the border between orthographic errors and errors affecting pronunciation also include errors in voicing. In Czech, consonants in clusters assimilate in voicing, but their spelling preserves voicing based on their morphological and phonological structure (the word *fotbalista* ‘soccer player’ is pronounced as a [fɔbɔlɪstɔ] due to voicing assimilation of /t/ with the voiced /b/). Errors caused by “phonetic” spelling are labeled as **formVcd0** (replacing a voiced character with its voiceless counterpart, e. g., *skouší* → *zkouší* ‘tries’) or **formVcd1** (replacing a voiceless character with its voiced counterpart *fɔbɔlɪstɔ* → *fɔtɔbɔlɪstɔ* ‘soccer player’). Czech voiced obstruents are devoiced word-finally, but they also preserve their voicing in spelling. Incorrect use of voiceless consonants in such cases is labeled with **formVcdFin0** (*kdyš* → *když* ‘when’). Prepositions ending in a voiceless consonant optionally assimilate with the voiced consonant of the following word (*přes hodinu* ‘over an hour’ is pronounced as *přezhodinu* or *přeshodinu*), but their spelling does not change either.

The errors such as *přez* → *přes* ‘over’ are labeled as **formVcdFin1**. However, we use the **formVcdFin1** tag for any incorrect use of a word-final voiced consonants, even for those that are not related to preposition voicing assimilation, including errors that a native speaker is unlikely to make (*svěd* → *svět* ‘world’). The remaining errors caused by an incorrect use of voiced consonants instead of voiceless ones and vice versa, which are not related to consonant cluster voicing assimilation and are not word final, are labeled as **formVcd** (*přehod* → *přechod* ‘crossing’, *sůstala* → *zůstala* ‘stayed.FEM’).

Sometimes, under the influence of the spelling rules in other European languages, *j* is incorrectly replaced with *y* (*yá* → *já* ‘I’, *žíyu* → *žiju* ‘live.1SG’, **formYJ0**), and *k* with *c* (*clientovi* → *klientovi* ‘client.SG.DAT’, *culturu* → *kulturu* ‘culture.SG.ACC’ **formCK0**). The words would be pronounced incorrectly if we followed the rules of Czech pronunciation, but it is likely that the author pronounces it correctly and just used an incorrect spelling. In Czech, the letter *y* is always pronounced as the vowel [i], never as the glide [j]. The character *c* in Czech words (original Czech words and not recent borrowings) is always pronounced as /ts/ (voiceless alveolar affricate). It is used for /k/ only in recently borrowed words.

Another phenomenon that includes both orthographic and pronunciation changes involves double phonemes. In Czech, double consonants are sometimes pronounced as two phones and sometimes as one. Double pronunciation is used especially for vowels separated by a morpheme boundary (*poloostrov* ‘peninsula’, *individu* ‘individual.DAT’). Often they are separated by a glottal stop ([ʔ]). Double pronunciation of consonants is also sometimes used to distinguish meaning (*racci* ‘seagulls’ vs. *raci* ‘crayfish’). Two identical consonants when one is in a prefix and another in a root are also often pronounced as two (*oddálit* ‘postpone’, *dvojjazyčný* ‘bilingual’), but not always (*leccos* ‘all sorts of things’). Double pronunciation across a root-suffix boundary is rather rare (*vyšší* ‘taller’, *činnost* ‘activity’, *babiččin* ‘grandma’s’). Errors in spelling of double and single letters are labeled with **formGemin**: **formGemin0** is used for characters that should be doubled but are not (*povinnost* → *povinnost* ‘duty’, *poloostrov* → *poloostrov* ‘peninsula’), and **formGemin1** is used for consonants that are doubled but should not be (*sobbota* → *sobota* ‘Saturday.ACC’, *rukoppis* → *rukopis* ‘manuscript’). The designation of error is slightly misleading as Czech does not have a real gemination.

5.4.5.5 Formal errors influencing pronunciation

One of the types of automatically identified formal errors affecting pronunciation (by native speakers) are errors caused by inappropriate writing of diacritics. Czech, has three diacritical marks: caron (wedge, hacek), acute accent and ring. Acute accent

and ring indicate a vowel is long (*síla* ‘silo.PL’ vs. *síla* ‘force’, *půl* ‘half.IMPER’ vs. *půl* ‘half’). Caron indicates so-called softening of consonants, i. e., shifting the place of articulation of alveolar consonant backwards: to postalveolar as in *z* /z/ → *ž* /ʒ/ or to palatal as in *d* /d/ → *ď* /ɟ/. The pronunciation of the *ě* character depends on the previous consonant. Errors in vowel quantity are labeled with **formQuant**: missing diacritics with **formQuant0** (*libí* → *líbí* ‘likes’), extra diacritics with **formQuant1** (*vyprávěl* → *vyprávěl* ‘narrated’). A missing caron is labeled with **formCaron0** (*pojď* → *pojd* ‘come.IMPER’, *neco* → *něco* ‘something’), a superfluous caron is labeled with **formCaron1** (*kteřých* → *kterých* ‘which’; *věnkově* → *venkově* ‘country’). The incorrect use of caron instead of acute or vice versa above the character *e* is labeled with **formDiaE** (*možně* → *možné* ‘possible’, *obchodé* → *obchodě* ‘shop’).

Palatalization is a historically motivated consonant alternation. Velar and glottal consonants (*k*, *h* /fi/, *ch* /x/, and *g* in borrowed words) change when followed by a morpheme originally containing the vowel yat, typically realized as *e*, *ě* or *í* in modern Czech: *k* → *c* /ts/, *h* /fi/ → *z*, *ch* /x/ → *š* /ʃ/, *g* → *z*. Errors in palatalization occur most often in the declension of nouns whose stem ends in one of the above consonants. For example, the paradigm *žena* ‘woman’ has the ending *-ě* in dative and local singular. The noun *řeka* ‘river’ belonging to this paradigm has the form *řece* (stem-final *k* changes to *c*, and the spelling of the ending *-ě* changes to *-e* in these cases). Missing palatalization is labeled with **formPalat0** whether the author uses *-ě* (*řekě* → *řece*) or *-e* (*řeke* → *řece*). Non-native speakers sometimes make also the opposite error: applying palatalization in places where it should not be: *koníčkem* ‘hobby.INST’ has the ending *-em* that historically does not contain yat and thus there is no palatalization. All cases of unjustified palatalization are assigned the error tag **formPalat1** (*koníčcem* → *koníčkem* ‘hobby.INST’, *pracovnícem* → *pracovníkem* ‘worker.INST’).

The following formal error also has a historical connection. Yers, Proto-Slavic vowels, disappeared in Old Czech, some transforming into *-e-* and some disappearing completely. This is the cause of alternating forms with and without *-e-* (*pátek* ‘Friday.NOM’ – *pátku* ‘Friday.GEN’). Synchronically, this is manifested as an epenthesis *-e-* making it easier to pronounce some words that would otherwise contain a consonant cluster both morpheme internally (*kra* ‘iceberg.NOM’ – *ker* ‘icebergs.GEN’ not *kr*) and across morpheme boundaries (*roz* + *brát* – *rozebrat* ‘take apart’). Forms with a missing *-e-* are labeled with **formEpentE0** error (*odbereme* → *odebereme* ‘remove.1PL’). However, the error is defined broadly: it applies to any missing *-e-* between two consonants, and most occurrences of this error are thus only loosely related to the original *-e-* epenthesis: *odpoldne* → *odpoledne* ‘afternoon’, *přijla* → *přijela* ‘arrived.FEM.SG’, *televizi* → *televizi* ‘TV.ACC’. An extra *-e-* between two consonants is labeled with **formEpentE1**. This is used both in cases where *-e-* occurs in

other forms of the paradigm (*dářeky* → *dárky* ‘presents’ cf. *dárek* ‘present’, *páteku* → *pátku* ‘Friday.GEN’, cf. *pátek* ‘Friday.NOM’), and in cases that are due to pronunciation difficulty of consonant clusters (*jmenuje* → *jmenuje* ‘is named’, *čtvrtek* → *čtvrtek* ‘Thursday’).

In spoken Czech, in Bohemia and Central Moravia, word-initial *o-* is often preceded with a prothetic *v*. For example, some speakers pronounce the word *okno* ‘window’ as *vokno* and sometimes they even write it in that nonstandard way (but the phenomenon is probably currently declining). In the *CzeSL* project, we evaluate the written text against the rules of SCz, so we label occurrences of prothetic *v* as errors with the **formProtV1** tag: *vobčas* → *občas* ‘sometimes’, *vpravdu* → *opravdu* ‘really’.

During the evolution of Czech from Proto-Slavic, the original phoneme *g* changed into *h*. However, this process did not occur in most other Slavic languages. This is a cause for another type of error when non-native speakers mostly of Slavic origin confuse the letters *g* and *h*. The incorrect use of the letter *g* in place of *h* is labeled with **formGH0**: *glavní* → *hlavní* ‘main’, *mного* → *mnoho* ‘many’, *gasič* → *hasič* ‘firefighter.INST’. Czech uses *g* in newly borrowed words. An incorrect replacement of such *g* with *h* is labeled with **formGH1**: *ciharetu* → *cigaretu* ‘cigarette.ACC’, *prohramů* → *programů* ‘programs.GEN’, *hrafička* → *grafička* ‘graphic artist.FEM’.

Character metathesis is a relatively common type of error for both non-native and native speakers. We automatically identify two types of metathesis: swapping adjacent characters (*sulnce* → *slunce* ‘sun’, *dobrodružství* → *dobrodružství* ‘adventure’), and swapping two characters separated by another character (*provůdce* → *průvodce* ‘guide’, *ojelů* → *olejů* ‘oil.PL.GEN’, *zicích* → *cizích* ‘foreign.PL.GEN’). These errors are labeled with **formMeta** error tag.

5.4.5.6 Other types of errors

The variability of errors in the texts of non-native speakers is too great, so it is not possible to systematically handle all cases. In this section we focus on automatically assigned tags that cannot be classified into any of the above categories. The tags attempt to provide at least some information about the nature of the difference in the original and corrected word. They almost always affect pronunciation.

There are three error tags for labeling single-character mistakes that cannot be classified with any of the more descriptive tags above. The **formSingCh** tag indicates cases where one character is replaced by another: *specifické* → *specifické* ‘specific.NEUT’, *existije* → *existuje* ‘exists’, *ofjevit* → *objevit* ‘appear.INF’. The **formMissChar** tag is assigned to cases where a single character is missing: *učiteka* → *učitelka* ‘teacher.FEM’, *výjmečnému* → *výjmečnému* ‘exceptional.MASC.DAT’, *zбудil*

→ *vzbudil* ‘woke up’. The **formRedunChar** tag is used in cases with an extra character: *usmrdcení* → *usmrčení* ‘killing’, *prvního* → *prvního* ‘first’, *kugličky* → *kuličky* ‘marbles’.

Another error tag marks mistakes due to a missing or extra prefix. Therefore, the error tag expresses that there are one or several characters missing or extra word-initially and that the characters are equal to one of the commonly used Czech prefixes. Errors where the prefix is missing in the original word are labeled with **formPre0**: *hledu* → *pohledu* ‘view.GEN’, *žaduje* → *vyžaduje* ‘requires’, *znamil* → *seznámil* ‘introduced.MASC.SG’. Words with extra prefixes in the original are labeled with **formPre1**: *pojet* → *jet* ‘drive.INF’, *přezačít* → *začít* ‘start.INF’, *potravíme* → *trávíme* ‘spend.1PL’. In some cases, the tag is also assigned to incorrectly fused words that were manually annotated in a wrong way: the incorrectly fused word *semnou*, corrected to *se mnou* ‘with me’, should be labeled with the error tag **wbdPreJoined** and linked with each of the T1 words *se* and *mnou*. But because it was only linked to the word *mnou*, the automatic annotation incorrectly assigns the error tag **formPre1**.

Word-initial errors where the difference cannot be classified as a common Czech prefix are labeled with **formHead** tags. The tag **formHead0** is used for missing word-initial characters (*busovou* → *autobusovou* ‘bus.ADJ’), **formHead** is used for different word beginnings (*prověděl* → *dozvěděl* ‘learned’, *chiny* → *Číny* ‘China.GEN’) and **formHead1** for extra characters. However, the last situation is always a result of errors in manual annotation: incorrectly fused words were properly corrected (*conejlíp* → *co nejlíp* ‘as good as possible’) but instead of splitting the original T0 word into two (or more) T1 words, linked to the source and labeled together with the **wbdOtherJoined** tag, one of the T1 words was labeled as a correction of the original and the other word was inserted as a missing word.

The opposite case, when the original and corrected words start in the same way but end differently, is labeled with **formTail** tags (**formTail0**, **formTail1**, **formTail**). The **formTail0** tag is used when the original word is missing some characters at the end (*t* → *tam* ‘there’, *ž* → *žit* ‘live’; there are only few meaningful examples in the corpus), **formTail** is used for words with different ends (*několiku* → *několika* ‘several.GEN’, *šansu* → *šanci* ‘chance.ACC’), **formTail1** for extra characters (no meaningful examples).

Even less information is contained in the **formLen** tags, which are used for words that (1) differ significantly (but that still do not cross the threshold when we give up on marking differences), and (2) that they also differ in length. The tag **formLen0** is used when the original word is shorter than the corrected word (*nákem* → *nějakém* ‘some’, *diš* → *když* ‘when’), and **formLen1** when the original word is longer (*vidňanami* → *Vídeňany* ‘Viennese’, *recat* → *říct* ‘say’).

Cases when the original T0 word significantly differs from the corrected T1 word are labeled with the `formUnspec` error tag: *omevy* → *umývá* ‘washes’, *choubů* → *hub* ‘mushrooms.GEN’, *ěště* → *ještě* ‘still’. In retrospect, we should not have introduced the tags `formLen`, `formTail` and `formHead` – for words where partial differences cannot be easily automatically identified, it would be better to give up recognition completely and always use the `formUnspec` tag.

5.4.5.7 Automatic classification of word-boundary errors

Word-boundary errors include words either incorrectly fused (*semsi* → *sem si* ‘AUX.1SG REFL.DAT’) or split (*ne chodila* → *nechodila* ‘wasn’t going’). During correction, these errors are manually labeled with `wbd` tags: `wbdPre` is used for prepositions fused with the following words and separated prefixes, `wbdOther` is used for other word-boundary errors. The automatic procedure adds a tag to mark whether the forms were incorrectly separated (`wbdPreSplit` / `wbdOtherSplit`) or incorrectly joined (`wbdPreJoined` / `wbdOtherJoined`).

5.5 Implicit error annotation

Either of the two components of error annotation (classification and correction) may be omitted. The decision to refrain from assigning error tags or from providing correct forms speeds up manual error annotation. However, such a decision can also be made due to theoretical reasons.

Some authors intentionally avoid categorizing errors, advocating correction as sufficient error annotation. They see categorization as an interpretation model, influencing access to the data, while correction is viewed as an implicit explanation for the errors (Fitzpatrick and Seegmiller 2004; Mendes et al. 2016). Notwithstanding this theoretical argument, if correction is the only approach to error annotation, its advantage is the easier task of the annotator due to the absence of an error classification scheme (Fitzpatrick and Seegmiller 2001). The annotator does not need to learn any classification rules, which speeds up the annotation task and avoids misclassification.

On the other hand, corrections without error labelling may not be sufficient to describe the error properly or substantiate the correction. The resulting annotation could then be too vague for specific queries or analysis by quantitative or statistical methods. As a compromise, corrections could be specified for specific annotation tiers, resulting in an implicit error classification, with an option to derive error tags automatically (Rio and Mendes 2019).

Error type	Error description	Example
Cap0	capitalization: lower→upper case	<i>evropě</i> → <i>Evropě</i> ; <i>štědrý</i> → <i>Štědrý</i>
Cap1	capitalization: upper→lower case	<i>Staré</i> → <i>staré</i> ; <i>Rodině</i> → <i>rodině</i>
Je0	<i>ě</i> → <i>je</i>	<i>ubjehlo</i> → <i>uběhlo</i> ; <i>Největší</i> → <i>Největši</i>
Je1	<i>je</i> → <i>ě</i>	<i>vjeděl</i> → <i>věděl</i> ; <i>vjeci</i> → <i>věci</i>
Mne0	<i>mě</i> → <i>mně</i>	<i>zapoměla</i> → <i>zapomněla</i>
Mne1	<i>mně, mňe, mňě</i> → <i>mě</i>	<i>mněla</i> → <i>měla</i> ; <i>rozumněli</i> → <i>rozuměli</i>
Dtn	<i>de, te, ňe; di, ti, ni</i> → <i>dě, tě, ně; di, ti, ní</i> :	<i>kuchyně</i> → <i>kuchyně</i> ; <i>vyměnit</i> → <i>vyměnit</i>
DiaU	diacritics: <i>ú</i> ↔ <i>ů</i>	<i>nemůžeš</i> → <i>nemůžeš</i> ; <i>úkoly</i> → <i>úkoly</i>
EpentJ0	missing <i>j</i> between vowels	<i>pieme</i> → <i>pijeme</i> ; <i>žiou</i> → <i>žijou</i>
EpentJ1	superfluous <i>j</i> between vowels	<i>fotografijemi</i> → <i>fotografiemi</i> ; <i>dijamant</i> → <i>diamant</i>
Y0	<i>i</i> → <i>y</i> ; <i>í</i> → <i>ý</i>	<i>kdiž</i> → <i>když</i> ; <i>svími</i> → <i>svými</i>
Y1	<i>y</i> → <i>i</i> ; <i>ý</i> → <i>í</i>	<i>hystorek</i> → <i>historek</i> ; <i>ostatným</i> → <i>ostatním</i>
ProtJ0	protethic <i>j</i> : missing <i>j</i>	<i>sem</i> → <i>jsem</i> ; <i>menoval</i> → <i>jmenoval</i>
ProtJ1	protethic <i>j</i> : extra <i>j</i>	<i>jse</i> → <i>se</i> ; <i>jmé</i> → <i>mé</i>
Vcd0	voicing assimilation: incor. voiced	<i>stratíme</i> → <i>ztratíme</i> ; <i>nabítka</i> → <i>nabídka</i>
Vcd1	voicing assimilation: incor. vcleless	<i>zbalit</i> → <i>sbalit</i> ; <i>nigdo</i> → <i>nikdo</i>
VcdFin0	word-final voicing: incor. voiceless	<i>kdyš</i> → <i>když</i> ; <i>vztach</i> → <i>vztah</i>
VcdFin1	word-final voicing: incor. voiced	<i>přez</i> → <i>přes</i> ; <i>pag</i> → <i>pak</i>
Vcd	voicing: other errors	<i>protoše</i> → <i>protože</i> ; <i>hodili</i> → <i>chodili</i>
YJ0	<i>y</i> → <i>j</i>	<i>yá</i> → <i>já</i> ; <i>žiyu</i> → <i>žiju</i>
CK0	<i>c</i> → <i>k</i>	<i>clientovi</i> → <i>klientovi</i> ; <i>cultura</i> → <i>kultura</i>
Gemin0	incor. single char. instead of double	<i>povinnost</i> → <i>povinnost</i> ; <i>polostrov</i> → <i>poloostrov</i> ;
Gemin1	incor. double char. instead of single	<i>sobbota</i> → <i>sobota</i> ; <i>rukoppis</i> → <i>rukopis</i> ;
Quant0	diacritics: missing vowel accent	<i>vzpominám</i> → <i>vzpomínám</i> ; <i>doufam</i> → <i>doufám</i>
Quant1	diacritics: extra vowel accent	<i>ktěrá</i> → <i>kteřá</i> ; <i>hledát</i> → <i>hledat</i>
Caron0	diacritics: missing caron	<i>vecí</i> → <i>věcí</i> ; <i>sobe</i> → <i>sobě</i>
Caron1	diacritics: extra caron	<i>břečel</i> → <i>brečel</i> ; <i>bratřem</i> → <i>bratrem</i>
DiaE	diacritics: <i>ě</i> → <i>é</i> , or <i>é</i> → <i>ě</i>	<i>usměvavé</i> → <i>usměvavé</i> ; <i>poprvé</i> → <i>poprvé</i>
Palat0	missing palatalization of <i>k, g, h, ch</i>	<i>ameriké</i> → <i>Americe</i> ; <i>matké</i> → <i>matce</i>
Palat1	incor. palatalization of <i>k, g, h, ch</i>	<i>koníčcem</i> → <i>koníčkem</i> ; <i>pracovnícem</i> → <i>pracovníkem</i>
EpentE0	<i>e</i> epenthesis: missing <i>e</i>	<i>domček</i> → <i>doměček</i> ; <i>najdnou</i> → <i>najednou</i>
EpentE1	<i>e</i> epenthesis: extra <i>e</i>	<i>rozeběhl</i> → <i>rozběhl</i> ; <i>účty</i> → <i>účty</i>
ProtV1	protethic <i>v</i> : extra <i>v</i>	<i>vosm</i> → <i>osm</i> ; <i>vopravdu</i> → <i>opravdu</i>
GHO	switch error <i>g</i> → <i>h</i>	<i>glavní</i> → <i>hlavní</i> ; <i>mного</i> → <i>mnoho</i>
GH1	switch error <i>h</i> → <i>g</i>	<i>ciharetu</i> → <i>cigaretu</i> ; <i>hrafička</i> → <i>grafická</i>

Table 5.8: Formal errors on T1 – part 1 of 2. All the error tags are prefixed with the string `form`, e. g., `formCap0`. We omit this prefix for space reasons.

Error type	Error description	Example
Meta	character metathesis	<i>dobrodružství</i> → <i>dobrodružství</i> ; <i>provůdce</i> → <i>průvodce</i>
SingCh	other erroneous character substitution	<i>otevřila</i> → <i>otevřela</i> ; <i>vezmíme</i> → <i>vezmeme</i>
MissChar	other missing character	<i>protže</i> → <i>protože</i> ; <i>oňostroj</i> → <i>ohnostroj</i>
RedunChar	other extra character	<i>opratrně</i> → <i>opatrně</i> ; <i>zrdcátko</i> → <i>zrcátko</i>
Pre0	missing prefix	<i>hledu</i> → <i>pohledu</i> ; <i>žaduje</i> → <i>vyžaduje</i> ;
Pre1	superfluous prefix	<i>pojet</i> → <i>jet</i> ; <i>přezačit</i> → <i>začit</i>
Head0	other missing characters word-initial	<i>busovou</i> → <i>autobusovou</i> ;
Head1	extra characters word-initial	
Head	different word beginnings	<i>prověděl</i> → <i>dozvěděl</i> ; <i>chiny</i> → <i>Číny</i>
Tail0	missing characters word-final	<i>t</i> → <i>tam</i> ; <i>ež</i> → <i>žit</i> ;
Tail11	extra characters word-final	
Tail	different word endings	<i>šansu</i> → <i>šanci</i> ; <i>nezajína</i> → <i>nezajímá</i>
Len0	orig. word shorter & unspec. errors	<i>ňákem</i> → <i>nějakém</i> ; <i>diš</i> → <i>když</i>
Len1	orig. word longer & unspec. errors	<i>vidňanami</i> → <i>Vídeňany</i>
Unspec	too many differences	<i>kreěnu</i> → <i>kterěnu</i> ; <i>choubů</i> → <i>hub</i>

Table 5.9: Formal errors on T1 – part 2 of 2. All the error tags are prefixed with the string **form**, omitted here for space reasons.

It is harder to argue that error classification can be done alone without assuming one or more implicit target hypotheses. Lüdeling et al. (2005) argue that it is impossible to tag an error without interpretation, i. e., without the assumption of one or more THs. Error classification is always based on a yardstick, represented at least as implicit alternatives or even a vague concept of what the author means. On the other hand, if a sentence or an expression resists interpretation as a whole and only spelling corrections are possible, there is an implicit target hypothesis involved in the orthographically correct form.

This is where the concept of interlanguage comes into play. If there are any linguistic components and phenomena of interlanguage which can be analyzed and annotated independently of a standard based on the target language of native speakers, then error classification could stand on its own, without an implicit target hypothesis. However, we have not annotated any part of *CzeSL* this way.

The 2T scheme can also be used without error tags to manually annotate texts only with target hypotheses, while retaining the tier-based distinction of errors in spelling/morphemics corrected at T1 and other errors corrected at T2 (see §8.4 for *CzeSL-TH*, a *CzeSL* release annotated this way). Successive corrections are still possible and the same annotation toolchain can be used. The absence of error tags representing explicit error categorization may be a problem for many usage scenar-

ios, but the task of manual annotation is less demanding, the resulting annotation more consistent and still useful for some purposes. Last but not least, the error tags can be added in a second annotation round, while the target hypotheses can be checked and modified if necessary.

The lower cost of manual annotation is the practical reason why existing *CzeSL* texts without manual annotation and also new texts are annotated without explicit error tagging. The linguistic motivation is the somewhat suboptimally representative choice of texts in *CzeSL-man* – a disproportionately large share of native speakers of some languages and an uneven distribution of proficiency levels.

Rather than using the 2T scheme, new texts are annotated in the *TEITOK* environment. For more about the implicit annotation of new texts see §8.8 and §7.7.

5.6 Multi-dimensional error annotation (MD)

In comparison with the 2T scheme, the MD scheme differs mainly in the following aspects:

- It is concerned primarily with morphology, spelling and morphonology.
- All annotation is applied to the source text (T0), relative to a single target hypothesis (T2).
- The annotated unit can be a single character, or a string of characters. The string can span multiple tokens, even in a discontinuous sequence.
- A single error may receive multiple alternative error categories.

5.6.1 Focus on morphology

The MD annotation scheme is focused on errors in spelling, morphonology and morphology, although the tagset allows to capture other types of errors in order to cross-check annotation of the same phenomena done in the 2T scheme.¹⁸ Moreover, a single error can be classified by more than one error category. This is why the annotation scheme can also be interpreted as “multi-domain”. However, the acronym MD is meant to be read as “multi-dimensional” – the dimensions are three views of each annotated phenomenon:

1. Location (prefix/stem/suffix and character range)

¹⁸All existing annotation schemes can be merged in a single corpus (see §9.2.3).

2. Linguistic domain (spelling, morphonology, morphology, syntax, lexicon)
3. Additional cross-domain characteristics (register, follow-up error, problem)

5.6.2 All annotation applied to the source text

There is only one target hypothesis, corresponding to T2. Annotation is provided with respect to T2, however, it is T0, the source text, which is annotated. Error tags are assigned to those parts of the source text which are different from the TH.

For texts annotated in the 2T scheme, the MD scheme adopts the existing TH. The annotators are instructed to modify the TH only in cases when the hypothesis seems to be mistaken.

5.6.3 Extent of the annotated unit

The minimal text unit to be annotated in the 2T scheme is a token (see §5.4.2.3). The only cases where the annotation is concerned with sub-token units is the hand-annotated distinction between *incorBase* vs. *incorInfl*, errors in word boundaries and – in formal tags – an implicit focus on morphs or characters. Even in cases of multiple unrelated instances of errors within the same form the 2T scheme does not allow to identify an incorrect morph or a specific character. This stands out as a somewhat neglected territory in the 2T scheme, especially in contrast with the more detailed annotation of errors in morphosyntax. Another reason for more detail in the identification of the locus of the error is the fact that non-native Czech exhibits all sorts of morphological idiosyncrasies, occurring in ill-formed stems, derivational affixes and inflectional endings. These are some of the reasons why the 2T scheme is complemented by the MD scheme capturing phenomena related to morphs or even individual graphemes.

The following examples illustrate the phenomena. In (16), the learner used a form showing accusative but failed to apply a rule of *-e* epenthesis to the stem *lev* ‘lion.NOM’ (i. e., to its declension paradigm).

- (16) T0: *vidím *leva*
 see.1SG lion.(ACC)
- T2: *vidím lva*
 see.1SG lion.ACC
 ‘I see a lion’

Morphemic error hypothesis seems to be preferable even in (17). In many, but not all, nouns of an otherwise identical declension paradigm, the genitive singular

ending *-a* with the *-u* ending, cf. *lesu/lesa* ‘forest.GEN’ vs. **Petřínu/Petřína* ‘the Petřín hill.GEN’ vs. *Řípu/*Řípa* ‘the Říp hill.GEN’. At the same time, the form *Petřínu* could be interpreted as the proper dative case, even though the learner’s intention to use the dative case is unlikely. The example is glossed according to the latter hypothesis of a morphosyntactic error, but the MD annotation scheme leaves room for both interpretations.

- (17) T0: *pohled z *Petřínu*
 view from Petřín.*DAT
 T2: *pohled z Petřína*
 view from Petřín.GEN
 ‘a view from the Petřín hill’

A morphemic error can be combined with a morphosyntactic error as in (18), where the ill-formed word *levy* can be interpreted as including two errors: (i) the learner used a form showing accusative even though the verb takes a dative object, and also (ii) failed to apply a rule of epenthesis to the stem.

- (18) T0: *rozumím *levy*
 understand.1SG lions.(*ACC)
 T2: *rozumím lvům*
 understand.1SG lions.DAT
 ‘I understand lions’

5.6.4 Alternative error domains

Sometimes a single error can be interpreted in a number of ways despite a single target hypothesis, as in (19).

- (19) T0: *Tam žije 200 *lidi.*
 there lives 200 people.*NOM
 T2: *Tam žije 200 lidí.*
 there lives 200 people.GEN
 ‘Two hundred people live there.’

A syntactician’s explanation of why the form *lidi* ‘people’ is wrong could be that nouns quantified by cardinal numbers higher than 4 are obligatorily assigned the genitive rather than the nominative case.¹⁹ However, the causes of the deviation from the correct form can be several:

¹⁹The wrong and the correct forms differ in the quantity of the final vowel. The acute accent over *i* denotes length.

- Spelling: the student forgot to mark the appropriate diacritic over the character *i* (a common error for non-native learners of Czech).
- Phonology: the student does not register (hear) the phonological difference between a short (*i*) and a long vowel (*í*).
- Morphology: the student considers the ending *-i* as correct for genitive plural of *lidé*.
- Syntax: the student assumes that the correct case in this context is nominative plural (because the noun phrase *200 lidí* is the subject of the sentence).

Although the uncertainty about the cause of the error is usually limited to one or two domains, multiple possible causes are not too rare in learner texts. In the 2T scheme the problem of deciding at which tier or by which error tag a specific error should be hand-annotated is resolved by a general rule to prefer a ‘more sophisticated’ explanation, i. e., by applying the preference for a target hypothesis and an error tag at T2 rather than at T1. In the 2T scheme, the wrong form in (19) is annotated unambiguously at T2 as an error in case assignment (**dep**).²⁰

On the other hand, the MD scheme allows for multiple error domains and the form is assigned error tags in all those four domains listed above.

The sentence in (20) further illustrates the need to distinguish between the linguistic domains, and sometimes combine some of them.

- (20) T0: *Během *dovoleny *šla s *kamaradku na *pláži každý den.*
 during (holiday) went.*PERF with (friend).(*ACC) on beach.*LOC every day
- T2: *Během dovolené **chodila** s kamarádkou na pláž každý den.*
 during holiday went.IMPf with friend.INS on beach.ACC every day

‘On holiday, she used to go to the beach with a friend every day.’

(HRD_LV_206 zh B1)

There are four incorrectly used words in (20): *dovoleny* → *dovolené*, *šla* → *chodila*, *kamaradku* → *kamarádkou* and *pláži* → *pláž*. One of these words, *kamaradku* ‘friend’, has two independent errors (*a* → *á* and *u* → *ou*).

²⁰A similar solution disregarding this uncertainty about the origin of the error is used also in the *MERLIN* project (see Boyd et al. (2014) and Wisniewski et al. (2014)).

The form *dovoleny* ‘holiday’, where only the ending *-y* differs from the appropriate form *dovolené*, is apparently formed from a correct word base and an incorrect ending *-y*, which is a correct genitive singular ending for a different feminine paradigm, or a misspelled or mispronounced variant of the otherwise correct CCz genitive singular form *dovolený*. The error in *dovoleny* → *dovolené* is therefore undoubtedly an error in morphology, phonology or spelling, as the author of the text apparently fails to use the correct ending for the word she uses, but the case seems to be correct.

The form *šla* ‘went.PERF’, used instead of the correct form *chodila* ‘used to go.IMPF’, is a correctly formed Czech word. However, its perfective aspect is inappropriate in the context of the expression *každý den* ‘every day’. It is replaced by the imperfective form and classified as a lexical error.²¹

The omission of the diacritic on the vowel *a* in the word *kamaradku* → *kamarádkou* ‘friend’ can be classified as an error in spelling or phonology (non-native speakers of Czech often do not distinguish between short and long vowels). The second error in the word *kamaradku* ‘friend’, i. e., the inappropriate use of the ending *-u* (which is correct for accusative singular) instead of *-ou* (the ending for instrumental singular, required here after the preposition *s* ‘with’), is either an error in morphology (the author of the text does not know what ending to choose to form the instrumental case), or an error in syntax (the author does not know which case should be used with the preposition *s* ‘with’).

The last incorrect form *pláži* ‘beach’, used instead of *pláž*, is a correct form of locative singular, a form that can be used after the preposition *na* ‘on’ (so that both *na pláži* and *na pláž* can be correct, depending on the context), but it is inappropriate with a verb of movement, such as *jít/chodit* ‘go’, so the error can be interpreted as an error in syntax. However, the ending *-i* is used in other feminine paradigms to form the accusative case, so we cannot exclude the possibility of an error in morphology (the appropriate accusative case is formed incorrectly).

5.6.5 Source text, target hypothesis, annotated strings

- Error tags are assigned to those parts of the source text which are different from the single target hypothesis, corresponding to T2 in the 2T scheme.
- A tag can annotate a part of a word, a whole word, or even multiple words. More than one tag can annotate a single text string. A tag may annotate a string which includes a shorter string annotated by a different tag, i. e., a tag

²¹The two forms are actually forms of two different verbs, because aspect in Czech is a lexical rather than morphological or syntactic category.

can be embedded in another tag. The spans of characters or words annotated by different tags may overlap.

- As a rule, the shortest possible strings are annotated, a sequence of incorrect characters, sometimes just one character. Only lexical errors are annotated on the full word form.

5.6.6 Domains and features

The MD scheme is based on five general categories of errors – *domains* – and a number of subcategories (4–13) – *features* – for each of the domains. For the full list of domains and features with examples see Tables 5.10–5.12 on pages 115–117.²²

- Each error is assigned to at least one domain and to one feature appropriate for the domain.
- Features are unique across the domains. Domains and detailed categories (domain-feature pairs) are thus identifiable by the feature tag.
- Multiple domains assigned to an error are interpreted as alternative explanations of the error.

The domain of orthography covers errors caused by ignoring the conventions of Czech writing, such as capitalization (*praha* → *Praha* ‘Prague’), conventions of transcription of some combinations of phonemes, e. g., *ě* representing the phonemes *j* and *e* in *vjec* → *věc* ‘thing’, the use of diacritics (*deti* → *děti* ‘children’) etc. Many of such errors are fairly common even among native speakers.

The domain of morphonology includes errors in phonology, e. g., the transcription of voiced and voiceless consonants (*sůstala* → *zůstala* ‘stayed’), or the distinction between the consonants *r* and *l* (sometimes ignored by native speakers of Chinese or Japanese, e. g., *na kluku* ‘on the boy’ vs. *na krku* ‘on the neck’, and incorrect forms of morphemes unrelated to inflection, e. g., *učiteka* → *učitelka* ‘teacher’).

As errors in morphology we classify only errors related to nominal declension and verbal conjugation, including both non-words (*na Erasmuse* → *Erasmu* ‘on the Erasmus’; *studovám* → *studuju* ‘I study’) and existing forms of the given word, inappropriate in the given context (in this case, the error can be either morphological or syntactic).

²²More details can be found in the (Czech) annotation manual (Škodová et al. 2019).

The domain of syntax covers errors caused by the incorrect use of word forms and function words (including prepositions) in a given context. Typically, this is where errors in valency, agreement, quantification and word order belong.

Errors in the lexical domain concern cases when the original word is replaced in the correction by a different word with a different meaning and it is not the result of a random morphological error. If necessary, two or more error domains can be used for the classification of any error.

The alternative explanation of a single error by parallel annotation in multiple domains results in some regularities (see Table 5.3 on page 118) or frequent co-occurrence (see Figure 5.13 on page 118) of some error tags. In addition to some linguistic interest, relations of implication and predictable coincidence of some domains and features are used to alleviate the task of manual annotation. In addition to a partial segmentation into morphs and the assignment of automatically identifiable features, usually corresponding to the formal error tags with counterparts in the *ORT* and *MPHON* domains, preceding the manual annotation, some annotation is added in a post-processing step, based on the such relations.

For more details about the annotation process concerning the MD scheme see 7.6 on page 140.

²³except for voiced↔unvoiced

Feature	Gloss	Examples
ORT	Spelling only, not pronounced, except for some GEM errors	
IY	$i \leftrightarrow y$	<i>analízovat</i> → <i>analyzovat</i> , <i>odpovídají</i> → <i>odpovídají</i> , <i>myšlenka</i> → <i>myslenka</i> , <i>babyčka</i> → <i>babyčka</i> , <i>viděl vili</i> → <i>vily</i>
ME	$mě \leftrightarrow mně$ etc.	<i>dítie</i> → <i>dítě</i> , <i>njekdo</i> → <i>někdo</i> , <i>tjišeji</i> → <i>tišeji</i> , <i>mněsíc</i> → <i>měsíc</i> , <i>jědí</i> → <i>jedí</i> , <i>pjet</i> → <i>pět</i> <i>rohlíků</i> , <i>obět</i> → <i>objet</i> <i>náměstí</i> , <i>konie</i> → <i>koně</i> , <i>dítě</i> → <i>dítě</i> , <i>díle</i> → <i>dítě</i> , <i>pro mně</i> → <i>mě</i> , <i>telo</i> → <i>tělo</i> <i>půjdu</i> → <i>půjdu</i> , <i>úzký</i> → <i>úzký</i> , <i>domů</i> → <i>domů</i>
AT	$ú \leftrightarrow ů$	
DIA	diacritics (other)	<i>ob ějet</i> → <i>objet</i> , <i>řikat</i> → <i>řikat</i> , <i>unava</i> → <i>únava</i> , <i>músel</i> → <i>musel</i>
GEM	gemination	<i>denník</i> → <i>deník</i> , <i>pana</i> → <i>panna</i> , <i>rozlobit</i> → <i>rozzlobit</i> , <i>odálit</i> → <i>oddálit</i>
SUBST	substitution (other)	<i>yako</i> → <i>jako</i> , <i>dal to Mariji</i> → <i>Marii</i>
CAP	capitalization	<i>praha</i> → <i>Praha</i> , <i>Maminka</i> → <i>maminka</i>
PUN	punctuation	<i>máma</i> , <i>a táta jsou doma</i> → <i>máma</i> , <i>a táta jsou doma</i> ; <i>přišla</i> <i>aby se rozloučila</i> . → <i>přišla</i> , <i>aby se rozloučila</i>
SEG	word boundary	<i>ne jsem</i> → <i>nejsem</i> , <i>byses</i> → <i>by ses</i> , <i>smaminkou</i> → <i>s maminkou</i>
MPHON	Morphology – errors altering non-native pronunciation	
VOC	vocalization	<i>z</i> → <i>ze</i> <i>školy</i> , <i>v</i> → <i>ve</i> <i>škole</i> , <i>se</i> → <i>s</i> <i>Marií</i>
ASIM	voiced↔unvoiced assimilating	<i>gdyž</i> → <i>když</i> , <i>noz</i> → <i>nos</i> ; <i>ktyš</i> → <i>když</i> , <i>hrat</i> → <i>hrad</i> , <i>bes</i> → <i>bez tebe</i> , <i>f</i> → <i>v</i> <i>kruhu</i> , <i>naschledanou</i> → <i>na shledanou</i>
NASIM	voiced↔unvoiced non-assimilating	<i>grad</i> → <i>hrad</i> , <i>uglí</i> → <i>uhlí</i> , <i>výhodní</i> → <i>východní</i> , <i>roglík</i> → <i>rohlík</i> , <i>bod židlí</i> → <i>pod židli</i> ; <i>uchlí</i> → <i>uhlí</i> , <i>chlidám</i> → <i>hlídám</i> , <i>rochlík</i> → <i>rohlík</i> , <i>sjištovat</i> → <i>zjištovat</i>
SIB	sibilants, affricates ²³	<i>vajes</i> → <i>vajec</i> , <i>noc</i> → <i>nos</i> , <i>mucím</i> → <i>musím</i>
PAL	palatalization <i>g/k/ch</i> , <i>l/r</i>	<i>ledničke</i> → <i>ledniče</i> , <i>článkech</i> → <i>článcích</i> , <i>ruke</i> → <i>ruce</i> , <i>páre</i> → <i>páře</i> , <i>Prahe</i> → <i>Praze</i> , <i>soši</i> → <i>sochy</i>
SOFT	softening $\acute{e}, d, t, n, r, s, c, z$	<i>telo</i> → <i>tělo</i> , <i>tělefon</i> → <i>telefon</i> , <i>něbe</i> → <i>nebe</i> , <i>veda</i> → <i>věda</i>
QUANT	vowel length	<i>Práha</i> → <i>Praha</i> , <i>učitelka</i> → <i>učitelka</i> , <i>počítač</i> → <i>počítač</i>
MET	metathesis <i>r/l/m/n</i>	<i>pernamentka</i> → <i>permanentka</i> , <i>lefrektor</i> → <i>reflektor</i> , <i>verlyba</i> → <i>velryba</i> , <i>žlička</i> → <i>lžička</i> , <i>lorák</i> → <i>rolák</i>
EPENT	epenthesis (including both <i>i</i> and <i>e</i>)	<i>volb</i> → <i>voleb</i> , <i>pesa</i> → <i>psa</i> , <i>ptáček</i> → <i>ptáčka</i> , <i>lžička</i> → <i>lžička</i>
PROT	prosthetic <i>v-</i>	<i>vlakno</i> → <i>okno</i> , <i>vobjednat</i> → <i>objednat</i> , <i>von</i> → <i>on</i> , <i>vošklivej</i> → <i>ošklivej</i>
CNTR	contraction	<i>děcký</i> → <i>dětský</i> , <i>bohactví</i> → <i>bohatství</i> , <i>czeský</i> → <i>český</i> , <i>morže</i> → <i>moře</i> , <i>bicze</i> → <i>biče</i> , <i>šok</i> → <i>šok</i>
ALT	other alternations	<i>ve vůze</i> → <i>voze</i> , <i>koup</i> → <i>kup to</i>
CHAR	additional or missing sounds	<i>večře</i> → <i>večeře</i> , <i>nesem</i> → <i>nejsem</i> , <i>cera</i> → <i>dcera</i> , <i>sedum</i> → <i>sedm</i>

Table 5.10: An overview of domains and features in the MD scheme, part 1 of 3

Feature	Gloss	Examples
MORPH Morphology – errors due a wrong choice of affixes and stems		
NAFF	affix incompatible with the stem	<i>spám</i> → <i>spím</i> , <i>neplavujeme</i> → <i>neplaveme</i> , <i>přečet</i> → <i>přečetl</i> , <i>Číné</i> → <i>Číny</i> , <i>Úvalách</i> → <i>Úvalech</i> , <i>rodičema</i> → <i>rodiči</i> , <i>dědečko</i> → <i>dědečka</i> , <i>Erasmusu</i> → <i>Erasmu</i>
FLEX	inappropriately used affix in a paradigm	<i>uvidím dědečkovi</i> → <i>dědečka</i>
VBX	compound verb forms	<i>zítra jsem</i> → <i>budu spát</i> , <i>jsem spát</i> → <i>spím</i> , <i>budu napsat</i> → <i>napišu</i> , <i>musíš přijdeš</i> → <i>přijít</i> , <i>učít</i> → <i>učil ses</i>
RFL	reflexives	<i>raduje si</i> → <i>se</i> , <i>má ráda její</i> → <i>své</i> děti, <i>směju</i> → <i>směju se</i>
PREP	preposition	<i>bydlím na</i> → <i>v</i> Praze, <i>vystup v</i> → <i>na</i> <i>konečné</i> , <i>půjdu v</i> les → <i>půjdu do</i> lesa
SYN Syntax		
AGR	agreement	<i>můj tatínek je už stará</i> → <i>starý</i>
DEP	dependents	<i>pozdravuj Honza</i> → <i>Honzu</i> , <i>bojím se jí zavolám</i> → <i>zavolat</i>
SUBJ	subject – missing or redundant pronoun	<i>dopoledne já čtu a já odpočívám</i> → <i>dopoledne čtu a odpočívám</i> ; <i>já ne, ale to uděláš</i> → <i>já ne, ale ty to uděláš</i>
COMPL	complement – missing (pronominal) object	<i>potřebuju tužku, maminka koupí</i> → <i>maminka ji koupí</i> ; <i>přivedla muže, Jana neznala</i> → <i>kterého Jana neznala</i>
COP	missing <i>be</i> , esp. copula	<i>země moc velká</i> → <i>země je moc velká</i> ; <i>teď v Číně</i> → <i>teď je v Číně</i>
CONST	other constituent	<i>přeju mu cestu</i> → <i>přeju mu šťastnou cestu</i>
CONJ	connecting expression, including relative pronoun	<i>mám ráda, že</i> → <i>když</i> <i>prší</i> ; <i>mám hlad, protože</i> → <i>proto se najím</i> ; <i>mám rád Prahu, proto</i> → <i>protože je přátelská</i> ; <i>chtěl, kdyby</i> → <i>abych</i> <i>přišel</i> ; <i>Petr ale</i> → <i>a</i> <i>Lucie se mají rádi</i> ; <i>doporučuju každému, který</i> → <i>kdo</i> <i>má zájem</i>
WO	word order	<i>mají hezký velmi dům</i> → <i>mají velmi hezký dům</i>

Table 5.11: An overview of domains and features in the MD scheme, part 2 of 3

Feature	Gloss	Examples
LEX	Lexicon	
CHOICE	wrong lexeme	<i>jeli pěšky</i> → <i>šli</i> ; <i>nudím se po domově</i> → <i>stýská se mi po domově</i>
ASP	aspect	celý den <i>chytili</i> → <i>chytali</i> ryby; denně <i>vstanu</i> → <i>vstávám</i> brzo
MOD	modality – verb, adverb, particle	v pondělí <i>může</i> → <i>musí jít do práce</i> ; <i>hodně</i> <i>myslím</i> , že byl <i>hladový</i> → byl <i>určitě</i> <i>hladový</i>
NEG	negation	<i>půjdu neráno</i> → <i>nepůjdu ráno</i> ; <i>on ne</i> → <i>není velký</i> ; <i>půjdu ne do školy</i> → <i>nepůjdu do školy</i> ; <i>mám</i> → <i>nemám žádný čas</i> ; <i>máma ani táta kouří</i> → <i>nekouří</i>
COIN	coinage – innovative word formation	<i>slíchtování názory</i> (?); <i>štopínky špičurkatý</i> (?); <i>je to smíchovní</i> → <i>legrační</i>
FGN	foreign or macaronic	<i>jdu do shopu</i> ; <i>byla v hangu</i> ; <i>to byl shock</i> ; <i>hledám kleenexy</i>
USE	suboptimal choice of (variant) forms, lexemes, collocations, independent categories	<i>říkám moje</i> → <i>svoje názory</i> ; <i>dělám studium</i> → <i>studuji</i> ; <i>ráno přišla rýma</i> → <i>ráno se mi spustila rýma</i> ; <i>moucha chodila</i> → <i>lezla po stole</i> ; <i>vidí jeho</i> → <i>ho v zrcadle</i> ; <i>dívá se na ho</i> → <i>něj</i> ; <i>na jaru</i> → <i>jaře všechno kvete</i> ; <i>čte už dlouze</i> → <i>dlouho</i>
POS	part of speech	<i>je to hezky</i> → <i>hezký muž</i> ; <i>učím se český</i> → <i>česky/češtinu</i> ; <i>moc rád pomoc</i> → <i>pomůže</i> ; <i>jsem český</i> → <i>Čech</i> ; <i>je to dobře</i> → <i>dobrá lekce</i> ; <i>máma jméno Dana</i> → <i>máma se jmenuje Dana</i>
PHR	construction	<i>Petr má rád lyžovat</i> → <i>Petr rád lyžuje</i> ; <i>mám 17 let</i> → <i>je mi 17 let</i> ; <i>já líbím Prahu</i> → <i>líbí se mi Praha</i> ; <i>večer dostanu bolest hlavy</i> → <i>večer mne začne bolet hlava</i>
XDOM	Cross-domain	
REG	register	<i>koláč je dobrej</i> → <i>dobrý</i> ; <i>přijdu s rodičema</i> → <i>rodiči</i> ; <i>to je ale maglajz</i> → <i>zmatek</i>
SEC	secondary (follow-up, subsequent) error	<i>jde na menzu</i> → <i>jde do menzy</i> ; <i>Saná má úžasný klimat</i> → <i>Saná má úžasně klíma</i>
PROBL	problem	

Table 5.12: An overview of domains and features in the MD scheme, part 3 of 3

- **MPHON:QUANT** \implies **ORT:DIA**
An issue in the quantity of a vowel is also an issue in diacritics, i. e., missing or redundant acute accent or “ring”.
 - **MORPH:NAFF** \implies no annotation in the **SYN** domain
An incompatible affix results in a non-word, which excludes a syntactic error.
 - **MORPH:FLEX** \iff **SYN:AGR** or **SYN:DEP**
An inappropriate ending in a form, still correct within a paradigm, is always due to an issue in syntax, either in agreement or in case assignment or other requirements of a syntactic head.
 - if **MORPH:NAFF** or ... (i. e., incompatible affix)
if (**SYN:DEP** or **SYN:AGR**) and ... (i. e., syntactic error)
if no other **MPHON** or ... (e. g., **SOFT/QUANT/NASIM**)
if no **ORT:IY/MNE/U/GEM/CAP** (i. e., if it’s not spelling only)
- \implies
- MPHON:ALT** or ... (when replacing characters)
 - MPHON:CHAR** (when deleting or adding characters)

Figure 5.3: Relations of implication and equivalence between features across error domains

ORT	MPHON	MORPH	SYN	Example
IY	SOFT	FLEX	AGR	<i>všichni děti</i> \rightarrow <i>všechny děti</i>
IY		FLEX	AGR	<i>byli</i> \rightarrow <i>byly</i>
DIA	QUANT	FLEX	DEP	<i>děti</i> \rightarrow <i>děti, přátele</i> \rightarrow <i>přátelé</i>
	ALT	FLEX	AGR	<i>druhém</i> \rightarrow <i>ruhým, jeden</i> \rightarrow <i>jedno</i>
	ALT	FLEX	DEP	<i>tradicí</i> \rightarrow <i>tradice</i>
	CHAR	FLEX	AGR	<i>byl</i> \rightarrow <i>bylo</i>
	CHAR	FLEX	DEP	<i>noh</i> \rightarrow <i>nohy, noc</i> \rightarrow <i>nocí</i>

Table 5.13: Co-occurrence of features across error domains

Chapter 6

Linguistic annotation

Standard corpora documenting contemporary written native language are commonly annotated by POS tags, morphological categories, lemmas, sometimes syntactic structure and functions, or even by information about named entities or word senses. For many languages, tools and training data are available to perform these tasks with an error rate sufficient for many purposes. The result is a corpus more useful in a number of ways. A linguistically annotated corpus can be searched more efficiently. It may even be impossible to make some queries without such annotation. The same applies to statistical analyses. Also, linguistically annotated corpora are vital in the development of NLP tools, especially those based on machine learning, which require extensive training and testing.

Learner corpora are no exception: together with error annotation, linguistic annotation helps to make them more useful. In fact, the error tagset used in the 2T scheme assumes that the texts are annotated at least by POS tags (see §5.4.3.3). However, linguistic annotation of learner corpora is not a straightforward task. This is due to several reasons:

1. Available tools are trained on standard language, mainly because it is difficult to obtain sufficiently large training data, comparable with the texts to be annotated in terms of text types and specifics of the learner language. Therefore, annotating learner texts by tools intended for standard language means that the reliability of automatic annotation could be lower than reported for native texts. The drop in success rate depends mainly on how far the texts diverge from the standard language. Even within a single learner corpus, texts authored by learners at different levels of proficiency and L1s can be annotated with various success.

2. In addition to the higher error rate, adopting the standard language approach to learner language arouses a conceptual concern: categories and structures suited to the standard language might not suit learner language.
3. To avoid such problems, we can annotate the target hypothesis instead. This assumes that the source text is reconstructed completely, to a grammatically correct version, including the correction of follow-up errors. However, some properties of the learner language may be lost in the annotation. As a possible solution, both the source text and the target hypothesis can be linguistically annotated.

This chapter has two parts:

1. The first part (§6.1) focuses at automatic linguistic annotation performed with tools for Standard Czech. This is straightforward for target hypothesis. Exploiting the fact that the words in the 2T scheme are interlinked, the result is projected to the partial target hypothesis (T1). Applying existing tools on the source text is theoretically less sound, but for practical purposes, the results are still useful.
2. The second part (§6.2) describes manual syntactic annotation of a portion of *CzeSL* using the Universal Dependencies annotation scheme.¹ We argue that the more abstract syntactic categories are a more intuitive and less arbitrary alternative to morphological annotation of the source learner text. The annotated corpus is too small to train machine learning tools on it in the usual way, but it can be used for benchmarking such tools.

6.1 Annotation with tools for Standard Czech

In order to make it easier for users to work with *CzeSL*, i. e., to enable a comfortable search for words according to base forms and grammatical categories, and to produce statistics based on linguistic categories, the words in the corpus were lemmatized and annotated with morphosyntactic tags.

6.1.1 Annotation of target hypothesis

Because the target hypothesis at T2 is a native-like Czech sentence, we could apply a standard lemmatizer and tagger to assign all T2 words non-ambiguous annotation. We use the “Prague” positional tagset (Hajič 2004) in the version modified for the

¹<https://universaldependencies.org>

Czech National Corpus. Each of the 16 positions corresponds to a morphosyntactic category, e.g., the first position stands for POS, the fifth position for case.² Following lexical look-up, the disambiguation step proceeds in two stages: first a rule-based system removes most of the ambiguity, then a stochastic tagger resolves the remaining cases. For more details see Hnátková, Petkevič, and Skoumalová (2011).

Moreover, the target hypothesis of *CzeSL-man v1 searchable* has a syntactic annotation according to the PDT standards, parsed automatically with *TurboParser* (Martins, Almeida, and Smith 2013).

6.1.2 Annotation of T1

The words in the 2T scheme are interlinked. We use this information to project lemmas and tags from T2 to T1, the intermediary target hypothesis, in the following way:

1. If the T1 word is identical to its T2 counterpart, it gets its lemma and tag.
2. Otherwise:
 - a) If the T2 lemma is one of the possible T1 lemmas, we use that T2 lemma and the set of all tags associated with it which are consistent with the T1 form. For example, the homonym *jí* is either a form of the verb *jíst* ‘eat’ or dative singular of the personal pronoun *ona* ‘she’. Let us assume the ambiguous form on T1 was corrected as *jedí* ‘eat.3PL’ on T2. Because *jedí* is non-ambiguously a form of the verb *jíst* ‘eat’, *jí* on T1 is considered only as the form of this verb and it is assigned tags for the 3rd person plural and Common Czech 3rd person singular.
 - b) Otherwise: T1 gets all possible lemmas and all possible tags.

6.1.3 Annotation of source texts

In several releases of *CzeSL* (*CzeSL-SGT*, *CzeSL-man v1*, *CzeSL-man v2*, *CzeSL in TEITOK*), automatic lemmatization and morphosyntactic tagging is available for T0, the source version of the text. For this task, we used *MorphoDiTa* (Straková, Straka, and Hajič 2014), trained on standard native Czech data of the *Prague Dependency Treebank* (PDT, Hajič et al. 2018), rather than the hybrid tagger applied to native Czech texts of the Czech National Corpus and used also for several TH

²See <https://wiki.korpus.cz/doku.php/en:pojmy:tag> for a description of the tagset.

versions of *CzeSL* corpora. No adaptation of the tagset or training data to the learner text was performed – this annotation is meant as a simple aid for searching the corpus and generating rough statistical data rather than as an accurate analysis of the interlanguage.

6.2 Annotation of interlanguage in the Universal Dependencies framework

As mentioned in §6.1.3, the linguistic annotation of the source text was performed by tools designed for native Czech. To fill in the gaps and to put the linguistic annotation of the source text on a more solid ground, we have focused on syntactic annotation of the non-native text within the framework of Universal Dependencies. Our ideal goal is to annotate according to the non-native grammar in the mind of the author, not according to the standard grammar. However, this brings many challenges. First, we do not have enough data to get reliable insights into the grammar of each author. Second, many phenomena are far more complicated than they are in native languages.

Universal Dependencies (UD) is a unified approach to grammatical annotation that is consistent across languages and that currently dominates the annotation projects all over the world.³ It is an established framework used for more than 150 treebanks in 92 languages.⁴ The common guidelines make the data easily accessible to a large audience of researchers and comparable across languages. Also, following the UD schema and format makes it easier to train and test NLP tools on the basis of our annotation.

We follow the basic annotation principle of the SALLE project (Dickinson and Ragheb 2013), and attempt to annotate literally. Our annotation principles include:

1. When form and function clash, form is considered less important.
2. When lacking information, we make conservative assumptions.
3. We focus on syntactic structure and the most important grammatical functions, annotating unclear functions with an underspecified label.

Consider the sample essay in Figure 6.1. The text is perfectly understandable, yet there are errors in nearly every sentence and in about every other word. Some of

³<https://universaldependencies.org>

⁴The latest version 2.6 treebanks are available at <https://hdl.handle.net/11234/1-3242>.

Source	Target hypothesis	Translation
<i>Jmenujese Adam.</i>	<i>Jmenuji se Adam.</i>	My name is Adam.
<i>Ja jsem Mongolska.</i>	<i>Jsem z Mongolska.</i>	I am from Mongolia.
<i>Mongolska ma 21 kraji.</i>	<i>Mongolsko má 21 krajů.</i>	Mongolia has 21 regions.
<i>Moje rodina je hezka ještě velka.</i>	<i>Moje vlast je hezká a velká.</i>	My country is nice and large.
<i>Mongolska je 3000 million lidí.</i>	<i>Mongolsko má 3 miliony lidí.</i>	Mongolia has 3 million people.
<i>Ma tradiční píseňka, taneční.</i>	<i>Má tradiční písničky, taneční.</i>	It has traditional dancing songs.
<i>Mongolska tradiční píseňka je hezka.</i>	<i>Tradiční mongolská písnička je hezká.</i>	Traditional Mongolian song is nice.
<i>Ještě ma "Morin khuur".</i>	<i>Ještě máme "Morin chuur".</i>	We also have "Morin Khuur".
<i>Morin Khuur to je muzika.</i>	<i>Morin Chuur je muzika.</i>	Morin Khuur is music.
<i>Ten hezka tradiční pohádka, píseň.</i>	<i>Je to hezká tradiční pohádka, píseň.</i>	It is a nice traditional folk tale, a song.
<i>Mongolska má mnoho tradiční svátek.</i>	<i>Mongolsko má mnoho tradičních svátků.</i>	Mongolia has many traditional festivals.
<i>Třeba Naadam, Tsagaarsur.</i>	<i>Třeba Nádam, Cagán sar.</i>	Such as Naadam, Tsagaan Sar.
<i>Ještě mnoho Velbloud, Kůn, Kravá, Koza, Ovce.</i>	<i>Ještě mnoho velbloudů, koní, krav, koz, ovcí.</i>	Also many camels, horses, cows, goats, sheep.
<i>Mongolsky lidi dobrý.</i>	<i>Mongolský lidi jsou dobrý. (Common Czech) Mongolští lidé jsou dobří. (Official Czech)</i>	Mongolian people are good.
<i>Mongolsko ma mnoho horý a nemam ocean.</i>	<i>Mongolsko má mnoho hor a nemá oceán.</i>	Mongolia has many mountains and does not have an ocean.
<i>Mongolska hlavní naměsto. Ulaanbaatar.</i>	<i>Mongolské hlavní město je Ulánbátar.</i>	Mongolian capital is Ulaanbaatar.
<i>ADAM, 18 Let</i>	<i>ADAM, 18 let</i>	ADAM, 18 years
<i>Bydlím v Čechagh už 6 měsíc.</i>	<i>Bydlím v Čechách už 6 měsíců.</i>	I have been living in Czechia for 6 months already.

Figure 6.1: An essay written by a male student on the topic "My family" (BRY_B9_001 mn A2)

these deviations from native language make annotation with traditional grammatical categories quite complicated. Consider the second sentence: *Ja jsem Mongolska* meaning ‘I am Mongolian’ or ‘I am from Mongolia’. The word *Mongolska* can be interpreted at least in the following three ways:

1. As an adjective (*mongolská* or *mongolský*) and thus syntactically a predicative nominal
2. As a name of an inhabitant (*Mongol*), a noun, syntactically a predicative nominal
3. As a place (*z Mongolska* ‘from Mongolia’), a noun (actually a prepositional phrase lacking the preposition), syntactically an adverbial or (in some frameworks) an adjunct

It is not clear whether the language of the speaker actually distinguishes all of these categories. In this case, the UD framework makes the situation simple: all three interpretation lead to a structure with *Mongolska* being a non-verbal predicate and *jsem* ‘am’ being a copula.

6.2.1 Tokenization

There is an established tokenization used by Czech UD corpora that builds on the general UD tokenization rules. However, we used the original *CzeSL* tokenization to make the UD structures compatible with their error annotation. The differences affect mostly hyphenated words, certain numerical expressions and alternatives offered by the author due to their uncertainty (e. g., as in *b(y/i)l*, where the teacher-please-choose spelling of *byl* ‘was’ or *bil* ‘beat’ is treated as one token).

6.2.2 Part-of-speech and morphology

As mentioned in [Chapter 1](#), Czech, as most other Slavic languages, is richly inflected. Therefore, corpora of SCz are usually annotated with detailed morphological tags (for example, the “Prague” positional tagset for Czech has more than 4000 different tags). We have decided not to perform such annotation in *CzeSL-UD*. There are several reasons for this decision, mainly:

- Many endings are homonymous; therefore it is not obvious which form was used if we wanted to annotated according to the form. For example, the ending *-a* has more than 10 different morphological functions depending on the paradigm (cf. [Table B.1](#) on page 235). Therefore if somebody used the

form with the ending *-a* in a place requiring the accusative case with the different ending, we are not sure whether they made a mistake in the case, as in *Matemateka Angličtina* (21) or used the incorrect paradigm, as in *koťatka* (22).

- These complications do not always correlate with intelligibility. Some texts are easy to understand, yet they use wrong or non-existing suffixes, mix morphological paradigms etc. Consider the example in Figure 6.2. It is easy to understand, yet for most forms, most of the case endings are incorrect. It is not even clear if the author’s interlanguage includes the category of case.
- The corpus can still be searched for pedagogical reasons using the information derived by the approaches described in the previous section (see §6.1.1).

- (21) T0: *Chci *si naučit dobře *Česky a *Matemateka *Angličtina.*
 want.1SG REFL learn well Czech and (math) (English).
- T2: *Chci se naučit dobře česky a matematiku a angličtinu.*
 want.1SG REFL learn well Czech and math and English.
 ‘I want to learn Czech and math and English well.’ (HRD_CH_054 zh A2+)
- (22) T0: *Když mi *byli 4 roky, dostala jsem k *narozeninam *toho*
 when me (were) 4 years got AUX.1SG for (birthday) (that)
**slavného *koťatka.*
 (famous) (kitten)
- T2: *Když mi byly 4 roky, dostala jsem k narozeninám to slavné*
 when me were 4 years got AUX.1SG for birthday that famous
koťátko.
 kitten
 ‘When I was 4 years old, I got that famous kitten for my birthday.’
 (HRD_1G_309 ru B2)

Instead, we have limited ourselves to the Universal POS Tagset (UPOS; Petrov, Das, and McDonald 2011).⁵ When form and function clash, form is considered less important. For example, if a word functions as an adjective, we annotate it as an adjective even if it has a verbal ending.

An interesting example is provided by Díaz-Negrillo et al. (2010): the word *during* is used as a preposition in native English, but it is used as a subordinate conjunction in (23). We would annotate it as a subordinate conjunction.

⁵See <https://universaldependencies.org/u/pos/index.html>.

Source	Target hypothesis	Translation
<i>Moje rodina má 6 lide.</i>	<i>Moje rodina má 6 lidí.</i>	My family has 6 people.
<i>Oni všichni do škola.</i>	<i>Oni všichni chodí do školy.</i>	They all go to school.
<i>Máme velký barak, vni jsou 5 pokoj 3 zachod.</i>	<i>Máme velký barák, v něm je 5 pokojů a 3 záchody.</i>	We have a big house, there are 5 rooms and 3 toilets in there.
<i>Mam rad tělocvik a matika.</i>	<i>Mám rád tělocvik a matiku.</i>	I like gym and math.
<i>Chci si naučit dobře Česky a Matemateka Angličtina</i>	<i>Chci se naučit dobře česky a matematiku a angličtinu</i>	I want to learn Czech, math and English well

Figure 6.2: The sentences from the essay “My introduction” written by a 15 years old female student, a native speaker of Chinese, after learning Czech for 1–2 years and over 2 years of stay in Czechia. (HRD_CH_054 zh A2+)

(23) *RED helped him during he was in the prison.* (NOCE GR-1-A-EN-025-F)

One of the common deviating characteristics of learner Czech is the neutralization between adjectives and adverbs. In (24), the adjective *rychlé* ‘quick’ is used instead of the correct adverb *rychle* ‘quickly’.

(24) T0: *Kvalita života by se zlepšila moc *rychlé.*
 quality life.GEN would REFL improve too quick

T2: *Kvalita života by se zlepšila moc rychle.*
 quality life.GEN would REFL improve too quickly

‘Life quality would improve too quickly.’ (AA_JW_002 de A2)

This is similar to German or colloquial English. Unfortunately, UPOS force us to choose between adjectives and adverbs even for speakers who clearly use the same word for both. We annotate such words as adjectives with an additional note.

6.2.3 Lemmata

Ideally, we would use lemmata from the author’s interlanguage. For example, in (25), we would use the lemma *Praga* (correctly *Praha*). The situation is clear, because the word is in the lemma form already (nominative singular).

(25) T0: **Praga je hezké město.* → lemma: *Praga*
 (Prague) is nice city

T2: *Praha je hezké město.* → lemma: *Praha*
 Prague is nice city
 ‘Prague is a nice city.’

Often knowing the native language of the author helps. For example, in (26), the lemma of *krasivaja* is *krasivyyj*, based on Russian.

(26) T0: **Praga je *krasivaja.* → lemma: *krasivyyj*
 (Prague) is beautiful

T2: *Praha je krásná.* → lemma: *krásnýj*
 Prague is beautiful
 ‘Prague is beautiful.’

Sometimes we can see that the author declines a word using a paradigm of another word. For example, for the non-word form *večeřem* in (27) we hypothesize the masculine lemma *večeř*. The form is intended most likely as the instrumental case of *večeře* ‘dinner’. The correct instrumental form of the feminine *večeře.NOM* is *večeří*. The form *večeřem* is built in analogy with the word *oběd* ‘lunch.NOM’ – *obědem* ‘lunch.INS’ (28).

(27) T0: **Začínáme *večeřem.* → lemma: *večeř*
 (start).1PL dinner.(INS)

T2: *Začínáme večeří.* → lemma: *večeře*
 start.1PL dinner.INS

‘We start with dinner.’

(AA_JN_001 de B1)

(28) *Začínáme obědem.* → lemma: *oběd*
 start.1PL lunch.INS

‘We start with lunch.’

However, in many cases, the situation is much more complicated and it is not clear whether a certain deviation is due to a spelling error, incorrect case, wrong paradigm (Czech has at least 14 basic noun paradigms) or simply a random error. Sometimes, we can see particular patterns in the whole document, e. g., the author does not distinguish between adjectives and adverbs, uses only certain morphological cases or certain spelling convention (Russian speakers sometimes use ‘g’ instead of Czech ‘h’), etc. These patterns can help us to deduce lemmata in concrete cases. Unfortunately, sometimes we simply do not have enough data to reliably deduce the correct lemma.

In that case, we try to be as conservative as possible and assume as little as possible: we use the form of the word as its lemma and mark it as unclear in the note field.

The alternative is to use the correct lemma (*Praha* in (25) and *večeře* in (27)). Obviously, this would make the situation clearer and the annotation more reliable. However, the benefit would be minimal: error annotation already provides us with the correct forms so we can easily derive their lemmata using available approaches for standard native language.

6.2.4 Syntactic Structure

In annotating syntactic structure, we again follow the rule of annotating the structure of interlanguage. For example, if the learner uses the phrase (29), the word *místnost* ‘room’ is annotated as a direct object (OBJ), even though a native speaker would use an adverbial (OBL) *do místnosti* ‘into room’ as in T2.

- (29) T0: *vtoupit *místnost.OBJ*
 enter room
 T2: *vtoupit do místnosti.OBL*
 enter into room
 ‘enter a/the room’

Examples (30) and (31) illustrate the difficulties we encountered during the annotation. Each example is followed by the corresponding sentence in standard native Czech.

Missing *že* ‘that’

- (30) T0: *Myslím, velmi málokdo dělá, co chce.*
 think.1SG very few-one does what wants
 T2: *Myslím, že velmi málokdo dělá, co chce.*
 think.1SG that very few-one does what wants
 ‘I think hardly anybody is doing what they want.’ (AA_IK_001 hu B1)

- *Annotation with interpretation.* In the corresponding grammatically correct Czech sentence in T2, the connector *že* ‘that’ follows the verb *myslím* ‘I think’. This makes *velmi málokdo dělá* ‘hardly anybody is doing’ a subordinate complement (object) clause, and thus the verb *dělá* ‘wants’ would be annotated as *ccomp*.

- *Annotation without interpretation.* Without interpretation, we consider the clause *velmi málokdo dělá, co chce* to be coordinated with the previous *Myslím*, connected to it with *conj*.

There is another possibility: the author uses a structure parallel to English *I think very few people know ...* without *that*. Then the form *dělá* ‘wants’ would be annotated as *ccomp* as well.

Using *abych* instead of *že*

(31) T0: *Rozhodla jsem se, *abych se naučila nějaký*
 decided.F.SG AUX.1SG REFL CONJ+AUX.1SG REFL learned.F.SG some
zajímavý jazyk, který je blízko nás ...
 interesting language which is close we.DAT ...

T2: *Rozhodla jsem se, že se naučím nějaký zajímavý*
 decided.F.SG AUX.1SG REFL COMP REFL learn.1SG some interesting
jazyk, který je blízko nás ...
 language which is close we.DAT ...

‘I decided to learn some interesting language that is close to us ...’

(AA_IK_001 hu B1)

- *Annotation with interpretation.* The verb *rozhodla jsem se* ‘I decided’ in the first clause requires a complement clause connected via the complementizer *že* ‘that’: *že se naučím ... jazyk* ‘that I learn ... language’, instead of an adverbial clause *abych se naučila ... jazyk* ‘in order to learn ... language’. Therefore the predicate *naučím* ‘learn’ of the subordinate clause would be annotated as *ccomp*.
- *Annotation without interpretation.* The subordinate clause is considered a complement clause as well but marked with the conjunction *aby* ‘so-that’. This is parallel to (32). In this case, UD helps by not forcing us to make spurious distinctions.

(32) Native Czech:

Požádal jsem ji, aby se naučila.CCOMP nějaký jazyk.
 asked.M.SG AUX.1SG her COMP+AUX.3 REFL learn.F.SG some language

‘I asked her to learn some language.’

6.2.5 Evaluation

The manual annotation of *CzeSL-UD* was done by two annotators: an annotator with a philological background and a secondary-school student.

They did not undergo any special training prior to the annotation, but instead relied on a secondary-school grammar training and the guidelines for Czech available at the UD project site.⁶ When they were not sure about a particular construction, they referred to existing Czech and English UD corpora, compiling a shared guide and a cheat sheet⁷ in the process. Technically, we used the *TrEd* editor with the *ud* extension to do the annotation.⁸ The annotators corrected a default structure obtained by running *UDPipe* (Straka and Straková 2017) on target hypothesis (T2) text and projecting the output to learner text (T0).

For a pilot annotation, we have randomly selected 100 sentences from *CzeSL-man* shorter than 15 tokens. We measured their IAA using Cohen’s *kappa* (see §5.4.4.1) on part-of-speech labels, syntactic labels and unlabeled heads. The IAA scores 0.934, 0.89, 0.927, respectively, are good but not perfect. However, we believe that the most important result of the pilot UD annotation is not the actual annotation, but the guidelines that can be used as a basis for other non-native languages.

The annotation is still a work in progress. Our goal is to eventually annotate the whole *CzeSL-man* corpus. So far more than 1600 sentences have been annotated.

⁶<https://universaldependencies.org/guidelines.html>

⁷<https://bit.ly/UDCheat>

⁸<https://ufal.mff.cuni.cz/tred>

Chapter 7

Annotation process

In this chapter, we discuss technical, procedural and practical aspects of the compilation and annotation of the *CzeSL* corpus. For a discussion of the annotation scheme, see §5 in case of error annotation, and §6 in case of linguistic annotation. The supporting tools are covered in §9.

7.1 Overview of the annotation process

The whole annotation process proceeds as follows. Most of the texts are processed in batches, and the individual steps are separated in time, space and people responsible for their correct execution. More recently, the texts can be processed as needed, also one by one. Thus the corpus can grow incrementally, with the texts ready for on-line searching soon after they become available as source documents (see §9.2.3).

1. Collection of texts: The original texts and their metadata are collected (see §3.1 and §4.4); handwritten manuscripts are scanned. For most texts, collection and procurement of metadata has been done in cooperation with the teacher or examiner. In some cases, especially in the assessment of the learner's proficiency, the metadata item is based on the teacher's estimate rather than on the performance of the learner in the text. This is why the information about the learner's CEFR level in the corpus is not always completely reliable.
2. Transcription: Scanned manuscripts are manually transcribed and anonymized (see §4.3 and §4.4); each transcription is checked by a supervisor (see §7.2).

3. Error annotation and linguistic annotation:

- Two-tier error annotation, including linguistic annotation of target hypothesis (2T; see §7.3.1)
- Multi-dimensional annotation (MD; see §7.6)
- Implicit error annotation (see §7.7)
- Syntactic annotation of the source text in the Universal Dependencies framework (UD; see §7.8)

Although these types of annotation are theoretically independent, there are some practical dependencies between them. The MD and UD annotation depend on the two-tier annotation: they use tokens derived from T1 and a default annotation based on T2. The implicit annotation is independent but compatible with the other annotation schemes. Implicit annotation can also be based on the annotation in one or more other annotation schemes. The annotation schemes can be integrated and used by a single corpus annotation, maintenance and search tool (see §9.2.3).

7.2 Transcription and anonymization of manuscripts

To transcribe hand-written texts, at first we used off-the-shelf editors supporting HTML with simple transcription codes. Later, we switched to XML markup and an XML-aware editor. For details about the transcription formats see §4.2.

The HTML-based format allows the transcribers to use a tool they are familiar with, which means that not much technical training is required. Some of the codes are supported via macros of the editor. This is how most of the hand-written texts in *CzeSL* were transcribed.

The decision to use HTML produced by an off-the-shelf editor was made intentionally to minimize training time and not to limit the pool of potential transcribers – it is hard enough to find people who know the rules of handwriting of speakers of language X, it is even harder to find experts who are also able to transcribe into XML. However, in retrospect we feel this was not a correct decision, because the efforts needed to review the transcripts clearly outweigh the benefit of using a widely known tool. First, it is really important to minimize the occurrence of errors in transcription as they influence all the subsequent annotation steps. It is easier to enforce formal correctness in an XML editor such as *XMLmind* than in an HTML editor. Second, the ability to learn to use an XML editor is actually a good indication of other abilities that are important in the transcription process, for example the ability to follow the formal rules of a transcription manual.

Starting with a new batch of collected manuscripts, all transcripts have been done in the *TEITOK* tool in the XML-based format. Although the tool does not validate the content of the markup according to an XML schema, the use of transcription and anonymization codes is facilitated by pre-defined keyboard shortcuts and the XML syntax is checked on the fly. The texts previously transcribed and anonymized in the HTML-based format are converted into the XML-based format.

7.3 Tiered error annotation

The tiered error annotation proceeds in the following steps:

1. Preprocessing: The transcript is converted into a format where T0 roughly corresponds to the tokenized transcript and T1 is set as equal to T0 by default. Both are encoded in PML, an XML-based format (see §7.3.3). The conversion includes basic checks for incorrect or suspicious transcription.
2. Manual error annotation: Errors in the text are manually corrected and tagged; each annotation is checked by a supervisor (see §7.3.1). Some texts are independently annotated twice (see §5.4.4).
3. Automatic annotation checks: Manually annotated texts pass through a series of automatic checks. Suspicious annotations are marked and manually reviewed.
4. Manual adjudication: Each doubly annotated text should be checked and adjudicated, resulting in a single annotated version. However, except for a small pilot, this has not been done yet, so a part of the corpus actually contains two independent annotations.
5. Linguistic (morphological) annotation: Target hypothesis is automatically annotated with lemmas and morphological tags, both full hypothesis on T2 (see §6.1.1) and individual words on T1 (see §6.1.2).
6. Automatic error annotation: Error information that can be inferred automatically is added by comparing original and emended words: type of spelling alternation, missing/redundant expressions, and inappropriate word order (see §5.4.5).

Conversion to PML (see §7.3.3), annotation, supervision and adjudication are done with the help of *feat*, an annotation editor designed as a part of the project (see §9.1.1). The storage of the documents and their flow within this process is managed by *Speed*, a purpose-built text management system (see §9.1.2).

7.3.1 Manual error annotation

Some of the transcribed texts are error-annotated manually according to the 2T scheme described in §5.4. The annotation was done in *feat* (see §9.1.1). The annotator corrected the text on appropriate tiers, modified relations between elements (by default all relations are 1:1) and annotated relations with error tags as needed. Figure 7.1 shows the annotation of a sample sentence as displayed by the tool. The top of the window shows the currently annotated part of the sentence, displaying the source text above the two annotation tiers. The context of the annotated text is shown both as a transcribed HTML document (bottom left of the window) and as a scan of the original document (bottom right).

In the annotated part of the window, forms identified by the tool as non-words are underlined, corrections done by the annotators are in red. Unless the annotator decides otherwise, vertical links align the words across the tiers 1:1. When the error type cannot be identified automatically, the annotator is supposed to replace the X label on the link between the incorrect form and its correction by one or more error tags. For some error tags, such as **agr** or **dep**, the annotator is instructed to provide a reference link to another word to substantiate the correction. It is usually the agreement source or the syntactic head of the corrected word.

Each annotation was reviewed by a supervisor, who could approve it, modify it, or return it to the annotator with comments for revision.

A subset of the texts annotated this way was independently annotated twice to assess the reliability of the annotation and the robustness of the tagset and the annotation scheme. After a pilot annotation, we used the result of the comparison to improve the annotation guidelines. See §5.4.4 for a detailed analysis of errors.

The annotation guidelines do not make any strict requirement about the sequence of steps in the error annotation, or about the relation of normalization and categorization as the two error annotation tasks. They only make an assumption that the two tasks are done by the same annotator, typically while annotating the whole text in one go. The annotators tend to normalize and categorize errors at the same time anyway. The advantage of this approach is that error tags reflect THs (see §5.1, p. 62 about the relation between TH and error categorization). On the other hand, separating annotation tasks in time and/or in the person of the annotator can result in better control of the annotator's judgments and thus more robust annotation. We followed this idea in *CzeSL-TH* (see §8.4), a part of *CzeSL*, which is corrected at T1 a T2 according to the 2T scheme, without error categorization. Error tags can be assigned in a separate step at any time later.

Given the 2T scheme, the annotators can also choose between annotating whole sentences, paragraphs or texts first on T1 and only then on T2, or annotating each

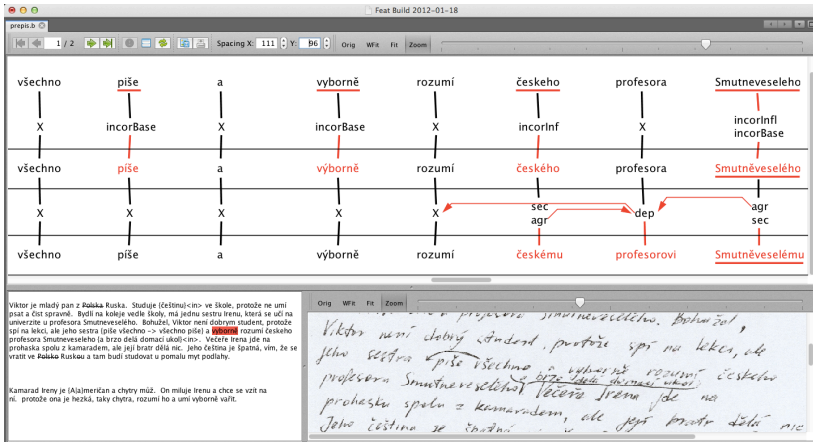


Figure 7.1: A sentence displayed in the *feat* annotation tool (see Table 5.1 on page 74 for the whole text) (NEM_GD_008 ru B2)

text in parallel on both tiers. Some annotators prefer to annotate by paragraphs, first annotating the whole paragraph on T1 and then on T2, while others annotate by sentences, annotating a sentence on both tiers in parallel before moving to the next sentence.

Despite the proofread status of the 2T annotation, additional checks by a different annotator as a part of the MD and implicit annotation have proved useful. This applies even to the doubly annotated part of *CzeSL-man*, due to the as yet unrealized plan of its adjudication. Manual categorization in the MD scheme is based on the TH made in the 2T scheme (more precisely, on its T2, see §5.6). The annotator can modify a TH which is clearly not correct. However, annotators are discouraged from substituting a more appropriate TH unless the existing TH is obviously wrong. In the implicit error annotation scheme (see §7.7), the annotators are free to use a TH suggested in 2T (if the text was annotated in 2T) or to use their own TH.

7.3.2 Automatic annotation checking

The system designed for automatic error tagging is also used for evaluating the quality of manual annotation, checking the result for tags that are probably missing or incorrect. For example, if a T0 form is not known to the morphological analyzer,

it is likely to be an incorrect word which should be corrected. Also, if a word was corrected and the change affects pronunciation, but no error tag was assigned, an `incorBase` or `incorInfl` error tag is probably missing. This approach cannot find all problems in error annotation, but provides a good approximate measure of the quality of annotation and draws the annotator's attention to potential errors.

7.3.3 Data format for the tiered annotation scheme

To encode the tiered annotation used in *CzeSL-man* (see §5.4), we have developed an annotation schema in the Prague Markup Language (PML).¹ PML is a generic XML-based data format, designed for the representation of rich linguistic annotation organized into tiers. Each of the higher tiers contains information about words on that tier, about the corrected errors and about relations to the tokens on the lower tiers.

We had also considered using a TEI format.² However, at least for stand-off layered annotation, the support offered by PML was superior to that of TEI, mainly in the availability of tools and libraries. This concerns tasks such as validation, structural parsing, corpus management and searching. While some of those libraries do exist for TEI, many would have to be developed.

More recently, we started using *TEITOK* §9.2.3 as the annotation and search tool, which explicitly supports some parts of the TEI standard. However, although it allows for stand-off annotation, its core uses the inline annotation format.

T0 does not contain any relations, only links to the neighboring T1. In [Figure 7.2](#), we show a portion (first two words and first two relations) of T1 of the sample sentence from [Figure 5.2](#), encoded in the PML data format.

To allow for data exchange, the *feat* editor supports import from several formats, including *EXMARALDA* (Schmidt 2009; Schmidt et al. 2011); it also allows export limited to the features supported by the respective format.

7.4 Automatic error tagging

After the manual error annotation in the 2T scheme the texts are automatically assigned tags identifying formal errors on T1 (see §5.4.5.2 for details). At the same time, some manually assigned error tags on T2 are automatically refined.

The tool (see §9.1.5) compares the source and the T1 forms. Any difference is assigned a formal error tag following rules implemented as an algorithm. On T2,

¹See Pajas and Štěpánek (2006) and <https://ufal.mff.cuni.cz/pml>.

²<https://tei-c.org/>


```

<?xml version="1.0" encoding="UTF-8"?>
<adata xmlns="http://utkl.cuni.cz/czesl/">
  <head>
    <schema href="adata_schema.xml" />
    <references>
      <ref file id="w" name="wdata" href="r049.w.xml" />
    </references>
  </head>
  <doc id="a-r049-d1" lowerdoc.rf="w#w-r049-d1">
    ...
    <para id="a-r049-d1p2" lowerpara.rf="w#w-r049-d1p2">
      ...
      <s id="a-r049-d1p2s5">
        <w id="a-r049-d1p2w50">
          <token>Bál</token>
        </w>
        <w id="a-r049-d1p2w51">
          <token>j sme</token>
        </w>
        ...
      </s>
      ...
      <edge id="a-r049-d1p2e54">
        <from>w#w-r049-d1p2w46</from>
        <to>a-r049-d1p2w50</to>
        <error>
          <tag>incorInfl</tag>
        </error>
      </edge>
      <edge id="a-r049-d1p2e55">
        <from>w#w-r049-d1p2w47</from>
        <to>a-r049-d1p2w51</to>
      </edge>
      ...
    </para>
    ...
  </doc>
</adata>

```

Figure 7.2: A part of T1 of the sample sentence (Figure 5.2 on page 85) encoded in PML (VOB_KA_049 kk A1)

```

<Meta>
  <task>
    <id>KAR_MI_005</id>
    <date>2010-04-21</date>
    <medium>manuscript</medium>
    <limit_minutes/>
    <aid>yes|dictionary</aid>
    <exam/>
    <limit_words/>
    <title>Jakou barvu má život?</title>
    ...
  </task>
  <student>
    <id>KAR_MI</id>
    <sex>f</sex>
    <age>19</age>
    <age_cat>16-</age_cat>
    <other_langs/>
    <cz_CEF>B2</cz_CEF>
    ...
    <l1>ru</l1>
    <l1_group>S</l1_group>
  </student>
  <annotation>
    <supervisorId>36</supervisorId>
    <annotatorId>15</annotatorId>
  </annotation>
</Meta>

```

Figure 7.3: A part of the metadata file `KAR_MI_005.meta.xml`, accompanying the text `KAR_MI_005` in the PML format

the tool is concerned mainly with errors in compound verb forms, selecting one of three subcategories using information on lemmas and morphosyntactic tags.

In *CzeSL* releases without the manual error annotation based on the 2T scheme, formal errors are identified by comparing source text forms with their THs, i. e., with corrections suggested by an automatic tool (see §7.5) or by a human annotator in the implicit annotation scheme (see §7.7). In the MD scheme (see §7.6), a semi-automatic error annotation method is used. Like formal errors, suggestions of some MD errors proposed to the annotator are generated by comparing the source and target forms (see §7.6.2). As in the 2T scheme, some of the manually assigned annotation is refined in a post-processing step (see §7.6.5).

7.5 Automatic correction

In this section, we discuss automatic correction as a way to normalize texts for which manual error annotation is not available. The results depend on the error type and on the immediate context of the error: the closer to standard Czech, the better chance of success (see §10.4.2).

One of the options to (partially) automate the correction task is to use a proofing tool – a spell checker or a grammar checker. So far, we have tested and applied *Korektor*,³ a spell checker that has some functionalities of a grammar checker, using a combination of lexicon, morphology and a syntax model.⁴

Korektor does not provide successive or domain-specific suggestions to match any of the annotation schemes used for the *CzeSL* texts (2T, MD or implicit). Specifically, its output does not match the successive corrections on the two tiers of the 2T annotation scheme. Although the tool does provide an n -best option to generate several suggestions ordered according to their assumed plausibility, in practice we always use a single suggestion for each error, i. e., the *autocorrect* mode. As long as the source form is a non-word, the suggestion often fits the definition of T1, without any successive correction (e. g., lexical) to fit T2. However, the tool can also correct real-word errors which are due to incorrect agreement and other morphosyntactic reasons, thus bypassing T1.

This is how the *CzeSL-SGT* corpus (see §8.2) was annotated. All non-native texts transcribed in the first round (2009–2012), including those error-annotated manually in the 2T scheme, were corrected using a single TH proposed by the tool and released with linguistic annotation. Automatic correction using the tool as a web service is also available in *TEITOK* as a substitute for manual annotation or to provide suggestions for the annotator.

The tool is concerned mainly with errors in orthography and morphemics, and handles some errors in morphosyntax, including real-word errors, as long as they are detectable locally, within a reasonably small window of n -grams. Corrections are limited to single words, targeting a single character or a very small number of characters by insertion, omission, substitution, transposition, addition, deletion or substitution of a diacritic. Errors that involve joining or splitting of word tokens or word-order errors of any type are not handled at the moment.

³See <https://ufal.mff.cuni.cz/korektor> and Richter (2010), Richter, Straňák, and Rosen (2012), Ramasamy, Rosen, and Straňák (2015), and Náplava and Straka (2019). There is a win-win cooperation between *Korektor* and *CzeSL*: the tool uses hand-annotated *CzeSL* texts as training data, see §10.4.2.

⁴Flor and Futagi (2012) report similar results for *ConSpel*, a tool used to detect and correct non-word misspellings in English, using n -gram statistics based on the *Google Web1T* database.

If a form detected as incorrect does not correspond to any Czech word (i. e., is a non-word), *Korektor* decides it is a spelling error. If it does correspond to a Czech word but is incorrect in the context (due to errors that produce a word which seems to be correct out of context), *Korektor* decides it is an error in grammar (i. e., a real-word error).

The performance of *Korektor* was evaluated first by Štindlová et al. (2012) with about 20% error rate on the set of non-words, and later by Ramasamy, Rosen, and Straňák (2015) on a larger sample. Form errors (resulting in non-words) were detected with a success rate of 89%. For grammar errors (real-word errors) the detection rate was much lower, about 15.5%. The detection of accumulated errors (a non-word corrected into a real word with a successive correction) was similar to form errors (89%). Náplava and Straka (2019) achieve even better results with a new system, which has not been used to annotate a released *CzeSL* corpus yet. For more about *Korektor* and its evaluation (see §10.4.2).

7.6 Multi-dimensional error annotation

Multi-dimensional (MD) error annotation (see §5.6) is designed to complement the 2T annotation. So far, all texts selected for the MD annotation have already been hand-annotated in the 2T scheme. The annotation proceeds in the following steps:

1. Texts annotated in the 2T scheme are converted to the vertical format with error annotation represented as structural markup (see §9.1.6). As far as the format allows, both T1 and T2 are preserved as the target hypotheses together with the corresponding error tags.
2. The source text (T0) and the final target hypothesis (T2) are extracted from the vertical format, preserving the word-to-word alignment links.
3. Inflected words in the source text are subjected to an automatic partial morphemic analysis, i. e., split into stems and inflectional affixes (see §7.6.1).
4. Source forms which are corrected, i. e., for which there is a different target hypothesis, are automatically assigned an error tag from the MD error tagset (see §7.6.2).
5. The automatically assigned error tags are checked, modified and extended by annotators in the *brat* annotation editor (see §7.6.4). An MD-annotated pilot dataset, revised by a supervisor, is available as the *CzeSL-MD* corpus (see §8.5).

6. The results of manual annotation are checked and extended in a post-processing step (see §7.6.5).

7.6.1 Morphemic analysis

The process of error classification, both manual and automatic, can be simplified by a preliminary automatic rule-based analysis, which can add useful information about each error. It can determine whether an error meets the criteria for an orthographic error or distinguish between errors in word stems from errors in inflectional affixes. The input to the automatic identification of errors is the source text and the final target hypothesis (TH), aligned word-to-word. We use the manual corrections from *CzeSL-man* for now, but we expect to use automatically corrected word forms in the future.

All identified inflectable words⁵ in the source texts (i. e., all identified inflectable words at T0) undergo a simple morphemic analysis. The analysis is simple in the sense that its ambition is not to divide the whole word into individual morphs, but only to mark inflectional prefixes and suffixes. For example, for the form *po|běž|í* ‘run.FUT’,⁶ we mark the prefix *po-* and the suffix *-í*.⁷ For verb participles, we mark both the participle suffix and the ending expressing gender and number: in *připraví|l|i* ‘prepared.MASC.PL’ *-l-* is the past participle suffix and *-i* is the plural masculine animate ending.

The procedure compares the source words on T0 and the corresponding words on T2. When the T0 word is correct, i. e., the T0 and T2 words are equal, we assume that they have the same morphological properties, and we use the morphological tag from T2 and segment the T0 word according to the T2 word. When the T0 is corrected, i. e., the T0 and T2 words are different, the situation is more complicated.

The automatic morphemic analysis proceeds as follows:

1. If the source form is not corrected (i. e., T0 = T2), we determine the inflectional suffix from the lemma and the tag assigned to the form, according to the rules of Czech inflection. The form *ledniče* ‘refrigerator.DAT’ with the lemma *lednička* and the tag NNFS6 (feminine noun in local singular) is analyzed as *lednič|ce*, based on its lemma and the phonetic alternation *k* → *c* in the local case of feminine singular.

⁵Inflectable words are identified according to the morphological tag automatically assigned to the word on T2 – every word tagged on T2 as a noun, adjective, pronoun, numeral, verb or a graded adverb.

⁶We use | to mark relevant morpheme boundaries.

⁷Here, *po-* is inflectional because it marks future tense of the verb *běžet* ‘run’; unlike in the verb *pomoci* ‘help’, where the prefix is derivational and is present in all forms of the verb. The suffix *í* expresses the 3rd person singular and plural for this verb.

2. If the source form on T0 is corrected on T2, but it is still an existing word (as determined by the morphological analysis) with the T2 lemma among its possible lemmas, the interpretation and therefore its morphemic analysis is based on the lemma of the corrected T2 form and the morphological tag of the source form. For example, the form *je* can be either the 3rd person singular present tense of the verb *být* ‘be’, or the accusative plural of the pronoun *on* ‘he’. In (33), *je* is corrected as *jsou* ‘are’ and is therefore analyzed as the verb *je* ‘is’ (*Vánoce* ‘Christmas’ is a plurale tantum).

(33)	<p>T0: <i>Vánoce</i> *je <i>nejdůležitější svátek.</i> Christmas is most-important holiday</p> <p>T2: <i>Vánoce</i> jsou <i>nejdůležitější svátek.</i> Christmas are most-important holiday</p> <p>‘Christmas is the most important holiday.’ (HRD_UE_347 ru A2)</p>
------	--

3. If the source form is not an existing (orthographically correct) Czech word,⁸ we attempt to guess its correct morphemic analysis based on the comparison with its correction. We perform the morphemic analysis of the word on T2 and compare its stem and inflectional affix with the word form on T0.
4. If the beginning of the source form matches the stem of the correction, we mark what follows as the source word’s inflectional suffix: *měsíc|y* → *měsíc|e* ‘month.PL.ACC’, *lázně|e* → *lázně|ě* ‘baths/spa’, *jmenu|em* → *jmenu|í* ‘name.1SG’, *zasp|ám* → *zasp|ím* ‘oversleep.FUT.1SG’.
5. If the source form ends with the characters of the corrected suffix and what precedes is similar to the corrected stem we assume that the source form has the same suffix as its correction: *poměnk|a* → *pomněnk|a* ‘forget-me-not’, *vopic|e* → *opic|e* ‘monkey’, *vyjadř|it* → *vyjádř|it* ‘express’, *glavn|í* → *hlavn|í* ‘main’.
6. If the source form and its correction differ in both their beginnings and ends, but their beginnings are similar enough and the stem-final consonant from T2 can also be found at a similar position on T0, we assume that the stem boundary follows this consonant on T0 as well: *spív|ají* → *zpív|á* ‘sing.3SG’, *cistěj|í* → *čistěj|í* ‘cleaner’, *vlastn|ou* → *vlastn|í* ‘own.ADJ’. If the number of differences is small, we also assume a partial match between the stem-final consonants (e. g., ignoring diacritics) *kultúr|u* → *kultūr|e* ‘culture.DAT/LOC’.

⁸In that case, it is assigned the morphological tag “unknown word” and a lemma identical to its form.

7. In all other cases, morphemic analysis of the form is skipped.

7.6.2 Automatic error annotation

Another part of the program for pre-processing learner texts before the manual error annotation is the automatic detection of errors (i. e., of differences between the forms on T0 and T2) and preliminary determination of their types. It is performed for corrected words: first the strings are compared and their differences are located, then the domain and type of error (“feature” – see §5.6) is identified. The program primarily classifies the detected errors into two domains: **ORT** (orthographic errors) and **MPHON** (morphological errors, i. e., errors affecting pronunciation). In a limited number of cases, it also identifies other errors, such as **CHOICE** (a lexical error). If the detected error type belongs to both the **ORT** and **MPHON** domains, only one of the domain (and one error type) is chosen and the annotators are instructed not to add the second error type, which will be filled in automatically after manual annotation.

The program compares the original and the corrected word forms from the beginning, character by character. If the forms begin differently (for example, if a prefix is missing in the source form: *rozumívát* → *dorozumívát* ‘understand’, or the prefix for the original and corrected form is different: *skončít* → *dokončít* ‘finish’), the program looks for the first position where the forms match (characters could be substituted, omitted, inserted or transposed, or the whole word may be different). To identify positions where the two forms match, a complete match is not required. It is sufficient if the characters are similar (graphically or in pronunciation): characters differing in upper/lower case, diacritics (*š* → *s*, *á* → *a*), voicing (*s* → *z*), *i* and *y* are considered similar. Similar partial matches are considered as half-errors. To consider a match of form portions, a half-matching character must be followed by another half-matching or a full-matching character. If the words are successfully aligned, all differences are marked.

Using simple rules, these differences might be split into several individual errors. For example, if the source word form *kultúru* is corrected as *kultuře* ‘culture.DAT/LOC’, the difference is at the fifth, sixth and seventh character: *úru* → *uře*. However, the program determines that these are three separate, unrelated errors: (i) an error in the vowel quantity of *ú* → *u* (**MPHON:QUANT** and **ORT:U**); (ii) an error in a missing diacritics *r* → *ř*, resulting in the failure to “soften” the consonant *r* (**MPHON:SOFT** and **ORT:DIA**); and (iii) an error where the ending *-u* is incorrectly replaced with *-e* (**MPHON:ALT**). The latter two errors are probably related to the use of an existing but wrong ending, but the program cannot determine this yet and such errors have to be marked manually. The program uses schemata (rules) to

identify several dozens of error subtypes (mainly in the domain of orthographic and morphonological errors). Some schemata are very simple (such as labeling errors in uppercase/lowercase letters, missing/inappropriate quantity), others are more complex, dependent on the phonetic environment, stem of the governing word, etc. (palatalization; decision whether to use error mark CHOICE etc.).

7.6.3 Experiments with automatic identification of errors in inflection

We experimented also with automatic identification of errors in inflection, which (if reliable enough) would significantly reduce the workload on manual annotators. The experiments had promising results, but we decided not to implement this module before the manual annotation would provide enough data to test it automatically. The following text describes this experiment and its (partial) results.

The experiment targets those words in the source text whose corrected form was identified as an inflectional word. Morphemic analysis, described in §7.6.1, was used to split both the source and the corrected word forms into a stem and an inflectional suffix (and sometimes prefix).

For example, if the incorrect source form is *stromom* ‘tree’ and the TH is *strom* with an empty inflectional suffix, the system does not compare only the two last characters of both words (which are identical by chance), but compares the entire stems and determines that the suffix of the original word is *-om* (*strom|om*). Using the stems and inflectional affixes for both the original and the TH forms, a two-dimensional comparison of the stems and the affixes is then performed. If the stems (original and TH) differ, two facts are checked: whether there are any minor errors in the stem (orthographic, phonological), and whether the source stem is an allomorph of the stem of the TH form, as in *v Prahe* → *v Praze* ‘in Prague’, where the original stem *Prah* (incompatible with the *-e* suffix) is used to form other (correct) forms of the same lemma, e. g., *Prah|a*, *Prah|y* etc.

If the affixes differ, they are also checked for minor changes (orthography, e. g., diacritics). Another check tests whether the incorrect affix is used within the given paradigm for other morphosyntactic properties, or whether the affix is used with other paradigms to express the same morphosyntactic properties. The observed differences correspond roughly to the proposed error classification scheme: all errors in orthography and most of the errors in morphonology can be identified automatically. Incorrect affixes indicate an error in morphology; if the incorrect ending is an existing one, expressing the same morphosyntactic properties, it may be an error only in morphology, otherwise it has to be seen as a possible error in syntax as well. If the original word is correct and has the same morphosyntactic properties as the

	A1	A2	B1	B2	C1	Total
Number of tokens	6,961	42,252	39,987	28,182	5,522	122,904
Percentage of the data	5.66	34.38	32.54	22.93	4.49	100.00

Table 7.1: Data distribution by language proficiency

	A1	A2	B1	B2	C1	Total
Correct	61.32	68.15	77.45	78.01	95.09	74.25
Incorrect ending	9.10	10.68	6.30	6.30	0.97	7.72
Incorrect stem	18.91	14.20	11.18	11.23	3.17	12.31
Incorrect whole	19.77	17.65	11.37	10.76	1.74	13.44
Total	100.00	100.00	100.00	100.00	100.00	100.00

Table 7.2: Proportion of correct and incorrect nouns by proficiency levels

TH word, but the lemma is different, the error may belong to the lexical domain (except for function words). The relationship between the automatic classification and the classification into error domains is not straightforward. A manual test on a sample of 500 learner errors shows that the approach is reliable with more than 90% of categories determined correctly. As the system is rule-based, it can be fine-tuned by modifying the rules.

We tested the rule-based system on nouns in the *CzeSL-man* corpus. Ill-formed nouns were identified as such using the disambiguated POS tags for corresponding corrected forms on T2. The texts were divided by language proficiency of the authors in terms of CEFR. The levels are not evenly distributed, as shown in [Table 7.1](#).

We performed two analyses of nouns in the *CzeSL-man* corpus: one more general, determining the proportion of incorrect nouns in the corpus, one detailed, focused only on errors in inflection suffixes of nouns.

[Table 7.2](#) shows the proportion of correct nouns, nouns with an incorrect suffix (*jeskyne* → *jeskyně* ‘cave’), with an incorrect stem and a correct suffix (*Prahe* → *Praze* ‘Prague.DAT/LOC’), with both stem and suffix incorrect (*delki* → *délky* ‘length.GEN’), and impossible to split automatically (*těmy* → *tématu* ‘topic.GEN’).

The proportion of correct nouns increases with the proficiency level, but there is little change between B1 and B2. On the other hand, there is an unexpectedly large difference between B2 and C1 in the proportion of correct nouns. The highest proportion of incorrect suffixes is in the A2 level texts.

We analyzed in more detail the errors in nominal suffixes: all nouns with either a correct stem, or with minor changes compared with the TH were examined. Two parameters were observed: whether the error in the suffix can be an error in orthog-

	A1	A2	B1	B2	C1	Total
Other paradigm	12.19	14.90	14.16	18.97	16.20	15.23
Other paradigm & spelling	8.87	4.51	7.08	6.91	7.82	6.83
Paradigm	19.38	30.77	22.83	24.03	24.02	25.34
Paradigm & spelling	4.03	3.24	5.25	5.14	11.73	4.40
Spelling	7.65	2.94	3.65	7.00	7.82	4.06
Other	47.88	43.64	47.03	37.94	32.40	44.14
Total	100.00	100.00	100.00	100.00	100.00	100.00

Table 7.3: Proportion of types of errors in endings of nouns

raphy, and whether the suffix is an existing Czech inflectional suffix used either to express the same case, number and gender in other paradigms, or is used in the same paradigm to express other morphosyntactic properties. Table 7.3 shows the analysis of errors in nouns in *CzeSL*. Six subtypes of nouns with incorrect suffix were registered:

Other paradigms: a suffix used in other paradigms (*Úvalách* → *Úvalech* ‘Úvaly.LOC’, a place name); likely syntactically correct

Other paradigms & spelling: a suffix used in other paradigms and with a spelling error at the same time (*Prázě* → *Praze* ‘Prague.DAT/LOC’)

Paradigm: a suffix used inside the paradigm for other morphosyntactic properties (*na procházka.*NOM* → *procházku.ACC* ‘on/for a walk’); this is probably an error in morphosyntax

Paradigm & spelling: as above, with a spelling error at the same time (*lidi* → *lidí* ‘people’)

Spelling: only a spelling error, none of the above (*pracé* → *práce* ‘work’)

Other: all other instances

We observe a steady decrease of “Other” errors, and an increase in the proportion of orthographic errors with language proficiency levels (the authors with a higher proficiency make less errors in general, but keep omitting diacritics). The system allows also for the analysis of individual suffixes: we observed, for example, that suffixes with high ambiguity such as *-e*, *-i*, *-í* are more prone to errors (already noted by Hudoušková 2014, 220).

7.6.4 Manual error annotation

The manual MD annotation is done in the *brat* annotation editor (see §9.1.3). The design and principles of the MD annotation scheme are described in §5.6 above.⁹

Figure 7.4 shows a text in *brat*, while Figure 7.5 shows the same text with a menu of error tags – labels of domains and features. The text has been pre-processed and manually annotated. As described above, pre-processing involves a partial morphemic analysis and automatic error annotation.¹⁰

1	C*o J*e PRO M*ně NEJ*DŮLEŽITĚJŠ*í ?	CAP CAP MNE DEP CAP
2	CO JE PRO MĚ NEJDŮLEŽITĚJŠÍ ?	
3	Vždycky , když s*e m*ně někd*o zept*á , c*o j*e pro m*ně nej*důležitějš*í v život*ě , vzpomínám s*i	MNE DEP MNE DEP LEX>ASP vid
	na rozhovor ze znám*ým polsk*ým filozof*em Leszk*ěm Kotakowskim , kter*ý js*em čet*i*a asi před dvěma let*y □	PREP NASIM aDIA NAFF
4	Vždycky , když se mě někdo zeptá , co je pro mě nejdůležitější v životě , vzpomenu si na rozhovor se známým polským filozofem Leszkiem Kotakowskim , který jsem četla asi před dvěma lety □	REPL
5	Kotakowski v něm vyprávě*i o život*ě , svět*ě a několika dalš*ích věc*ech a potom konečně řek*i , že nej*důležitějš*í pro něho js*ou (a vždycky by*i*i) jeho přátel*e □	aDIA NAFF ALT aDIA NAFF aDIA SOFT aDIA DEP
6	Kotakowski v něm vyprávěl o životě , světě a několika dalších věcech a potom konečně řekl , že nejdůležitější pro něho jsou (a vždycky byli) jeho přátelé □	aDIA SOFT aDIA DEP

Figure 7.4: A sample MD annotation in *brat*

(AA_A0_002 p1 B1)

Each sentence in Figure 7.4 is displayed twice. The pre-processed source version, corresponding to T0, comes first. Inflective words are split into morphs by asterisks and errors are tagged by error labels, corresponding to the feature name. Nearly all errors are detected and partially categorized automatically in the pre-processing

⁹For the MD annotation manual (in Czech) see Škodová et al. (2019).

¹⁰A pilot annotation of 18 texts, based on the annotation manual, can be viewed and searched using *brat* at https://quest.ms.mff.cuni.cz/brat/czesl.err/index.xhtml/#/anna_daniela/, or downloaded as a dataset in the *brat* format from <https://bitbucket.org/czesl/czesl-md/>.

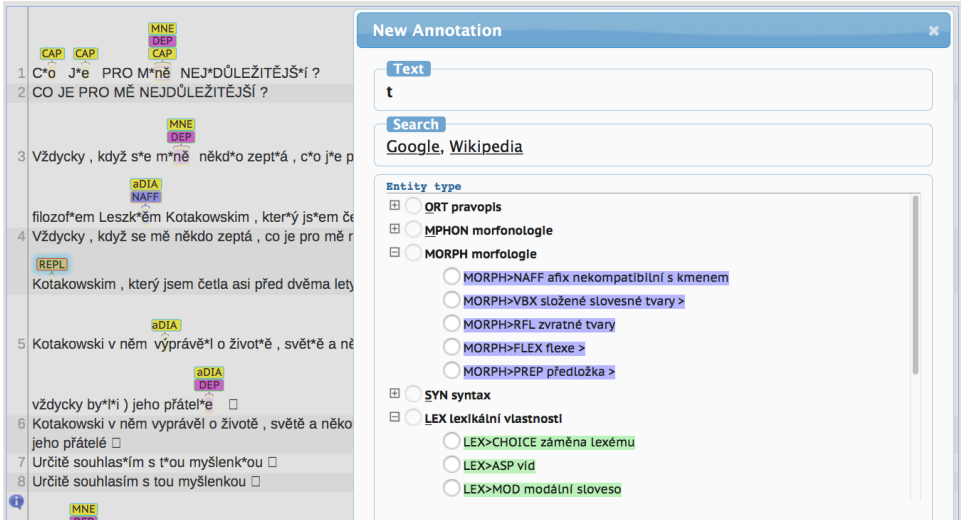


Figure 7.5: MD annotation in *brat* with the error tags menu (AA_A0_002 p1 B1)

step. The annotators proofread and modify the annotation by comparison with the target hypothesis of the sentence, shown below the source on the light grey background. The target hypothesis is adopted from the 2T annotation scheme. It is assumed to be correct, but can be modified when the annotator disagrees.

For space reasons, the error domains are distinguished by different colors rather than by more verbose display of the domain tag with a gloss. Error labels assigned in pre-processing are denoted by the letter “a” preceding the feature tag, as in **aDIA**. Some of the other **a**-type tags may have been modified by the annotator, other tags were added manually. The span of the error, shown below the error tags, may be identified correctly by the morphemic analysis, but the annotator is free to modify it or annotate a new error with its specific error span.

One of the main features of the MD annotation scheme is the option of multiple alternative interpretations of an error even in the case of a single target hypothesis. This might seem as an additional burden for the annotator. However, there are some regular patterns of co-occurring error tags, which are used in the pre-processing and post-processing steps. The patterns follow from the error taxonomy and most of them are easy to remember (cf. Figure 5.3 and Table 5.13 on page 118).

For example, a feature tag in the **MPHON** domain is deducible from an automat-

ically assigned tag in the `ORT` domain. As a result, the annotator need not worry about the annotation in the `MPHON` domain when an `ORT` domain tag is in place. The same rule applies also in the opposite direction. On the other hand, tags in the `LEX` domain, except for `CHOICE`, and in the `SYN` domain must always be specified by hand. However, if the `SYN` domain tag is `AGR` or `DEP`, then the `FLEX` tag in the `MORPH` domain is always appropriate and need not be specified, and – if either `ALT` or `CHAR` are the correct tags in the `MPHON` domain – they are not needed either.

To assist the annotator, the annotation editor is informed about possible combinations of tags and issues a warning whenever the annotator adds an incompatible tag for the same or overlapping span.

The annotation spans can be of arbitrary length, from a single character to a sequence of words, where some words need not be completely included in the span. For some error types (defined in the *brat* annotation setup), even discontinuous sequences of words or characters are allowed. These error types include `ORT:SEG`, `MORPH:VBX`, `SYN:WO` or `LEX:PHR`. For multiple errors concerning different parts of a single word form, it is useful to specify spans for multiple substrings of a word, i. e., a morph or even a single character. On the other hand, some tags can only be used for entire words. This applies mainly to the `LEX` and `SYN` domains, except for `SYN:AGR`, `SYN:DEP`, `LEX:NEG` and `LEX:USE`.

The MD annotation can be searched and viewed also in *TEITOK* (see §9.2.3). [Figure 7.6](#) shows the same text again, this time in the *TEITOK* stand-off annotation view. The view shows only the MD annotation. Annotated words are underlined, details of the annotation are shown on mouse-over. In the right-hand column the error codes used in the text are listed at the top. A click on the tag shows all words annotated by that tag. All annotated forms with the spans highlighted are shown in the list of similarly clickable forms below the error tags.

In *TEITOK*, the MD annotation can also be edited. Error tags can be deleted or added, and the span and error tag can be modified.

7.6.5 Post-processing of manually annotated texts

The manual annotation in *brat* is followed by post-processing. This step adds error tags identifying phenomena implied by the manually assigned tags. For example, some morphological tags entail orthographic errors: any voicing assimilation error, such as *hodně chyp* → *hodně chyb* ‘many errors’ is labeled as `MPHON:ASIM` (in this case, the spelling is incorrectly influenced by final devoicing), but it should be also labeled as `ORT:SUBST` (incorrect substitution of one letter by another). In cases when one error label implies another, the annotators are instructed to add only the former, the implied error is added by automatic post-processing.

CzeSL - multi-modální korpus
EN | CS
Hlavní nabídka

- Domů
- Dokumenty
- Hledat

user: AR

- Admin
- Help
- Custom annotation
- XML Files

Powered by TEITOK
© Maarten Janssen, 2014-

AA_AO_002
Error annotation using the multi-domain tagset

CO JE PRO MNĚ NEJDŮLEŽITĚJŠÍ?

Vždycky, když se mně někdo zeptá, co je pro mně nejdůležitější v životě, ze známým polským im, který jsem četla a mně v něm výprávěl o životě, a potom koněčně řekl, že nejdůležitější pro něho jsou (a vždycky byli) jeho přátelé. Určitě souhlasím s tou myšlenkou. Pro mně je taky nejdůležitější na světě přátelství. Citím to hodně v Praze, protože všichni moji přátelé zůstali v Krakově a proto se mi po nich moc stýská. Pišeme se často e-mailly, ale myslím, že to není stejné jako setkat se s někým tvář v tvář, jít na kávu a povídat si dvě hodiny nebo víc.

Pokud jde o další věci, moc důležitá je pro mně láska. Mohl by se někdo zeptat, proč není láska nejdůležitější?

Odpovím, že z toho důvodu, že každá láska má začátek a konec, ač přátelství je věčné. Člověk, kterého miluju je taky mým přítelem. Tuším ale, že láska může někdy skončit a potom zůstaně jenom přátelství.

Error code

- AGR
- ALT
- ASP
- CAP
- CHAR
- CONJ
- DEP
- DIA
- FLEX
- MNE
- NAFF
- NASIM
- PHR
- PREP
- PUN
- RFL
- USE

- Mám ráda
- ze
- a
- a
- Co
- ač
- ze
- se
- ale
- moje
- moje
- práci
- vzdělání
- ta
- ta
- Co
- je
- se
- výprávěl
- Citím
- Pišeme
- výprávět

Figure 7.6: A sample of MD annotation in *TEITOK* (AA_AO_002 p1 B1)

The post-processing algorithm has not been fully implemented yet. It is waiting for a sufficient amount of annotated texts and more feedback from the annotators. As a more distant step, we consider designing an algorithm that would replace the manual annotation of other, possibly all, MD error tags.

7.7 Implicit annotation

Manual error annotation can be easier when one of the two parts of the error annotation task is omitted (correction or categorization). In both of our two attempts to simplify error annotation this way we applied error correction, omitting catego-

rization.¹¹

The first approach is based on the previous experience with the 2T error annotation scheme. We used the same toolchain, including *feat* as the annotation editor, and the same annotation guidelines, including the distinction of T1 and T2 and the geometry of cross-tier links for splitting, joining and reordering tokens. However, no error tags were used.

In 2017, 1,300 texts (180 thousand tokens) were manually corrected. The texts were selected from the pool of texts annotated only by automatic tools in *CzeSL-SGT* to partially fill the under-represented groups of learners according to the combined L1 and CEFR specifications. The annotated texts are released as *CzeSL-TH* (see §8.4) and as a part of the *AKCES-GEC* dataset (see §8.7).

We have also tested and adopted an approach based on a sequence of target hypotheses corresponding to linguistic notions such as spelling, morphology, syntax or lexicon with a radically simplified error categorization part, using *TEITOK* as the annotation tool.¹²

The first major application of this type of annotation was in a corpus of native learners of Czech – *SKRIPT 2015* (see §8.9). The corpus is based on already existing transcripts. A part of the corpus overlaps with *SKRIPT 2012*, which was released without linguistic or error annotation. The texts were converted from transcripts using the original transcription markup into XML and, if necessary, manually anonymized. Then the texts were hand-corrected at four levels: (i) rectification of non-standard forms (resulting in a form that is still non-standard but spelled “correctly”), (ii) spelling and morphonology (correcting even “correctly spelled” non-standard forms), (iii) morphosyntax and (iv) lexicon. Most of the levels were tagged and lemmatized, and annotated with the formal error tags (see §5.4.5).

Due to a positive experience with this fairly large-scale manual annotation project, other *CzeSL* texts without manual annotation included in *CzeSL-SGT* are due to be annotated in the same way, while the already existing manually annotated parts of *CzeSL* will be integrated into the result – *CzeSL in TEITOK*. Importantly, the annotation in *CzeSL in TEITOK* is compatible with the 2T annotation scheme.

Based on experience and options, the following data can be used in a corpus built in *TEITOK*:

New texts (manuscripts or audio) can be transcribed and anonymized in *TEITOK* in the XML format

¹¹See §5.5 for more about implicit error annotation.

¹²See §9.2.3 for more about the tool. Several learner and historical corpora are available in *TEITOK*, with annotation based mainly on corrections.

Existing transcripts in the old format, possibly anonymized, can be converted into the *TEITOK* XML format using a conversion tool (see §9.1.6)

2T error-annotated texts – including texts without error tags, can be converted into the *TEITOK* XML format; some annotation can be expressed inline, other annotation (more complex cross-tier links) in a stand-off annotation format

MD error-annotation can be added to the *TEITOK* XML in the stand-off format

Once the data are included in a *TEITOK* corpus, they can be annotated in the following ways:

Error annotation

automatic: TH guessing (*Korektor* web service¹³), formal error tags

manual: successive corrections, implicitly specifying the error type (corresponding to 2T tiers and some 2T error tags or to MD error domains)

Linguistic annotation

automatic: lemmas and tags for the source and/or any correction level (*MorphoDiTa* web service¹⁴), syntactic structure (*UDPipe* web service¹⁵)

manual: checking and editing

7.8 Universal Dependencies

A syntactically annotated learner corpus, *CzeSL-UD*, was built according to the framework of Universal Dependencies as described in §6.2. The annotation proceeded in three steps:

1. Preprocessing: An automatic parse of the target hypothesis (T2) is projected to the source text (T0) as the default syntactic structure.
2. Manual annotation: The default syntactic structure is corrected as necessary by annotators using *TrEd*.

¹³<https://lindat.mff.cuni.cz/services/korektor/api-reference.php>

¹⁴<https://lindat.mff.cuni.cz/services/morphodita/api-reference.php>

¹⁵<https://lindat.mff.cuni.cz/services/udpipe/api-reference.php>

3. Adjudication: A double-annotated subset of data had differences resolved.

The manual annotation itself was done by two annotators: an annotator with a philological background and a secondary-school student. They did not undergo any special training prior to the annotation, but instead relied on a secondary-school grammar training and the guidelines for Czech available at the UD project site.¹⁶ When they were not sure with a particular construction, they referred to existing Czech and English UD corpora, compiling a shared guide and a cheat sheet¹⁷ in the process. Technically, we used the *TrEd* editor with the *ud* extension to do the annotation.¹⁸

As mentioned above, the annotation was not done from scratch, but the annotators corrected a default structure obtained by running *UDPipe* on target hypothesis (T2) text and projecting the output to learner text (T0). Obviously, using a default structure provides a certain bias, but we thought the bias to be minimal and the amount of manual work saved was quite large, so we decided it is worth the cost. Ideally, we would run a pilot study comparing the annotations resulting from annotations done from scratch and annotations based on correcting a default structure, but unfortunately this was not practically feasible.

However, we did perform a pilot annotation to validate a general reliability of the annotation. We double annotated a sample of the sentences and compared the results. We used the analysis of the results to improve the guidelines. The pilot also showed that the differences between independent annotations were relatively small. See §6.2.5 for more details.

¹⁶<https://universaldependencies.org/guidelines.html>

¹⁷<https://bit.ly/UDCheat>

¹⁸<https://ufal.mff.cuni.cz/tred>

Chapter 8

The *CzeSL* corpora

This chapter presents the tangible results: various releases of the corpus of non-native learners' Czech. The corpus releases are available via a concordancer or as full texts under the Creative Commons license.

Learner texts collected throughout the years have been released as several corpora: *CzeSL-plain*, *CzeSL-SGT*, *CzeSL-man* (in three versions), *CzeSL-TH*, *CzeSL-MD*, *CzeSL-UD*, *CzeSL-GEC* and *AKCES-GEC*. All these corpora include texts produced by non-native learners of Czech. Corpora including texts by native Czech learners built using similar methods and tools are only mentioned briefly (*AKCES 4*, *SKRIPT 2012*, *SKRIPT 2015* – see §8.9). Eventually, all non-native texts collected within the *CzeSL-man* project should be searchable from a single user interface with all available error and linguistic annotation (*CzeSL in TEITOK*).

The corpora differ in a number of aspects: (i) content, i. e., which subset of the whole pool they include; (ii) metadata, i. e., how much metadata about the texts and their authors they offer, if any; (iii) type of annotation, if any, i. e., the depth and method of linguistic and error annotation, (iv) whether they are annotated by hand or by automatic tools; (v) ways they can be accessed: whether they are downloadable as datasets and/or available for online searching.

Table 8.1 shows the content of available releases of *CzeSL*, including the size and the availability of annotation and metadata. The *CzeSL-* prefix in most of the names of the corpora is omitted for space reasons, except for *C(zeSL)-GEC* in contrast to *A(KCES)-GEC*. In the error annotation column, error categorization (Tags) and error correction (TH) is distinguished. Linguistic annotation is specified according to the 2T scheme. For corpora which do not follow the 2T scheme, T0 should be understood as the source text and T2 as the (only) TH. The abbreviations

have the following meaning:

- Tags: F – formal, G – grammar-based, MD – multi-dimensional, I – implicit
- TH: K – correction suggested by the proofing tool, 2T – successive corrections in the 2T scheme, T2 – correction at Tier 2, 2T+ – more than 2 successive corrections
- Linguistic annotation: M – morphology (lemmas and morphosyntactic tags), S – syntax (structure and functions)
- Access: S – searchable on-line, D – downloadable in full as a dataset
- Year: year of the first release

	Thousands of tokens in				Error annotation		Linguistic annotation			Meta-data	Access	Year
	Non-native Essays	Theses	Ethno-lect	Σ	Tags	TH	T0	T1	T2			
<i>plain</i>	1,315	732	428	2,475	–	–	–	–	–	–	SD	2012
<i>SGT</i>	1,147	–	–	1,147	F	K	M	–	M	yes	SD	2014
<i>man v0</i> , a1 ¹	134	–	192	326	F+G	2T	–	M	M	–	SD	2012
<i>man v0</i> , a2	59	–	149	208	F+G	2T	–	M	M	–	S	2012
<i>man v1</i>	134	–	–	134	F+G	T2	M	–	M+S	yes	SD	2016
<i>man v2</i>	134	–	–	134	F+G	2T	M	M	M	yes	SD	2020
<i>TH</i>	180	–	–	180	–	2T	–	–	–	yes	D	2018
<i>MD</i>	12	–	–	12	MD	T2	–	–	–	–	D	2018
<i>UD</i>	10	–	–	10	–	–	M+S	–	–	–	D	2018
<i>C-GEC</i>	?	?	–	20	–	2T	–	–	–	–	D	2017
<i>A-GEC</i> ²	336	–	168	504	G	2T	–	–	–	–	D	2019
<i>TEITOK</i> ³	299	–	–	299	F+I	2T+	M	M	M+S	yes	S	2020

Table 8.1: Available releases of *CzeSL*

¹Some texts in *CzeSL-man v.0* are doubly annotated. The texts annotated by an additional annotator are included in the *CzeSL-man v.0*, a2 part. See <http://utkl.ff.cuni.cz/learncorp/> for links and more details.

²Includes some texts annotated by an additional annotator.

³Work in progress, the number of tokens stands for manually annotated texts, which are not included in *CzeSL-man*. Eventually, non-native texts from the other *CzeSL* corpora will become part of *CzeSL* in *TEITOK*, together with newly collected texts. Currently, the corpus includes about 1300 new transcripts of written essays.

8.1 *CzeSL-plain* – without annotation and metadata

The *CzeSL-plain* corpus (plain = without annotation),⁴ released in 2012, contains about 12.4 thousand texts, totaling approximately 2.5 million tokens (about 2 million words). It includes transcripts of essays hand-written by non-native learners and pupils speaking the Romani ethnolect of Czech together with some bachelor's, master's and doctoral theses written in Czech by foreign students. Except for specifying the three groups above and a basic structural mark-up, this corpus does not include any annotation or metadata about the author or about the text itself. The three groups are identified in the *KonText* search interface and in the XML headers of the dataset as the following three text types:

ciz Transcripts of essays written by non-native speakers in language teaching classes of various types and levels; the size: 8,109 texts, i. e., 1,161 thousand tokens

kval Academic texts obtained from non-native speakers of Czech studying at Czech universities in Masters or doctoral programs; the size: 174 texts, i. e., 732 thousand tokens

rom Transcripts of texts written at school by pupils and students speaking the Romani ethnolect of Czech; the size: 4,105 texts, i. e., 428 thousand tokens

The first two subcorpora concern Czech as a second/foreign language, while the third part would be more appropriately viewed as a L1 acquisition subcorpus.⁵ This is the first publicly available corpus of this type for Czech.

The texts written by non-native speakers (the **ciz** part), extended by some newer texts, are available as the *CzeSL-SGT* corpus, together with metadata and automatically performed morphosyntactic and error annotation, including the identification of incorrect forms.

The corpus is on-line searchable via the web-based search interface of the Czech National Corpus,⁶ or available as full texts under the Creative Commons license from the LINDAT repository⁷ as two subcorpora: *AKCES 3* contains the **ciz** and **kval** parts while *AKCES 4* includes the **rom** part.

⁴<https://wiki.korpus.cz/doku.php/en:cnk:czesl-plain>

⁵Czech is not considered to be a foreign language for students speaking the Romani ethnolect of Czech (see §B.4).

⁶<https://kontext.korpus.cz>

⁷<https://lindat.mff.cuni.cz>

The corpus includes the HTML-based transcription markup, i. e., codes used in the transcription of the manuscripts and in the encoding of some foreign and non-standard characters §4.2. This is why the number of characters in the corpus is somewhat higher than in the original texts.

8.2 *CzeSL-SGT* – with automatic annotation

Essays written by non-native learners are available with automatic annotation as *CzeSL-SGT*.⁸ As the first public release of *CzeSL* with full metadata, this corpus extends the “foreign” part of *CzeSL-plain* by texts collected in 2013. The corpus includes 8,617 texts (1.1 mil. tokens or 958 thousand words in 111 thousand sentences) by 1,965 different authors with 54 different first languages. The text corresponds to T0 without transcription markup.⁹

The corpus can be searched in *KonText*, the search interface of the Czech National Corpus (see §9.2.2).¹⁰ The corpus can also be downloaded from the LINDAT data repository.¹¹

CzeSL-SGT includes both linguistic and error annotation. All annotation is provided by automatic tools. Each token is labeled by the attributes described in Table 8.2. The annotation consists of the following steps:

1. All source word forms, both correct and incorrect, are tagged by lemma and morphological tag. The tagger identified 9.23% of tokens as non-words.
2. Some forms are corrected by *Korektor*, a context-sensitive spelling/grammar checker (see §7.5 and §10.4.2). *Korektor* detected as incorrect and corrected 13.24% forms in *CzeSL-SGT*, including 10.33% labeled as a spelling error, and 2.92% as an error in grammar, i. e., a real-word error. The share of non-words (10.33%) detected by *Korektor* is slightly higher than by the tagger (9.23%) because the tagger uses a larger lexicon.
3. The corrected text (i. e., with some word forms corrected by *Korektor*, the rest of the forms copied from the source text) is tagged and lemmatized again.

⁸ *Czech as a Second Language with Spelling, Grammar and Tags*

⁹For more details see Rosen (2017).

¹⁰To query *CzeSL-SGT* go to https://kontext.korpus.cz/first_form?corpname=czesl-sgt. For help on using the search interface see https://wiki.korpus.cz/doku.php/en:manually:kontext:novy_dotaz.

¹¹<https://hdl.handle.net/11234/1-162>

4. For all corrected word forms the source form and the corrected form are compared and errors are assigned error tags (see §5.4.5).

word	source word form
lemma	lemma of word ; same as word if the form is not recognized
tag	morphological tag of word ; if the form is not recognized: X@----- ¹²
word1	corrected word form; same as word if determined as correct
lemma1	lemma of word1
tag1	morphological tag of word1
gs	information on whether the error was determined as a spelling (S) or grammar (G) error; for grammar errors, word is mostly recognized
err	error type, determined by comparing word and word1

Table 8.2: Attributes used in the annotation of tokens in *CzeSL-SGT*

Results of the automatic annotation of *CzeSL-SGT* are illustrated in the following two examples. The sentence in (34), annotated in Table 8.3, shows how spelling and morphological errors are annotated in *CzeSL-SGT*.¹³

(34) S: **Tén pes *míluje *svěcho *kamaráda – člověka.*
(that) dog (loves) (self’s) (friend) – man

T: *Ten pes míluje svěho kamaráda – člověka.*
that dog loves self’s friend – man

‘That dog loves its friend – the man.’ (ttt_G1_434 ru B1)

Example (35) shows the use of the annotation in a sentence with a real-word error (*postele* → *posteli* ‘bed’), analyzed by *Korektor* as an error in grammar (gs=G), more specifically in case (genitive → local). The word **Nejakij* ‘some’ requires diacritics (*e* → *ě*) and correction to its non-existent ending *-ij*. The most straightforward way is substituting *i* with *e*, resulting in the colloquial ending *-ej*. This is what *Korektor* did. On the other hand, the correction resulting in the standard Czech form *Nějaký* is more “costly”: it requires replacing two characters with one character.

¹³The Czech morphological tagset is described at https://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html or https://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/docc0pos.pdf. The positional tags are truncated to save space – any unspecified or irrelevant trailing positions are omitted.

word	lemma	tag	word1	lemma1	tag1	gs	err
<i>Tén</i>	Tén	X@	<i>Ten</i>	ten	PDYS1	S	Quant1
<i>pes</i>	pes	NNMS1	<i>pes</i>	pes	NNMS1		
<i>míluje</i>	míluje	X@	<i>miluje</i>	milovat	VB-S---3P	S	Quant1
<i>svécho</i>	svécho	X@	<i>svého</i>	svůj	P8MS4	S	Voiced
<i>kamarada</i>	kamarada	X@	<i>kamaráda</i>	kamarád	NNMS4	S	Quant0
-	-	Z:	-	-	Z:		
<i>člověka</i>	člověk	NNMS2	<i>člověka</i>	člověk	NNMS4		
.	.	Z:	.	.	Z:		

Table 8.3: *CzeSL-SGT*: annotation of a sentence with spelling errors (34)

(35) S: **Nejakij muž spí v *postele.*
 (some) man sleeps in bed.*GEN

T: *Nějakej/ý muž spí v posteli.*
 some.COLL/STD man sleeps in bed.LOC

‘Some guy is sleeping in the bed.’

(UJA2_4P_005 uk A1)

word	lemma	tag	word1	lemma1	tag1	gs	err
<i>Nejakij</i>	Nejakij	X@	<i>Nějakej</i>	nějaký	PZYS1-6	S	Caron0
<i>muž</i>	muž	NNMS1	<i>muž</i>	muž	NNMS1		
<i>spí</i>	spát	VB-S---3P	<i>spí</i>	spát	VB-S---3P		
<i>v</i>	v	RR--4	<i>v</i>	v	RR--6		
<i>postele</i>	postel	NNFP4	<i>posteli</i>	postel	NNFS6	G	SingCh
.	.	Z:	.	.	Z:		

Table 8.4: *CzeSL-SGT*: annotation of a sentence with a spelling and a grammar error (35)

CzeSL-SGT is released with all available metadata (see §4.4).¹⁴ For the number of texts authored by students according to their first language and the CEFR proficiency level in Czech see Table 8.5. The language group abbreviations read as follows: S = Slavic, IE = non-Slavic Indo-European, nIE = non-Indo-European.

Some or even all metadata items may be missing for some texts: identification of the author is present in 96.7% texts, the first language in 96.3% texts. Missing

¹⁴For a list of all attributes and values in Czech and English see http://utkl.ff.cuni.cz/~rosen/public/meta_attr_vals.html. The numbers of documents, listed according to specific attribute values, are given here: http://utkl.ff.cuni.cz/~rosen/public/sgt_counts_by_meta_en.html.

	S	IE	nIE	unknown	Σ
A1	1783	199	622	5	2609
A1+	283	21	11	0	315
A2	1348	269	480	1	2098
A2+	403	54	113	0	570
B1	929	195	357	0	1481
B2	523	115	107	0	745
C1	82	17	24	0	123
C2	0	1	0	0	1
unknown	291	27	33	324	675
Σ	5642	898	1747	330	8617

Table 8.5: Number of texts by language group and proficiency level in *CzeSL-SGT*

items are represented as empty values. Some attributes may include multiple values, delimited by vertical bar (“|”).

The metadata are available in Czech and English. The Czech version is available in *KonText* from the Czech National Corpus site, while the LINDAT data repository offers the English version.

Metadata and structural annotation are represented as XML elements with corresponding attributes. In Release 1,¹⁵ the text itself is represented in the vertical format, i. e., as tab-delimited columns, in the order shown in Table 8.3. For a sample of the tabular format, see Figure 8.1. In Release 2,¹⁶ the whole corpus is an XML document with each text as a `div` element. Annotation of each word is represented as XML attributes of a `word` element, see (36).

```
(36) <word lemma="dival"
      tag="X@-----"
      word1="dival"
      lemma1="divat"
      tag1="VpYS---XR-AA---"
      gs="S"
      err="Quant0">
      dival
    </word>
```

¹⁵<https://hdl.handle.net/11858/00-097C-0000-0023-95B1-E>

¹⁶<https://hdl.handle.net/11234/1-162>

```

<doc t_id="UJA2_PH_003" t_date="2010-12-21" t_medium="manuscript" t_limit_minutes="45"
t_aid="none" t_exam="yes|interim" t_limit_words="25" t_title="E-mail kamarádce/kamarádovi"
t_topic_type="general" t_activity="" t_topic_assigned="specified" t_genre_assigned="specified"
t_genre_predominant="informative" t_words_count="30" t_words_range="-50" s_id="UJA2_PH"
s_sex="m" s_age="17" s_age_cat="16-" s_L1="vi" s_L1_group="nIE" s_other_langs=""
s_cz_SER="A1" s_cz_in_family="" s_years_in_CzR="" s_study_cz="university"
s_study_cz_months="" s_study_cz_hrs_week="15-" s_textbook="NCSS" s_bilingual="no">

<s id="1">
mám mít VB-S---1P-AA--- mám mít VB-S---1P-AA---
dobře dobře Dg-----1A---- dobře dobře Dg-----1A----
. . Z:----- . . Z:-----
</s>
<s id="2">
V v RR--4----- V v RR--4-----
neděli neděle NNFS4----A---- neděli neděle NNFS4----A----
dival dival X@----- dival divat VpYS---XR-AA--- S Quant0
jsem být VB-S---1P-AA--- jsem být VB-S---1P-AA---
se se P7-X4----- se se P7-X4-----
na na RR--6----- na na RR--6-----
televizi televize NNFS6----A---- televizi televize NNFS6----A----
a a J^------ a a J^------
uklízěl uklízěl X@----- uklízet uklízet VpYS---XR-AA--- S Quant0|Caron1
jsem být VB-S---1P-AA--- jsem být VB-S---1P-AA---
. . Z:----- . . Z:-----
</s>
<s id="3">
Ano ano TT----- Ano ano TT-----
přijdu přijít VB-S---1P-AA--- přijdu přijít VB-S---1P-AA---
se se P7-X4----- se se P7-X4-----
tebe ty PP-S2--2----- tebe ty PP-S2--2-----
do do RR--2----- do do RR--2-----
kina kino NNNS2----A---- kina kino NNNS2----A----
a a J^------ a a J^------
taky taky Db----- taky taky Db-----
mám mít VB-S---1P-AA--- mám mít VB-S---1P-AA---
čas čas NNIS4----A---- čas čas NNIS4----A----
jen jen TT----- jen jen TT-----
večer večer Db----- večer večer Db-----
, , Z:----- , , Z:-----
večer večer Db----- večer večer Db-----
půjdemě jít VB-P---1F-AA--- půjdemě jít VB-P---1F-AA---
do do RR--2----- do do RR--2-----
kina kino NNNS2----A---- kina kino NNNS2----A----
. . Z:----- . . Z:-----
</s>
<s id="4">
Tvoje tvůj PSHS1-S2----- Tvoje tvůj PSHS1-S2-----
kamarád kamarád NNMS1----A---- kamarád kamarád NNMS1----A----
. . Z:----- . . Z:-----
</s>
</doc>

```

Figure 8.1: A sample of the *CzeSL-SGT* data in the vertical format

(UJA2_PH_003 vi A1)

8.3 *CzeSL-man* – with manual annotation

There are three releases of *CzeSL-man*. All of them consist of transcripts of essays, hand-written in 2009–2013, annotated by humans using the two-tier (2T) error annotation scheme. The annotation includes corrections of the original text (manual), error types (manual and automatic), and morphosyntactic categories and lemmas for the corrected text (automatic).

For every original sentence there are two successive target hypotheses, the first correcting individual forms, disregarding the context; the second correcting the words in context, with a correct Czech sentence as a result. Every correction on both tiers has an error label, with some 30 error labels assigned manually, completed by some 50 automatically assigned error labels.

For details about the manually assigned tags in the 2T scheme see §5.4.2. For the formal error labels assigned automatically see §5.4.5.

8.3.1 *CzeSL-man v0*

CzeSL-man v0 includes subsets of the **ciz** and **rom** parts of *CzeSL-plain*, i. e., Czech texts by non-native learners and by speakers of the Roma ethnolect, the total of about 330 thousand tokens. Texts of about 208 thousand tokens are annotated independently by two annotators. *CzeSL-man v0* is accessible online without metadata via a purpose-built search tool (*SeLaQ*; see §9.2.1).

8.3.2 *CzeSL-man v1*

CzeSL-man v1 contains a subset of texts included in *CzeSL-SGT*. It is a collection of manually annotated transcripts of essays written by non-native learners of Czech, native speakers of 32 different languages. The total of 645 texts (128 thousand tokens), include 298 doubly annotated texts (59 thousand doubly annotated tokens). For a comparison with *CzeSL-SGT* see [Table 8.6](#).

For the number of texts authored by learners according to a combination of their first language and proficiency level in Czech see [Table 8.7](#) and [Table 8.8](#) below.

In addition to the number of tokens for the same category, [Table 8.9](#) shows also the frequency of errors of the **dep** type, i. e., valency errors in the broad sense, including errors in the number of complements and adjuncts or errors in their morphosyntactic expression. The rather frequent error type shows a considerable and expected decrease in higher proficiency levels.

Most texts are equipped with metadata about the author, the text and the annotation process. See §4.4 for details. Two additional metadata items are avail-

	<i>CzeSL-SGT</i>	<i>CzeSL-man v1</i>
Number of texts	8,600	645
Number of sentences (in thousands)	111	11
Number of words (in thousands)	958	104
Number of tokens (in thousands)	1,148	128
Number of different authors	1,965	262
Number of different native languages	54	32
Proficiency levels	A1–C2	A1–C1
Share of women:men	5:3	3:2
Number of words per text	100–200	100–200

Table 8.6: *CzeSL-man v1* and *CzeSL-SGT* compared

	S	IE	nIE	unknown	Σ
A1	49	6	4		59
A1+			3		3
A2	18	26	67		111
A2+	81	9	59		149
B1	123	26	30		179
B2	102	11	15		128
C1	10		2		12
unknown				4	4
Σ	383	78	180	4	645

Table 8.7: Number of texts by language group and proficiency level in *CzeSL-man v1*

able to keep track of the manual annotation: the ID of the annotator and the supervisor. Missing items are represented as empty elements. Some attributes may include multiple values, delimited by vertical bar (“|”). The items are included in the `*.meta.xml` files.

8.3.3 *CzeSL-man v1* downloadable

This release is in the PML format, generated by *feat*.¹⁷ Each text with its annotation consists of several related files – see Table 8.10. Some of the texts are independently annotated twice – the `annotation1` and `annotation2` folders contain two parallel annotations of the same set of documents. The `annotation2` folder contains a proper subset of the texts in `annotation1` folder.

¹⁷The dataset is available at <https://bitbucket.org/czesl/czesl-man/>.

	S	IE	nIE	Σ
A1	37	2	1	40
A1+			3	3
A2	5	23	47	75
A2+	21	6	49	76
B1	20	23	28	71
B2	7	11	12	30
C1	1		2	3
Σ	91	65	142	298

Table 8.8: Number of doubly annotated texts by language group and proficiency level in *CzeSL-man v1*

	A1	A2	B1	B2	C1	Σ
IE	227	7,336	5,311	2,340	0	15,214
dep	13	361	118	28	0	520
%dep	5.73%	4.92%	2.22%	1.20%		3.42%
nIE	439	17,640	7,606	4,219	760	30,664
dep	13	715	237	116	7	1,088
%dep	2.96%	4.05%	3.12%	2.75%	0.92%	3.55%
S	6,434	16,939	27,226	22,173	4,761	77,533
dep	225	470	652	443	17	1,807
%dep	3.50%	2.77%	2.39%	2.00%	0.36%	2.33%
Σ	7,100	41,915	40,143	28,732	5,521	123,411
dep	251	1,546	1,007	587	24	3,415
%dep	3.54%	3.69%	2.51%	2.04%	0.43%	2.77%

Table 8.9: Number of tokens and valency errors by language group and proficiency level in *CzeSL-man v1*

8.3.4 *CzeSL-man v1* searchable

This release is available for on-line searching using *KonText*, the search interface of the Czech National Corpus.¹⁸ The release differs from both *CzeSL-man v0* and *CzeSL-man v1 downloadable* in two aspects: (i) there are no texts with alternative error annotation: each text is annotated by a single annotator (just one version of each doubly annotated text is included), and (ii) the two-tier annotation scheme is radically modified to fit the token-based setup of the search tool. Apart from that, the content and metadata are identical to *CzeSL-man v1 downloadable* and the search options to those of *CzeSL-SGT*.

¹⁸https://kontext.korpus.cz/first_form?corpname=czesl-man

*.jpg	Scan of the handwritten original (not part of the distribution, for privacy reasons)
*.html	Transcription of the handwritten original (anonymized)
*.meta.xml	Metadata about the document, its author and annotation
*.w.xml	Tokenized text (T0)
*.a.xml	Annotation of the text at T1
*.b.xml	Annotation of the text at T2

Table 8.10: Files constituting an annotated text in *CzeSL-man v1 downloadable*

The main feature in the annotation of this release is the reversal of the source text and its annotation. The target hypothesis at T2, the corrected text, is assumed to be the basis for the annotation. The tokens of this corpus represent the words at T2. The original text is added as annotation of the T2 tokens. Each token of the corrected text receives its corresponding T0 form and a T2 error label as attributes. This annotation discards any T1 corrections and error tags, and simplifies other than 1:1 links between tokens at T0 and T2. See [Table 8.11](#) for a list of attributes representing the basic error and morphosyntactic annotation in this release. As in *CzeSL-SGT*, dynamic attributes derived from the morphosyntactic tags can be used in queries and visualization of the results, see [Table 8.12](#).

word	T2 corrected form; same as word0 if determined as correct
lemma	Lemma of word
tag	Morphological tag of word
err	T2 error tag of word , if any
word0	T0 form (the source)
lemma0	Lemma of word0 ; same as word0 if the form is not recognized
tag0	Morphological tag of word0 ; if the form is not recognized: X@-----

Table 8.11: Token attributes used in *KonText* for the annotation of *CzeSL-man v1 searchable*

This radical simplification of the two-tier error annotation scheme was designed to provide access to the manually annotated texts, especially to the morphosyntactic aspect of the annotation, including the context-based corrections. The supposedly correct Czech text also allowed for a more reliable application of a tagger and a parser trained on standard Czech texts. As a result, the *KonText* release of *CzeSL-man v1*, i. e., the T2 target hypothesis, is parsed in a way similar to some other Czech corpora searchable in *KonText*, such as *SYN2015*. For a list of syntax-related

k0, k	Word class (position 1 of the tag)
s0, s	Detailed word class (position 2 of the tag)
g0, g	Gender (position 3 of the tag)
n0, n	Number (position 4 of the tag)
c0, c	Case (position 5 of the tag)
p0, p	Person (position 8 of the tag)

Table 8.12: Dynamic attributes used in *KonText* for *CzeSL-man v1 searchable*

attributes assigned to each token see [Table 8.13](#).

proc	Disambiguation processing step
afun	Syntactic function
parent	Relative pointer to parent
eparent	Relative pointer to effective parent
prep	Preposition as parent
p_tag	Parent tag
p_lemma	Parent lemma
p_afun	Syntactic function of the parent
ep_tag	Effective parent tag
ep_lemma	Effective parent lemma
ep_afun	Effective parent afun
lc	Lowercase T2 word
lemma_lc	Lowercase T2 lemma
p_k	Parent category (POS)
p_c	Parent case

Table 8.13: Syntax-related attributes used in *KonText* for *CzeSL-man v1 searchable*

Some solutions to the problem of using a feature-rich corpus search engine, which is still not suited to the two-level annotation scheme of *CzeSL-man*, are presented in §11.

8.3.5 *CzeSL-man v2*

In this release the two-tier error annotation is represented as pairs of XML elements `err` and `corr`. An ill-formed portion of the source text is enclosed within the `err` structure, immediately followed by its correction, enclosed within the `corr` structure.

There can be multiple tokens within an **err** or **corr** structure to represent split or joined tokens. The elements can be embedded to cope with successive corrections. In the 2T scheme, an **err** and **corr** pair can be embedded to represent a T1 spelling correction within a larger **err** structure followed by an **corr** structure representing a T2 lexical correction.

Linguistic token-based annotation is possible together with error annotation spanning multiple tokens. However, word-order corrections are still not easy to represent, especially when they involve long-distance moves. This is the reason why some corrections represented in the two-tier scheme are not implemented even in *CzeSL-man v2*.

For details about the error annotation of *CzeSL-man v2* see §9.2.2. Apart from the error annotation, the content and metadata are the same as in *CzeSL-man v1* and the linguistic annotation (tags and lemmas) is provided for all tokens at T0 and T2.

8.4 *CzeSL-TH*

This corpus includes a subset of *CzeSL-SGT*, hand-corrected, but not error-tagged, in 2017–2018, according to the 2T scheme. The corpus includes about 1300 texts (180 thousand tokens), selected from those that had not been manually error-annotated before (i. e., are not part of *CzeSL-man*). The selection was meant to make the manually annotated part of *CzeSL* more balanced in terms of L1 and CEFR level.

8.5 *CzeSL-MD*

This corpus includes a subset of *CzeSL-man*, semi-automatically annotated by the MD tagset – see §5.6. The texts were annotated in multiple experimental rounds by different versions of the tagset. The current version is a corrected and adjudicated version of a dataset including 10 thousand words, doubly hand-annotated, using suggestions by a pre-processing module. The texts are available from <https://bitbucket.org/czesl/czesl-md> in the *brat* format (see §9.1.3).

8.6 CzeSL-UD

This corpus is a subset of *CzeSL-man v1* with syntactic annotation according to the Universal Dependencies (UD) standard.¹⁹ It includes 1645 annotated sentences, out of it 100 sentences are doubly-annotated with good inter-annotator agreement. It is probably the second largest UD-annotated learner corpus (after *TLE* – <http://esltreebank.org>). There is no error annotation in the texts, the texts are not normalized or assigned error labels. The words are annotated with POS, morphological categories, syntactic function and their syntactic head without an explicit target hypothesis.

8.7 CzeSL-GEC and AKCES-GEC

The *CzeSL Grammatical Error Correction Dataset (CzeSL-GEC)*²⁰ is a corpus containing pairs of original and corrected versions of Czech sentences (21 thousand sentences, including 13 thousand doubly annotated sentences, with 20.4% error rate per token), collected from essays written by both non-native learners of Czech and Czech pupils with Romani background. The corpus was built from *CzeSL-man v0*.

The corpus consists of several parts, each including the original text and its correction, aligned sentence-by-sentence or word-by-word. The word-aligned parts differ in whether the corrections is empty, consist of multiple tokens or differ in more than 50% in terms of edit distance. All those parts are split into training, development and testing sets and the annotator 1 and annotator 2 sets.

The *AKCES Grammatical Error Correction Dataset/seeAKCES-GEC (AKCES-GEC)*²¹ extends and supersedes *CzeSL-GEC*. It is generated from a subset of *AKCES*, but apart from the released *CzeSL-man*, *AKCES-GEC* includes additional hand-corrected (but not error-labeled) non-native texts. In comparison to *CzeSL-GEC*, this dataset is twice as large in the number of sentences (47 thousand sentences, 505 thousand tokens, with 21.4% error rate per token) and contains separate edits together with the type annotations in the M2 format.²² The datasets are split into training, development and testing sets.

¹⁹See §6.2, Hana and Hladká (2018a, 2018b), and <https://universaldependencies.org>. The corpus is available from the LINDAT repository <https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2927> and <https://bitbucket.org/czesl/czesl-ud> (Hana and Hladká 2019)

²⁰The corpus is available from <https://hdl.handle.net/11234/1-2143>, see Šebesta et al. (2017).

²¹See Náplava and Straka (2019). The corpus is available from <https://hdl.handle.net/11234/1-3057>, see Šebesta et al. (2019).

²²See Dahlmeier and Ng (2012).

```

S Tam nikoho nebylo celý den jsem byla jedna .
A 1 2|||dep|||nikdo|||REQUIRED|||-NONE-|||0
A 2 3|||agr|||nebyl|||REQUIRED|||-NONE-|||0
A 3 3|||unspec|||,|||REQUIRED|||-NONE-|||0
A 7 8|||lex|||sama|||REQUIRED|||-NONE-|||0

```

Figure 8.2: A sentence from *AKCES-GEC* annotated in the M2 format

(KKOL_A1_003 ru B2)

8.8 *CzeSL in TEITOK*

CzeSL-man, *CzeSL-TH*, *CzeSL-MD* and *CzeSL-UD* will eventually be merged with newly annotated texts into a single corpus with multiple types of annotation.²³ The tool that can accommodate the annotation and manage and search such a corpus is *TEITOK* (see §9.2.3).

Currently, the on-line searchable part of the corpus consists of 2,030 texts (about 300 thousand tokens), selected from the part of *CzeSL-SGT* which has not been hand-annotated yet, i. e., which are included in *CzeSL-man* or *CzeSL-TH*.²⁴ These texts were originally transcribed in the HTML format and later converted, together with the transcription and anonymization markup, into the new XML-based transcription format. Then they were manually error-annotated by successive corrections at several grammar-defined levels. They are also annotated linguistically. See §9.2.3 for details.

In addition to these previously transcribed texts, the corpus also includes 1,300 new transcripts. They will become searchable after they are manually error-annotated in the same way as the other part of the corpus.

All texts included in *CzeSL in TEITOK* are equipped with full metadata in the TEI-like XML format – see Figure 8.3.

8.9 Learner corpora of native Czech

There are several corpora of texts by native Czech learners which were built and made available using similar methods and tools: *AKCES 4*, *SKRIPT 2012* and *SKRIPT 2015*.

²³ *CzeSL in TEITOK* is available for viewing and searching at <http://utkl.ff.cuni.cz/teitok/czesl/>.

²⁴ Content-wise, *CzeSL-MD* and *CzeSL-UD* are subsets of *CzeSL-man*.

```

<teiHeader>
  <profileDesc>
    <particDesc>
      <listPerson>
        <person id="SME_M1" role="learner">
          <age group="16-">22</age>
          <sex type="f"/>
          <langKnowledge>
            <langKnown n="first" group="S" tag="ru"/>
            <langKnown n="native" bilingual="no" tag="-"/>
            <langKnown n="foreign" tag="-"/>
          </langKnowledge>
          <langLearning lang="cs">
            <proficiency assign="current">B2</proficiency>
            <duration months="60-"/>
            <intensity hpw="-3"/>
            <stay country="cz" duration="-12"/>
            <mode>TY|foreign|other</mode>
            <textbook>other</textbook>
          </langLearning>
          <note n="langContact" lang="cs">nobody</note>
        </person>
      </listPerson>
    </particDesc>
    <textDesc>
      <channel mode="manuscript"/>
    </textDesc>
    <taskDesc>
      <task>
        <title>Můj prázdninový zážitek</title>
        <date>2007-10-15</date>
        <taskSetting precedingActivity="-"
          referenceMaterial="yes|dictionary|textbook|other"
          timeLimit="-"
          exam="-"/>
        <extent>200-</extent>
        <topic topicType="general">specified</topic>
        <genre>free</genre>
        <desc genre="narrative" extent="450" extentRange="200-"/>
      </task>
    </taskDesc>
  </profileDesc>
  <notesStmt>
    <note n="docid">SME_M1_001</note>
  </notesStmt>
  <revisionDesc>
    <change when="2019-09-07">XML file created from HTML</change>
  </revisionDesc>
</teiHeader>

```

Figure 8.3: A part of a text header representing metadata in the TEITOK format

AKCES 4 consists of transcripts of hand-written essays written by pupils speaking the Romani ethnolect of Czech, initially collected within the *CzeSL* project.²⁵ The corpus consists of 4,527 texts in the HTML format with transcription markup, altogether 469 thousand tokens. It is not annotated in any way and does not include metadata.

SKRIPT 2012 includes the written language of Czech pupils and students at primary and secondary schools.²⁶ It consists of transcripts of student's written assignments which were produced during their language classes. It contains 1,694 texts, i. e., 709 thousand tokens; it is POS tagged and lemmatized, with metadata about the learner, the school and the text. It is not error-annotated.

SKRIPT 2015 is a balanced mix of essays extracted from *AKCES 4* and *SKRIPT 2012*. The corpus consists of 2,582 texts, i. e., 380 thousand tokens. The authors are pupils of primary and secondary schools of all types, aged 10–15. Metadata and facsimiles of the manuscripts, accessible to approved registered users, are attached. The texts were manually transcribed, anonymized and the writer's corrections marked up. Then they were semi-automatically annotated and revised in the *TEITOK* corpus tool. All texts are manually normalized on multiple levels: spelling and morphematics, morphosyntax and lexicon. The original text and all corrections are tagged and lemmatized, then the type of spelling and morphematic error is automatically identified. Registered users can correct and add texts and annotations in *TEITOK*. The corpus is searchable from LINDAT in the *TEITOK*²⁷ or *KonText*²⁸ environments.

²⁵ *AKCES 4* is downloadable from <https://hdl.handle.net/11858/00-097C-0000-000C-2293-0> (Šebesta et al. 2012). This corpus is searchable as a part of the *CzeSL-plain* corpus. A manually error-annotated subset is searchable as a part of the *CzeSL-man v0* corpus.

²⁶ The corpus is searchable from the CNC *KonText* interface at https://kontext.korpus.cz/first_form?corname=skript2012 (Šebesta et al. 2013) and downloadable as *AKCES 1* from <https://hdl.handle.net/11234/1-1741> (Šebesta et al. 2016).

²⁷ <https://lindat.mff.cuni.cz/services/teitok/skript2015/index.php?action=home>

²⁸ <https://lindat.mff.cuni.cz/services/teitok/skript2015/index.php?action=home>

Chapter 9

Tools

The annotation process described in §7, resulting in the resources discussed in §8, was supported by a number of tools. Other tools are used as corpus search tools to access some releases of the *CzeSL* corpora online. Some of the tools were developed as a part of the *CzeSL* project.

Tools that are commonly used for processing Czech texts produced by native speakers, such as taggers or parsers, are described in §6. Tools mentioned elsewhere in this book but not used in the *CzeSL* project are briefly described in §2. The tools described here include two annotation editors (*feat* and *brat*), an annotation manager (*Speed*) and several corpus search tools (*SeLaQ*, *TrEd*, *Sketch Engine*, *KonText*, and *TEITOK*. *TEITOK* is actually also an annotation editor.

9.1 Annotation tools

9.1.1 *feat*

The manual portion of error annotation in the 2T scheme is supported by *feat*,¹ an annotation tool we have developed. We did not re-use some other annotation tool because none of those available at that time fitted the 2T annotation scheme. Even a tool such as *EXMARaLDA*,² which supports tiered annotation, does not allow for cross-tier links – see §5.4.1.3 for more details.

The annotator corrects the text on appropriate tiers, modifies relations between elements on adjacent tiers (by default all relations are 1:1) and annotates relations

¹See: <https://bitbucket.org/czesl/feat>.

²<https://exmaralda.org/>

with error tags as needed. Figure 7.1 on page 135 shows the tool's user interface. The context of the annotated text is shown both as a transcribed HTML document and – optionally – as a scan of the original document. Both the editor and the data format accommodate various approaches towards the process of multi-tier annotation.

The tool is written in Java on top of the Netbeans platform.³ It automatically synchronizes with *Speed*, our text management system: the user receives (whether an annotator, supervisor or adjudicator) the assigned documents into their Inbox, processes them and moves them to Outbox. For adjudication, two documents are displayed in parallel, differences in their annotation are highlighted and the preferred option can be selected.

9.1.2 *Speed*

To coordinate work of a large project team and to control the passage of texts along the path from the scanned manuscript up to the annotated and adjudicated result, all versions of every document throughout the whole transcription, anonymization and annotation process were stored and maintained in *Speed*, a text management system, developed as a part of the project.⁴

The system distributes documents to transcribers, annotators, coordinators and adjudicators for processing and accepts the results, monitoring their workload and generating error-rate statistics on demand. Using this tool, coordinators could manage the team of 30 annotators efficiently, without wasting their time on administrative tasks.

User privileges are consistently applied both horizontally and vertically. Each user is assigned her views of the data and filters associated with those views. As a result, the annotator is prevented from seeing an interpretation used by a colleague. At the same time, the system is shielded from potential faults and inconsistencies within the users' local file systems.

The system was designed on top of a general workflow machine, intended as reusable for similar applications, and was linked with the off-line annotation tool *feat* using web services. The users could receive their tasks and deliver results without leaving the environment of the application. This included quality checking – through the same channel, the annotator could receive an inadequately annotated text for review with comments by the supervisor.

³<https://platform.netbeans.org/> and <https://netbeans.apache.org/>

⁴*Speed* is available from <https://bitbucket.org/czesl/speed/>. However, it is no longer maintained and there is no support available for its implementation and use.

9.1.3 *brat*

Another annotation editor we use is *brat*.⁵ It is intended for manual annotation of texts, based on a predefined set of tags and relations. A typical use of *brat* could be for annotating named entities, but it can be used for many other purposes, including purely linguistic categories, such as POS or syntactic relations.

A unique feature of *brat* is the notion of “annotation span”. A tag or relation is associated with an interval of positions in the text rather than with individual words or predefined units of the text. The annotation span can then be a single character, a string of characters including spaces, a contiguous sequence of words or parts of words, or even (using symmetric relations joining the components) discontinuous sequences of words. This flexibility is very useful for annotating morphs or errors specific to morphemes, graphemes, phonemes, and also strings consisting of these elements. Annotation spans can overlap, thus multiple errors can be tagged for a single string.

Pre-annotated texts can be imported and the results exported in a transparent format. *brat* can thus be used in a process involving several annotation steps. This is how *brat* is used in the *CzeSL* project. For screenshots of a text annotated in *brat* according to the MD annotation scheme see [Figure 7.4](#) on page 147 and [Figure 7.5](#) on page 148.

9.1.4 *TrEd*

TrEd is a fully customizable and programmable graphical editor and viewer for tree-like structures.⁶ Among other projects, it was used as the main annotation tool for syntactic annotation in *PDT*. We have used *TrEd* together with the *TrEd-ud* extension for annotating universal dependencies in the *CzeSL-UD* corpus.

9.1.5 Error annotation tools

9.1.5.1 Automatic error tagging in 2T

To supplement and facilitate manual annotation, and to provide error tags for texts without manual annotation, we developed a tool for automatic identification and tagging of errors. The tool is focused primarily on formal errors on T1 (see §5.4.5.2). Additionally, some manually assigned tags are specified in more detail.

From the 2T format the tool first extracts the source, T1 and T2 versions of each text, converting them into the tabular (vertical) format – one word per line. The

⁵See <https://brat.nlplab.org/>, <https://github.com/nlplab/brat>, Stenetorp et al. 2012.

⁶See <https://ufal.mff.cuni.cz/tred/> and Pajas 2009.

three data sets are connected via identifiers. Texts on T1 and T2 are provided with morphological annotation (see §6.1). The tool then compares the corresponding source and T1 words, and labels the differences with formal error tags. On T2, the tool uses morphosyntactic annotation to distinguish subcategories of some manually annotated error types.

9.1.5.2 Automatic error detection and tagging in MD

To facilitate manual error annotation in MD, a rule-based tool identifies errors as differences between the source text and the TH and marks them with the most likely error tag. The human annotator can accept or change the tag (see §7.6.2). The tool also performs a simple morphemic analysis of the source words (see §7.6.1).

The tool first extracts the source and the TH text from the *Sketch Engine* learner corpus format into the vertical format (one word per line), performs automatic lemmatization and morphological tagging (using an external tagger), keeping the links between the source and the TH words. Morphemic analysis and error annotation is based on the comparison of the source and the TH words.

The source text with the morphemic analysis and error tags is saved together with the TH text (without any markup) in the *brat* format (a text file and a file with the error tags, with numbers identifying the positions of the errors in the text).

The rule-based error identifier can assist the human annotator or be part of a NLP toolchain. It can also be adapted to provide feedback in an e-learning software.

9.1.6 Conversion tools

These tools help to re-use texts already annotated or transcribed in a different format.

2T scheme → vertical format. This conversion tool is used to format texts annotated in the 2T scheme for search tools based on the vertical format, i. e., *Sketch Engine*, *KonText* or *Corpus Workbench*. It is also used in the toolchain feeding the MD annotation in *brat*.

The format was devised for the *Sketch Engine* tool to cope with learner corpora such as the *Cambridge Learner Corpus* (see §9.2.2). A modified version of the format is used in the *CzeSL-man v2* corpus.

Vertical format → *brat*. This tool uses the vertical format (*Sketch Engine* learner corpus format) as the input to prepare texts for the MD annotation in

brat. Only the T2 target hypothesis is extracted and other error annotation is discarded. The tool is coupled with the MD error identifier (see §9.1.5.2).

Transcription in HTML → XML. This tool converts the HTML-based transcription markup and text format used until recently to transcribe hand-written texts into an XML-based format, used in *TEITOK*.

9.2 Search tools

9.2.1 *SeLaQ*

The only corpus search tool fully compatible with the two-tier error annotation scheme of *CzeSL* is *SeLaQ* (*Second Language Query*).⁷ It is a dedicated web-based tool, developed for the project. The tool is written in Perl on top of the *Dancer* framework⁸ and the PostgreSQL database.⁹

The user can build a query from boxes corresponding to nodes on different tiers. A new box is created by specifying its relation to an existing box (e.g., following/preceding, immediately following/preceding on the same tier, corresponding to a higher/lower tier node), its form, lemma, or tag can be further constrained by a condition (e.g., equal/not equal to, matching a regular expression, same as other box's). If the relation connects two tiers the error type can be also specified.

Figure 9.1 shows a simple query looking for a token at T2 (R2) with the morphosyntactic tag (*Mtag*) specified by a regular expression (\sim) as a cardinal or indefinite numeral (*C[lna]*).¹⁰ At the same time, the edge linking this T2 token to its corresponding counterpart on T1 ($\uparrow 1$ Error:) is labeled by the error code *agr*. The corresponding token at T0 (R0) should match another regular expression, namely the string *dv.**. The user can also specify the context size in the number of words on either side of the keyword(s) (*Velikost kontextu*) and the tier for which the concordance is produced, see Figure 9.2.

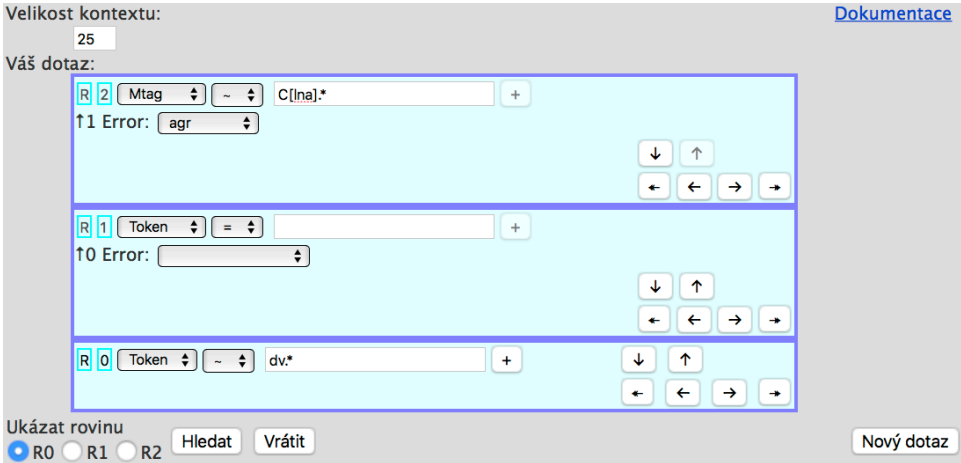
The strength of *SeLaQ* is in its adherence to the 2T scheme. A query is specified and processed in a way that closely follows the annotation and does not miss any of its details. On the other hand, it does not include some features available in more mature standard concordancers and fails to provide a truly user-friendly display of

⁷The tool can be used at <http://utkl.ff.cuni.cz/czesl/selaq.html> to search all texts hand-annotated in the *CzeSL* two-tier annotation scheme. The tool itself is available at <https://bitbucket.org/czesl/selaq>.

⁸<http://www.perldancer.org>

⁹<http://www.postgresql.org>

¹⁰Currently, the user interface is only in Czech.

Figure 9.1: A query in *SeLaQ*

o ještě těžší pro mě . Mám jen jednu sestru . Jmenuje se a je o dvě roky mladší než já . Není jen moje sestra , ale také moje r poskytuje hodně možností : divadlo , opera , koncert . Během dva roků , uviděla jsem víc než dese her , patnáct koncertů a . Nakonec jsem se velmi bavil tím , ale bylo to i moc práci – asi dvě roky jsem prováděl tého organizávání a realizace , však z a vejce , která mají . Když daly mu peníze a vejce , švec udělal dva hnízda různých částech domu , položil tam vejce a příkáz a . Ona bude přiset v Praze příští rok . Eva je moje sestra a má dve sin . Kdy ona ma čas ona libi chodit v parku spolu její ma počívát , protože cesta z mého národního města do Tokia trvá dvě dny autobusem , a nemohla jsem spát v autobusu a v leta bo devět roky zpátky . A samořejmě ona zapoměla . Měli jsme dvě těžké kufry , proto že jsme planovali navštěvu na měsíc . t celý den v firmě . Budu mit Na oběd až 11:30 ´ v jídelně . Za dva hodiny budu mit přestavka 10 minut . Budu končit pracov udje strojní fakulta na zapado česká univerzita bydlíme spolu dvě měsíce . Ačkoli se učim česky sest měsíců , ale je samozř

Figure 9.2: A concordance in *SeLaQ*

results. The most lamentable weakness is the absence of an option to filter texts using their metadata attributes. Another sorely missed feature is the option to display multiple annotation tiers in parallel, as in *feat*.

In an ideal world, the two-tier annotation, the querying options of *SeLaQ*, the features of the annotation editors *brat* (see §9.1.3) and *feat* (see §9.1.1), the statistical and collocation components of *Sketch Engine* and *KonText* (see §9.2.2), and the corpus development, maintenance and display options of *TEITOK* (see §9.2.3) would be combined in a single powerful device.

9.2.2 *Sketch Engine* and *KonText*

*Sketch Engine*¹¹ and *KonText*¹² are both corpus query tools based on the *Manatee* search engine.¹³ In addition to a concordancer, *Sketch Engine* includes various other components, such as *Word Sketch*, a lexical profiling tool. Basic features of *Sketch Engine* are available in *NoSketch Engine*, an open source project.¹⁴

KonText is developed primarily as the interface to corpora created, hosted and maintained by the Institute of the Czech National Corpus.¹⁵ It is used for searching various types of corpora: reference, speech, historical, dialectal or parallel. In its annotation options *KonText* depends on the search engine. Apart from the structural annotation of text elements, such as documents and their properties (text metadata), sections, paragraphs, sentences or tokens, all other annotation concerns individual tokens. Typical token attributes are POS and lemma. In a syntactically annotated corpus, a syntactic function and a pointer to the syntactic head may be added as additional attributes of a token. Syntactic structure can be displayed as a tree for a sentence and sound can be played for time-aligned transcript units in a speech corpus or a graphical representation. Queries can be made using the *Manatee* dialect of CQL or via a simplified user-friendly menu, including a section for restricting the set of queried texts according to metadata specifications. The support for parallel corpora includes parallel multilingual queries and the display of multilingual parallel concordances.

Like *Sketch Engine*, *KonText* offers a number of options for presenting the search results. In addition to concordances, various statistics can be produced. Typical collocations can be computed using several collocation measures.

¹¹*Sketch Engine* (<https://www.sketchengine.eu>, Kilgarriff et al. 2014) is a commercial product, developed and maintained as a web service by Lexical Computing, a research company founded in 2003 by Adam Kilgarriff. Until 1 April 2022, *Sketch Engine* is available at no cost for non-profit use to academic institutions within the European Union, as a part of the *ELEXIS* project (<https://elex.is>).

¹²<https://kontext.korpus.cz/>; Machálek 2017

¹³*Manatee* was developed by Pavel Rychlý as a part of *Bonito* (Rychlý 2007), an alternative to *CWB*, another corpus query tool (Christ 1994; Christ et al. 1999).

¹⁴<https://www.sketchengine.eu/nosketch-engine/>

¹⁵A production version of *KonText* has been available at <https://kontext.korpus.cz> since 2014, replacing *NoSketch Engine*. *KonText* was adopted as the search tool for the LINDAT/CLARIN repository (<https://lindat.mff.cuni.cz/services/kontext/corpora/corplist>) and is also used by other institutions, such as Clarin-PL or the Sorbian Institute in Bautzen, Germany. *KonText* can be used to search all accessible corpora at no cost for academic and educational purposes and is also available as an open source project at <https://github.com/czcorpus/kontext>.

9.2.2.1 Token-based error annotation

Both *Sketch Engine* and *KonText* can be used straightforwardly for querying a learner corpus, as long as its linguistic and error annotation can be represented as attributes of individual tokens. This is how *CzeSL-plain*, *CzeSL-SGT* and *CzeSL-man v1 searchable* are annotated (see, e.g., [Figure 8.1](#) on p. 162).

A simple query into the *CzeSL-SGT* corpus using *KonText* can be made from https://kontext.korpus.cz/first_form?corpname=czesl-sgt. With the Query Type set to **Basic**, a string entered in the Query field returns sentences where the form occurs in the original, uncorrected text. For more advanced queries, including references to tags, lemmas, error types, corrected forms and metadata attributes, the Query Type should be set to CQL¹⁶ and/or the settings in Restrict search should be modified.

In addition to the attributes listed for *CzeSL-SGT* in [Table 8.2](#), the *KonText* search interface offers “dynamic” attributes, derived from some positions of **tag** and **tag1**. Dynamic attributes can be used in queries to specify values of morphological categories without regular expressions, to stipulate identity of these values in two or more forms to require grammatical concord, or to compare values of a category for **word** and **word1**. Dynamic attributes available for the source and the corrected form are listed in [Figure 9.3](#). They are meant especially for CQL queries including a “global condition”. As in standard corpora, such queries target two or more word tokens with an arbitrary but equal value of an attribute such as morphological case to express grammatical agreement and similar morphosyntactic phenomena ([37](#)).

(37) 1: [] 2: [] & 1.c = 2.c

In a learner corpus, such queries make sense even for a single word token, e. g., for expressing identical or distinct values of the morphological case of the original form and of its corrected version ([38](#)).

(38) 1: [] & 1.c != 1.c1

The slightly more complex query shown in [Figure 9.4](#) makes sure that we target a token with an error in grammar (**[gs="G"]**) whose word class (**k**) and subcategory (**s**) stay unchanged even after the correction (**k1**, **s1**). Additionally, the CEFR levels of the learners are restricted to B1 and higher, which reduces the size of the texts to be searched to 379 thousand tokens (the entire *CzeSL-SGT* has 1,147 thousand tokens).

¹⁶For general help on CQL see https://wiki.korpus.cz/doku.php/en:pojmy:dotazovaci_jazyk.

- k, k1** word class (position 1 of the tag)
- s, s1** detailed word class (position 2 of the tag)
- g, g1** gender (position 3 of the tag)
- n, n1** number (position 4 of the tag)
- c, c1** case (position 5 of the tag)
- p, p1** person (position 8 of the tag)

Figure 9.3: Dynamic attributes, derived from morphological tags in *CzeSL-SGT*

Search in the corpus

Corpus: / ★

Query type: ⓘ

[Insert tag](#) | [Insert within](#) | [Keyboard](#) | [Recent queries](#)

Query: `1:[gs="G"] & 1.c != 1.c1 & 1.k = 1.k1 & 1.s = 1.s1`

You can use the "down arrow" key to view recent queries ▶

Default attribute:

▶ **Specify context**

▼ **Restrict search**

1 **doc.s_cj_SERR** ∈ {B1, B2, C1, C2}
378,688 positions

Figure 9.4: A query in *KonText* into the *CzeSL-SGT* corpus

Some results are shown in Figure 9.5. The highlighted keywords are followed by values of some of the token attributes: the morphological case of the source form and its correction,¹⁷ followed by the formal error label (see Table 5.8 and Table 5.9 on page 106–107) and the correction (provided by *Korektor*). Two selected metadata items are shown in the first column: L1 and CEFR level.

¹⁷Morphological cases are encoded as numbers: 1 stands for nominative, 2 for genitive, 3 for dative, 4 for accusative, 5 for vocative, 6 for local and 7 for instrumental.

Hits: 1,301 | l.p.m.: 1,133.79 (related to the whole corpus) | ARF: 685.27 | Result is shuffled 1 / 66 ▶▶▶

Line selection: simple +

<input type="checkbox"/>	uk →B2	Chodit do školy v naši /4/6/Quant0/naši době je povinnost pro každého .
<input type="checkbox"/>	ru →B1	Díky globalizaci , není to dnes problem , jet z České republiky do Francie na par /2/4/Quant0/pár dnů a zpátky .
<input type="checkbox"/>	ru →B2	Mým rodičům nic netřeba , a pro mě si vezmi neco /1/4/Caron0/něco k jídlu a taky hodně .
<input type="checkbox"/>	ru →B2	A když jsou daleko , tak to je moc těžké , protože se rodiče pořad strachují za svých dětí /1/2/Quant0/dětí , protože nevědí jak bydlí dětí , co dělají , kdy jedí , co jedí , kdy spí a dobře spí nebo není dobře .
<input type="checkbox"/>	ko →B1	Ve třináct hodin jsem se naobědvala v nějaké dobré restaurací /2/6/Quant1/restauraci .
<input type="checkbox"/>	it →B2	Říkám to , protože skoro nikdo neví jistě jestli bude nebo nebude ekologická katastrofa kvůli neekologickém /6/3/SingCh/neekologické chování .
<input type="checkbox"/>	ja →C1	Když navštívujeme supermarket nebo hypermarket , vidíme vždy řadu zubních past nebo šampónů , jejichž /7/1/Quant1/jejichž roždily vůbec rozpoznáme .
<input type="checkbox"/>	ru →B1	To je moc pekný a veselý svatek /2/1/Quant0/svátek .
<input type="checkbox"/>	ja →C1	Ale v současnosti více případů ve světe /5/6/Caron0/světě se souvisí s našim životem a určitě nemůže nám chybět média jako zdroje .
<input type="checkbox"/>	kk →B1	Taky muzeme jít do některých muzei /7/2/Quant0/muzei .
<input type="checkbox"/>	de →B2	Když návštěvník například jenom chce prohlédnout Vídeň tří /2/1/Quant1/tři nebo čtyři dny jsou dost .
<input type="checkbox"/>	lv →B2	Myslím , že to není normalní situace , obchody nemají co vydat spatky protože to musí být jejich /2/X/Quant1/jejich biznes a musí být připravení .
<input type="checkbox"/>	ru →B2	Když jsem přišel domu Tam byli moji /3/1/Quant1/moji rodiči a nekolik kamaradi , byli překvapeni a zeptali se " Co jsi dneska dělal ?
<input type="checkbox"/>	ru →B1	Koupil jsem v obchode různé chemický /1/2/SingCh/chemické komponenty , smichal je a pak jsem založil pod domem a zapalil .

Figure 9.5: Concordance in *KonText*, showing a partial response to the query in Figure 9.4

To see the frequency distribution of the source and target morphological cases, *KonText* can generate a table shown in Figure 9.6. Accusative seems to be the most frequent case that the learners failed to use, using nominative and genitive instead.¹⁸

Finally, Figure 9.7 shows the top 20 lemmas of forms where the learners erred most often in morphological case.¹⁹

¹⁸Here it is important to remember that the annotation of the source forms could be misleading for at least two reasons: (i) the tagger works less reliably on an incorrect text and (ii) the reason why the source form is not correct need not be due to the learner's decision to choose a wrong morphological case.

¹⁹The prepositions *s* 'with' and *z* 'from' are listed because their tag includes the case of the prepositional object. The caveat concerning the tags for the source forms applies here as well: the most frequent lemma *člověk* 'man' is often used in the suppletive plural forms *lidé* 'people.NOM', *lidi* 'people.ACC/NOM.COLL', *lidí* 'people.GEN', and also as a form of a different lemma *lid* – *lide* 'a people, nation.VOC', which can all be used incorrectly due to an error in spelling or morphonology rather than in morphosyntax.

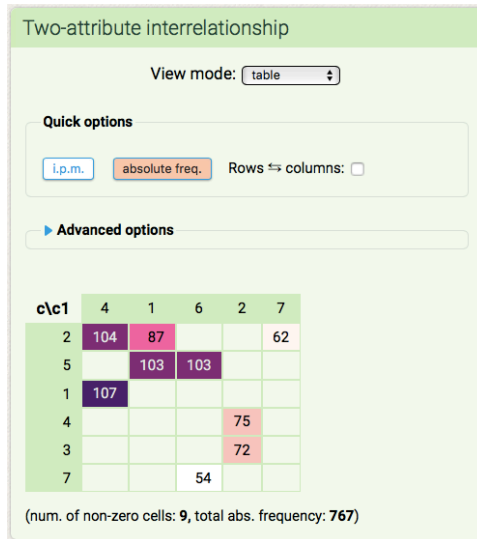


Figure 9.6: Frequencies of incorrect and corrected morphological cases for the concordance in Figure 9.4, generated by *KonText*

9.2.2.2 Error annotation using structures

To overcome the word boundary restriction, error annotation can be encoded as structural annotation in the token-based tabular (vertical) format. Once the corpus includes the appropriate structural elements, *Sketch Engine* offers a dedicated learner corpus search interface.²⁰

For an example of this annotation see Figure 9.8. There are just two tokens in the example, incorrectly split and misspelled: *při poměl* → *přípomněl* ‘(he) reminded’. In an appropriate context, the separated verbal prefix *při* can be interpreted as a preposition ‘next to’.

Linguistic annotation is represented in the horizontal dimension as columns of an imaginary table (i.e., as attributes of the corresponding token), while error annotation is represented in the vertical dimension as structural elements. The error annotation, following the *CzeSL* two-tier annotation scheme, is encoded as

²⁰See <https://www.sketchengine.eu/documentation/setting-up-learner-corpus/>. This type of error annotation was used in the *Cambridge Learner Corpus* – see https://www.cambridge.org/elt/corpus/learner_corpus2.htm, <https://www.cambridge.org/sketch/help/>.

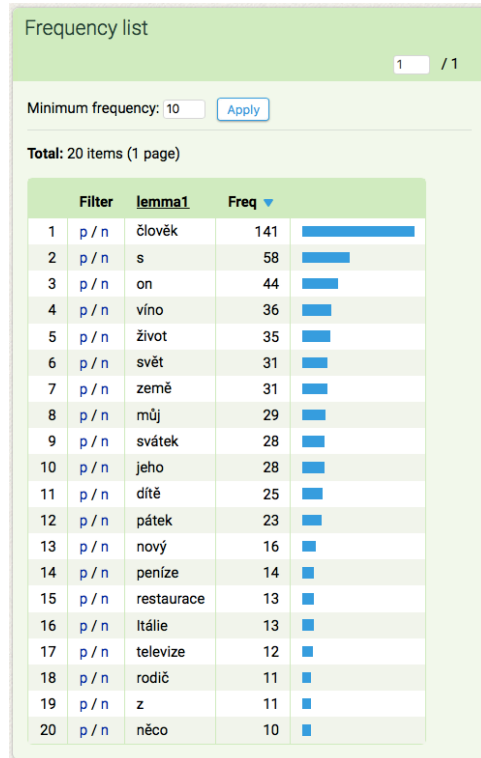


Figure 9.7: Frequencies of forms listed by their lemmas for the concordance in Figure 9.4, generated by *KonText*

structural elements **err** and **corr**.

The **err** element includes incorrect forms, with each token in a separate row of the imaginary table. The first column includes the form itself and the two following columns represent the token attributes: the positional morphosyntactic tag and the lemma.

The **err** element is immediately followed by a corresponding **corr** element with the same error **type** specification. The **corr** structure includes the forms within the preceding **err** structure after correction.

Structures can be embedded. In this example, the word boundary is removed in the embedded **err/corr** structures (`<err tier="1" type="wbdSplit">`) and


```

<err tier="1" type="incorBase">
<err tier="1" type="wbdSplit">
při      RR--6-----      při
poměl    VpYS---XR-AA---    pomít
</err>
<corr tier="1" type="wbdSplit">
připoměl X@-----      připoměl
</corr>
</err>
<corr tier="1" type="incorBase">
připomněl VpYS---XR-AA---    připomenout
</corr>

```

Figure 9.8: A wrongly split word form annotated by **err** and **corr** structures

then the spelling error in the joined form *připoměl* is corrected in the outermost **err/corr** structures (`<err tier="1" type="incorBase">`).

```

<err tier="1" type="wbdJoin">
přečístse X@-----
</err>
<corr tier="1" type="wbdJoin">
přečíst   Vf-----A----    přečíst
<err tier="2" type="lex">
se        P7-X4-----      se
</err>
<corr tier="2" type="lex">
si        P7-X3-----      se
</corr>
</corr>

```

Figure 9.9: A wrongly joined word form annotated by **err** and **corr** structures

Figure 9.9 shows a wrongly joined word and misspelled form *přečístse* → *přečíst si* ‘(to) read (for oneself)’. The incorrectly joined form *přečístse* is annotated as an error in word boundary by the `<err tier="1" type="wbdJoin">` structure and corrected within the immediately following `<corr tier="1" type="wbdJoin">` structure as two tokens: *přečíst* and *se*. While *přečíst* is already a correct form, the incorrect form of the reflexive particle *se* requires a subsequent correction at T2. This is why the `<corr tier="1" type="wbdJoin">` structure includes an additional embedded pair of **err** and **corr** structures. The form *se* is annotated as a lexical error by the `<err tier="2" type="lex">` structure and corrected as *si* within the im-

mediately following `<corr tier="2" type="lex">` structure. The result is a string consisting of *přečíst*, corrected in a single step already on T1, and *si*, detached in an intermediate step on T1 from *přečístse* as a separate token *se*.

The forms shown in Figure 9.8 and 9.9 are displayed in Figure 9.10 as concordances in a context. The concordances are produced by the *Sketch Engine* search tool following a CQL query `<err type="wbdJoin|wbdSplit"/>`. The query looks up all `err` structures with the `type` attribute `wbdJoin` or `wbdSplit`, i. e., any wrongly joined or split forms. The display of the structural elements used for error annotation is customizable. The notation in Figure 9.10 uses brackets to show the tier and the error type (immediately following the left bracket), the incorrect form(s) and the correction (following the `>` sign). Embedded `err` and `corr` structures are displayed within embedded brackets. Apart from a different graphical design, *KonText* displays the same result.

The screenshot shows the Sketch Engine search interface. At the top, a search bar contains the query `cql <err type="wbdJoin|wbdSplit"/>` and shows 2 results (134.31 per million). Below the search bar is a toolbar with various icons. The main area displays two concordance results:

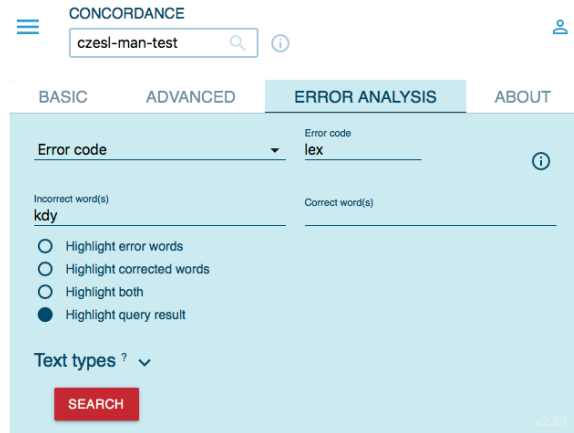
1. `XX_XX_001`
 Petr mi to včera [1:incorBase [1:wbdSplit při poměl > připoměl]> připomněl] .
při/RR-6----- pomit VpYS--XR-AA--/

2. `XX_XX_001`
 Jana už nestihla [1:wbdJoin přečístse > přečíst [2:lex se > si]] tu knížku celou .
přečístse/X@-----

Figure 9.10: Concordance in *Sketch Engine* showing word boundary errors

In addition to CQL, a learner corpus annotated by the `err` and `corr` structures can also be queried from a dedicated Error query (ERROR ANALYSIS) interface, shown in Figure 9.11. The interface is very intuitive: the user chooses an error type and a token or a string of tokens annotated as an error (Incorrect word(s)) and/or a correction (Correct word(s)). Tokens can also be specified using wildcard characters, as in `simple` query type, e. g., an asterisk stands for any string. For a query using only the error type, at least an asterisk is needed in the Incorrect word(s) or Correct word(s) field in addition to an error type in the Error code field. Figure 9.11 shows the result of the query.

Figure 9.13 shows the result of a CQL query combining linguistic and error annotation. A query in (39) searches for adjectives with an error in the ending due to incorrect morphosyntactic agreement. A query combining an error type with a

Figure 9.11: Error query in *Sketch Engine*Figure 9.12: Concordance in *Sketch Engine*: lexical errors in *kdy* ‘when’

POS specification using the morphosyntactic tag cannot be specified in the ERROR ANALYSIS interface.

(39) `[tag="A.*"] within <err type="agr"/>`

Other than error-annotated parts of the texts can be searched in the usual way, using all the query types, except for the Error query (ERROR ANALYSIS). To search for a sequence of tokens which might or might not be error-annotated, the query should include optional structural elements. The query in (40) looks up sequences of a verb followed by an adjective and a noun in the accusative case in the source text. The query matches the source text, both correct and incorrect. This is because the optional `corr` structures follow tokens unspecified for whether they are or are not embedded within a structure. If they are embedded within a structure, it must be an `err` structure. Thus the tokens are either incorrect or are not error-annotated.

Figure 9.13: Concordance in Sketch Engine: agreement errors in adjectives

(40) `[tag="V.*"] (<corr/>)? [tag="A...4.*"] (<corr/>)? [tag="N...4.*"] (<corr/>)?`

The query in (41) searches for the same sequence. However, if any tokens in the sequence are corrected, the query matches corrections rather than incorrect tokens. This is because the optional `err` structures precede tokens unspecified for whether they are or are not embedded within a structure. If they are embedded within a structure, it must be a `corr` structure. Thus the tokens either represent corrections or are not error-annotated.

(41) `(<err/>)? [tag="V.*"] (<err/>)? [tag="A...4.*"] (<err/>)? [tag="N...4.*"]`

For a query involving morphosyntactic features, it makes better sense to target corrections rather than incorrect forms. This is due to the high share of non-

words among incorrect forms, which cannot be matched by a query targeting the original text. The concordance shown in Figure 9.14, produced by *KonText*, can only be found using the query in (41) targeting corrections.²¹ However, for errors in morphosyntax – where corrections involve modifying an existing adjectival form to reflect NP-internal agreement in its case, number and gender ending – it could still be useful to use a query targeting incorrect forms.

The Error query/ERROR ANALYSIS interface of *Sketch Engine* inserts the optional **err/corr** structures into sequences of tokens specified as queries in the Incorrect or Corrected word(s) fields.

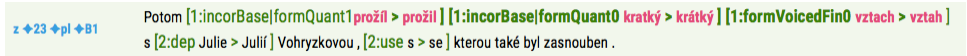


Figure 9.14: Concordance in *KonText*: a sample result of the query in (41)
(HRD_2E_225 p1 B1)

This type of error annotation, combining the token-based vertical format with the **err/corr** structures, is compatible with a large part of the two-tier *CzeSL* annotation scheme, namely with error annotation concerning contiguous sequences of tokens. For discontinuous word order errors and links from the incorrect form to a word “explicating” the error, additional constructs must be introduced, such as XML pointers and anchors.

9.2.3 TEITOK

Learner texts are often annotated in ways which are not readily compatible with standard corpus tools and annotation formats. Most of the time, it is because linguistic annotation is complemented by error annotation. Depending on the annotated text and goals of the error annotation, an error type and a target hypothesis may need to be specified even in cases where words in the original and the normalized text do not correspond 1:1 – the original words can be split, joined or reordered. Moreover, there can be multiple target hypotheses, successive or alternative, for a single stretch of text. It may also be useful to provide linguistic annotation for each of the hypotheses.²²

²¹In addition to error annotation, the view options of the search tool can be set to show some metadata of the text including the concordance. The leftmost string represents four metadata items: **z** stands for the gender of the author (female), **23** for her age, **pl** for her first language (Polish) and **B1** for her proficiency level in Czech.

²²For an overview of infrastructure issues and some proposals concerning learner corpora see, e.g., Stemle et al. (2019).

The need for annotation of such complexity is not specific to learner corpora. In fact, the Text Encoding Initiative (TEI)²³ provides guidelines for digitizing many features of texts such as medieval manuscripts, resulting in truly complex annotation. On the other hand, adherence to the TEI standard does not guarantee trouble-free use of the digitized text. Most corpus search tools index tokens as the smallest text unit and assume that linguistic or error annotation is represented as attributes of individual tokens. It is a challenging task to comply with this restriction and translate a complex TEI-compliant annotation, including annotation spanning multiple tokens, into annotation restricted to individual tokens.

*TEITOK*²⁴ is meant to bridge the gap between texts annotated in the TEI way on the one hand and the need to provide efficient access to the texts on the other. The access is similar to that provided by standard corpus search tools while preserving the options to search and view all properties represented in the annotation. The tool also allows for adding or editing texts, their annotation and metadata, and updating the searchable corpus accordingly. *TEITOK* is in fact a single environment for building, maintaining and using a corpus, suitable particularly for specialist corpora for which complex annotation or continuous development is useful or needed. Moreover, as a web-based tool it is easy to maintain and customize for individual projects in a collaborative setting.

The search module of *TEITOK* is based on the widely used *Corpus Query Processor (CQP)*, the main component of the *IMS Open Corpus Workbench (CWB)*,²⁵ and employs the same data format and structure. This is why the indexed corpus can be shared by *TEITOK* and other CQP-compatible corpus search tools such as *Manatee* with the *Sketch Engine* or *KonText* user interface.²⁶ Several corpora can now be searched on-line using *TEITOK* or *KonText* while being extendable and editable using *TEITOK*, e. g., *SKRIPT 2015*, the learner corpus of native speakers of Czech (Janssen 2020).²⁷

There are several other learner corpora available in *TEITOK*: the Learner Corpus of Portuguese as Second/Foreign Language (*COPLE2*)²⁸ with successive normalization levels instead of error tags, the *Croatian Learner Text Corpus (CroLTeC)*²⁹ with both successive normalization and error codes, the corpus of Baltic interlan-

²³<https://tei-c.org>

²⁴Janssen (2016, 2018), <http://www.teitok.org>

²⁵Christ (1994) <http://cwb.sourceforge.net>

²⁶Rychlý (2007) and Kilgariff et al. (2014), <https://nlp.fi.muni.cz/trac/noske>, <https://github.com/czcorpus/kontext/>

²⁷<https://lindat.mff.cuni.cz/services/teitok/skript2015/>

²⁸<http://teitok.clul.ul.pt/cople2/>, Mendes et al. (2016), Rio et al. (2016), and Rio and Mendes (2019)

²⁹<http://nlp.ffzg.hr/resources/corpora/croltec/>, Preradović, Berać, and Boras (2015)

guage with error tags (Znotina 2017), and the *CzeSL* corpus itself.³⁰ The main *TEITOK* site shows a list of other projects using *TEITOK* (spoken, historic, developmental, i. e., L1 learner corpora).³¹

TEITOK is in fact a graphical user interface used to create, visualize and edit TEI/XML files, and to search a corpus built from such files. Each text is represented as an XML file with a header, including all metadata, and the text itself, annotated with a potentially very rich combination of any predefined textual, linguistic and error markup. The elementary text unit is a token, an XML element with annotation related to that token represented as its attributes. The token serves as the link between the TEI annotation and the format required by the search engine. A single corpus can include written and spoken parts, searchable with a single query. This may be useful for comparing written and spoken language of specific learners in a learner corpus.

The graphical user interface can be used to customize many properties of the corpus and its search functionality, of the texts and their annotation, and of the user interface itself. Some typical attributes of a transcribed word (a token) in a learner corpus built from hand-written documents can be the word's written form (after any corrections made by the author), its normalized (corrected) form, the POS tag and lemma of the normalized form and an error tag.

Figure 9.15 shows the transcript and the facsimile of a manuscript in the Text View interface. The text (glossed in Table 5.1) is included in *CzeSL-man*. Its metadata are accessible and editable (by privileged users) after a click on edit header data. Some properties of the hand-written text are preserved in the transcription markup. Line breaks are retained as such, other properties are represented as colored text: deletions are in strikethrough red and additions in blue.

The text in Figure 9.16 is shown after all corrections, based on the *TEITOK* settings specified for the *CzeSL* corpus. The color of a word form represents a type of correction, in fact its level in a sequence starting from the transcribed form up to the level of Subsequent correction.

- Transcription (black) – the transcribed manuscript word form
- Written form (blue) – the word form after resolving the transcription markup (if different from transcription)

³⁰<http://utkl.ff.cuni.cz/teitok/czesl/> – work in progress at the time of writing

³¹<http://www.teitok.org/index.php?action=projects>

NEM_GD_008.xml

Bratr a sestra

- edit header data · view telHeader

View options

Text: [Transcription](#) | [Written form](#) | [Normalized form](#) | [Orthographic correction](#) | [Morphosyntactic correction](#) | [Lexical correction](#) | [Subsequent correction](#) | Show: [Colors](#) | [Formatting](#) | [<pb>](#) | [<lb>](#)
 Images | Tags: [POS tag](#) | [Lemma](#) | [POS tag after all corrections](#) | [Lemma after all corrections](#) | [UD POS tag](#) | [UD features](#) | [Error tag](#)

Edit the information about each word of this file by clicking on the word in the text below, or click [here](#) to edit the raw XML

Bratr a Sestra.

Viktor je mladý pan z **PolskaRuska**. Studuje **češtinu** ve škole, protože ne umí psát a číst **spravně**. Bydlí na koleje vedle školy, má jednu sestru Irenu, která se učí na univerzitě u profesora Smutneveselého. Bohužel, Viktor není dobrý student, protože spí na lekci, ale jeho sestra **píše-všechno všechno píše** a výborně rozumí českého profesora Smutneveselého a **brzo dělá domácí úkol**. Večere Irena jde na prochásku spolu z kamarádem, ale její bratr dělá nic. Jeho čeština je špatná, vím, že se vrátit ve **PolskoRusku** a tam budí studovat u pomalu myt podlahy.

Kamarad Ireny je američan a chytrý muž. On miluje Irenu a chce se vzít na ní, protože ona je hezká, taky chytra, rozumí ho a umí výborně vařit.

Kdo neumí nic a nechce studovat je bloubec. **buďi** Bohužel, bloubec je Viktor. Ty bratr a sestra jsou moc různ~~ye~~**ye**.

To je všechno.

Konec

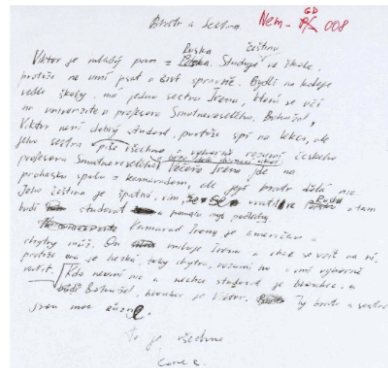


Figure 9.15: Transcription and facsimile view of a text in TEITOK

(NEM_GD_008 ru B2)

- Normalized form (brown) – correction proposed by an automatic tool,³² e.g., *spravně* → *správně* ‘correctly’
- Orthographic correction (magenta) – manual correction of spelling and morphemics (if different from Normalized form), e.g., *prohasku* → *procházku* ‘a walk.ACC’
- Morphosyntactic correction (red) – manual correction, e.g., *mu* ‘he.DAT’ → *ho* ‘he.ACC’

³²Currently, it is *Korektor* as a web service, see <https://ufal.mff.cuni.cz/korektor>; Richter, Straňák, and Rosen (2012).

- Lexical correction (green) – manual correction, e. g., *lekci* → *hodině* ‘class.LOC’
- Subsequent correction (cyan) – manual correction of a form correct in the original text but turned incorrect due to corrections in the context, usually because of morphosyntactic agreement, e. g., *českého* ‘Czech..ACC’ (*profesora* ‘professor.ACC’) → *českému* ‘Czech.DAT’ (*profesorovi* ‘professor.DAT’)

Following the setting of inheritance for the correction levels in this corpus, the annotator does not have to specify forms for all the correction levels. As a result, the order of levels determines which form is assumed for a specific level if the corrected form specification is missing. The missing value is then assumed to be the same as the value specified by the annotator or by the default rule for the immediately preceding level.

Depending on the View options, the color shows only the last correction level up to the level selected in the View options. So while the wrongly spelled *vrátit* is first corrected as *vrátit* ‘to return’, and then the infinitival form is changed to a finite form as a morphosyntactic correction *vrátí* ‘(he) returns’, the color shown is that corresponding to the morphosyntactic correction (red).³³

All token-based annotation is available on mouse-over. The error annotation is added mostly by a human annotator using *TEITOK*, while all linguistic annotation is provided by automatic tools. The tools can be run by the corpus administrator in batch mode for the whole corpus or launched by the annotator for a specific text from the interface. The interface can also be used to revise the automatic annotation.

Figure 9.17 shows the *TEITOK* Edit Token window for the word form *českeho* → *českého* ‘Czech.ACC’ → *českému* ‘Czech.DAT’. The correction tool guessed the Normalized form correctly, so the annotator does not have to make the orthographic correction. On the other hand, the annotator has to make the subsequent correction. The use of Rectified non-standard form is restricted to cases when a colloquial form is misspelled, e. g., as *dobrey* instead of *dobrej*, which is supposed to be *dobrý* ‘good’ in Standard Czech. Only the latter form is used for the orthographic correction, while the correct but colloquial *dobrej* ends up as the Rectified non-standard form.

The rest of the items available in the Edit Token window represent tags and lemmas rather than corrected forms. All of them are specified by automatic tools,

³³The correction levels and their colors are defined in the setting of the specific corpus. For another corpus a single or no correction level at all can be defined. Instead, a rich taxonomy of error types can be introduced, e. g., corresponding to error domains – spelling, morphemics, morphology, morphosyntax, lexicon, etc. A rich error taxonomy, designed by Ibrahim Mansour, is used in the *Corpus of Arabic learners of Czech* (<http://utkl.ff.cuni.cz/teitok/ima-lc/>).

Bratr a sestra

Viktor je mladý muž z Ruska. Studuje češtinu ve škole, protože neumí psát a číst správně. Bydlí na koleji vedle školy, má jednu sestru Irenu, která se učí na univerzitě u profesora Smutněveselého. Bohužel, Viktor není dobrý student, protože spí na hodině, ale jeho sestra všechno si zapisuje a výborně rozumí českému profesorovi Smutněveselému a rychle d

Kamarád Ireny je Američan a chytrý muž. On miluje protože ona je hezká, taky chytrá, rozumí mu a u

Kdo neumí nic a nechce studovat, je blbec. Bohužel sestra jsou hodně různí.

To je všechno.

Konec

Normalized form	piše
Lexical correction	zapisuje
POS tag	Unrecognized (X@-----)
Lemma	piše
POS tag after all corrections	Verb (VB-S---3P-AA---) Finite; Singular; 3rd person; F
Lemma after all corrections	zapisovat
UD POS tag	VERB
UD features	Aspect=Imp Mood=Ind Num
Error tag	formQuanto

Figure 9.16: A view of the corrected text in TEITOK (NEM_GD_008 ru B2)

including Error tag, the only error annotation item. POS tag and Lemma refer to the original text (actually to Written form), while POS tag after all corrections and Lemma after all corrections as well as UD POS tag and UD features refer to the fully corrected text.³⁴

The correction levels presented here correspond to the two-tier error annotation scheme used in the *feat* annotation editor (see §9.1.1). Levels up to Orthographic correction are annotated at T1 while levels from Morphosyntactic correction onward are annotated at T2.

Not all annotation is visible in *TEITOK*'s text view. In addition to the token-specific annotation, visible on mouse-over and editable in Edit Token window (Fig-

³⁴The first four items are identified using the *MorphoDiTa* web service at <https://lindat.mff.cuni.cz/services/morphodita/api-reference.php> (Straková, Straka, and Hajič 2014) with the default language model. For the positional tags *TEITOK* offers a tag builder. The *UD* categories are supplied by the *UDPipe* web service. The Error tag item is determined by a tool comparing Written form and Orthographic correction – see §5.4.5.2 and Jelínek (2017).

Edit Token

Filename	NEM_GD_008.xml	
Title	Bratr a sestra	

Token value (w-66): českeho

pform	Transcription (Inner XML)	českeho
form	Written form	
nform	Normalized form	českého
dform	Rectified non-standard form	
ort	Orthographic correction	
gram	Morphosyntactic correction	
lex	Lexical correction	
subs	Subsequent correction	českému

pos	POS tag	X@-----	tag builder
lemma	Lemma	česke	
spos	POS tag after all corrections	AAMS3----1A----	tag builder
slemma	Lemma after all corrections	český	
upos	UD POS tag	ADJ	
feats	UD features	Animacy=Anim Case=Acc Degree=Pos Gender=Masc Number=Sing Polarity=Po	
err	Error tag	formQuant0	

insert tok after: **attached / separate** • before: **attached / separate** • insert elm before: **paragraph ; linebreak** •
split in dtoks: **2 ; 3**
edit context XML • merge left to w-65 • create mtok left: **1 ; 2**
treat similar tokens

Bohužel, Viktor není dobrý student, protože spí na lekci, ale jeho sestra **piše všechno všechno piše** a vyborně rozumí **českeho** profesora Smutneveseleho **a brzo delá domácí ukol.**

Save Cancel Token Details

Figure 9.17: Edit Token window of TEITOK

(NEM_GD_008 ru B2)

ure 9.17), there can also be annotation spanning multiple (even discontinuous) tokens, useful for correcting word-order, multi-word units and constructions. This kind of annotation is visible and editable in the Stand-off error annotation window (Figure 9.18). Stand-off annotation is stored in a separate file.

The underlined sequences in the text can be assigned an error code and reordered or replaced by a different text. Figure 9.18 shows five such errors, three concerning word order, two concerning restructuring corrections. The correction shown in the

mouse-over box replaces a grammatically correct but stylistically clumsy sequence *a rychle dělá domácí úkoly* ‘and quickly makes homeworks’ by *a domácí úkoly má rychle hotové* ‘and homeworks has quickly finished’. Note that some of the words in the text have already been corrected as tokens.

Stand-off error annotation

Bratr a sestra

Text: [Transcription](#) [Written form](#) [Normalized form](#) [Orthographic correction](#) [Morphosyntactic correction](#)
[Lexical correction](#) [Subsequent correction](#)

Bratr a sestra

Viktor je mladý muž z Ruska. Studuje češtinu ve škole, protože neumí psát a číst správně. Bydlí na koleji vedle školy, má jednu sestru Irenu, která se učí na univerzitě u profesora Smutněveselého. Bohužel, Viktor není dobrý student, protože spí na hodině, ale jeho sestra všechno si zapisuje a výborně rozumí českému profesorovi Smutněveselému a rychle dělá domácí úkoly. Večer Irena jde na procházku spolu s kamarádem, ale její bratr nedělá nič. Jeho čeština je špatná, vím, že se vrátí do Ruska a tam bude studovat a rychle dělá domácí úkoly pomalu mýt podlahy.

Kamarád Ireny je Američan a chytrý muž. On miluje Irenu a chce si ji protože ona je hezká, taky chytrá, rozumí mu a umí výborně vařit.

Kdo neumí nic a nechce studovat, je blbec. Bohužel, blbec je Viktor. Ten bratr a sestra jsou hodně různí.

To je všechno.

Konec

Annotations

Error code

WO CONSTR

a rychle dělá domácí úkoly

- jsou různí
- na hodině spí
- jde Irena
- si všechno

[show as list](#) • [edit raw XML file](#)

Error code	Construction (CONSTR)
Corrected forms	domácí úkoly má rychle hotové

Figure 9.18: Word-order and restructuring corrections as standoff annotation in *TEITOK* (NEM_GD_008 ru B2)

If available, other types of annotation can be displayed for tokens or sentences. Figure 9.19 shows dependency syntactic structure and functions for a sentence from the text above in a linearized tree view (another option is the standard tree with the root at the top). Dependency relations and function labels can be modified using the interface. *TEITOK* can also provide access to the *CzeSL* two-tier error annotation in *feat* (see §9.1.1) and the multidimensional annotation (see §5.6).

Figure 9.20 shows the sentence from Figure 9.19 in the *TEITOK*-internal XML format, with each word annotated as a token (`tok`).³⁵ In addition to lemma, syntactic

³⁵Some attributes in the `tok` (token) elements are omitted for space reasons, namely the POS

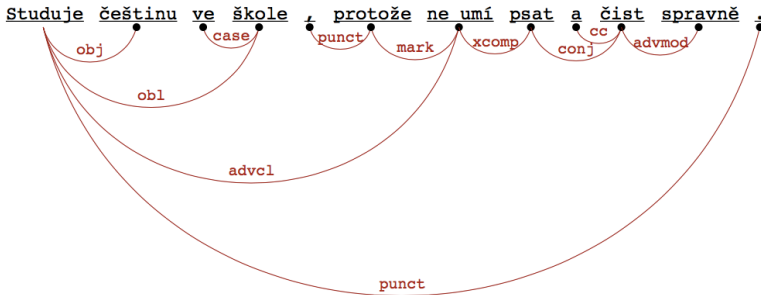


Figure 9.19: Linearized text tree view of a short sentence parsed by *UDPipe* (NEM_GD_008 ru B2)

function (`deprel`), UD POS tag (`upos`) and pointer to its syntactic governor (`head`), a corrected token can also include attributes such as Written form (`form`), Normalized form (`nform`), Orthographic correction (`ort`), Lemma after all corrections (`slemma`) or Error tag (`err`).

The `mtok` element is used to join two incorrectly split word forms. The `mtok` element from Figure 9.20 is shown again in (42). The two incorrectly split tokens *ne* ‘not’ and *umí* ‘knows, is able, can’ are embedded within a single `mtok` element. The `c` element marks up a space disappearing after removing the word boundary as a spelling correction.³⁶ Depending on the *TEITOK* settings, either `mtok` or `tok` is exported as an indexed token to the generated CQP corpus.

(42) `<mtok ort="neumí"><tok>ne</tok><c ort="--"></c><tok>umí</tok></mtok>`

For incorrectly joined forms such as *proni* → *pro ni* ‘for her’ *TEITOK* offers a solution based on `dtok` elements nested within a `tok` element (43).³⁷ Similarly as in the wrongly split case, either `dtok` or `tok` is handled as an indexed token in the generated CQP corpus, depending on the *TEITOK* settings.³⁸

`tag` (`pos`) and UD features (`upos`). The `s` element also includes a link to the sentence annotated in the *feat* format (`featid`) and the target hypothesis after all corrections for the whole sentence (`reg`). Properties of the handwritten text are represented by the `add` element introducing an inserted text and the `lb` element standing for a line break.

³⁶For brevity, only attributes relevant to the correction of a wrongly split form specified.

³⁷Again, only attributes relevant to the correction of a wrongly joined form are shown.

³⁸The `dtok` elements are also used to annotate contractions, such as *proň* ‘for him’, a contracted form of *for him*. In a learner text, one may wish to provide error annotation for the former case and linguistic annotation for the latter. In order to distinguish the error annotation of wrongly joined

```

<s id="s-3" featid="w-NEM_GD_008-d1p2s3"
  reg="Studuje češtinu ve škole, protože neumí psát a číst správně.">
  <tok id="w-12" lemma="studovat" deprel="root" upos="VERB">Studuje</tok>
  <add>
    <tok id="w-13" lemma="čeština" deprel="obj" upos="NOUN" head="w-12">češtinu</tok>
  </add>
  <tok id="w-14" lemma="v" deprel="case" upos="ADP" head="w-15">ve</tok>
  <tok id="w-15" lemma="škola" deprel="obl" upos="NOUN" head="w-12">škole</tok>
  <tok id="w-16" lemma="," deprel="punct" upos="PUNCT" head="w-177">,</tok>
  <lb id="e-2"/>
  <tok id="w-17" lemma="protože" deprel="mark" upos="SCONJ" head="m-1">protože</tok>
  <mtok id="m-1" form="ne umí" ort="neumí" slemma="umět" deprel="advcl" upos="VERB"
    head="w-12">
    <tok id="w-176" lemma="ne">ne</tok>
    <c ort="--"> </c>
    <tok id="w-177" lemma="umět">umí</tok>
  </mtok>
  <tok id="w-19" lemma="psat" nform="psát" err="formQuant0" deprel="xcomp"
    upos="VERB" head="m-1" slemma="psát">psat</tok>
  <tok id="w-20" lemma="a" deprel="cc" upos="CCONJ" head="w-21">a</tok>
  <tok id="w-21" lemma="čistý" nform="čistě" ort="číst" err="formQuant0"
    deprel="conj" upos="VERB" head="m-1" slemma="číst">číst</tok>
  <tok id="w-22" lemma="správně" nform="správně" err="formQuant0" deprel="advmod"
    upos="ADV" head="w-21" slemma="správně">správně</tok>
  <tok id="w-23" lemma="." deprel="punct" upos="PUNCT" head="w-12">.</tok>
</s>

```

Figure 9.20: A sentence in the *TEITOK* XML format (NEM_GD_008 ru B2)

(43) <tok ort="pro ni">proni<dtok form="pro"/><dtok form="ni"/></tok>

TEITOK also provides a user-friendly way to generate a CQP corpus from the annotated texts and a customizable corpus search interface. The user can enter CQL queries or use a menu-based query builder, which also shows the corresponding CQL query. Menu items in the query builder concern both linguistic and error annotation, as well as all metadata items to filter the searched texts. For the positional tagset, a tag builder is available. Concordances can be shown in the context, with all markup

words from the linguistic annotation of contractions, the element used for contractions should have a different name.

In fact, the *mtok* element, used to annotate incorrectly split word forms, could also appear as a part of linguistic annotation to identify multi-word expressions. Instead of *dtok* and *mtok*, elements such as *contr* and *mwe* could be used for their counterparts in linguistic annotation (Maarten Janseen, p.c.).

and a facsimile, if available. *TEITOK* can also generate collocations and statistics, including various graphs.

Chapter 10

Using the corpus

Over the past twenty years, learner corpora have changed our perspective on the acquisition and use of foreign languages. One of the questions raised by the existence of learner corpora is how to use them directly for language learning. Tono (2003) presented the possibilities of using learner corpora, which can be summarized in five points: (1) description of developmental levels of the learner interlanguage, (2) studying the influence of mother tongue and language transfer, (3) defining the overuse and underuse of linguistic expressions in the learner language, (4) distinction between universal errors and errors due to the learner's mother tongue; and (5) distinction between elements of communication in native and non-native speakers that are responsible for the foreign touch.

After more than twenty years of research in this field, this thematic distribution is still valid and the identified topics are still relevant. What has changed is the number and diversity of learner corpora in the global context. As illustrated, e. g., by Granger, Gilquin, and Meunier (2015), the use of learner corpora in teaching and SLA research is a fast-growing field.

The aim of the chapter is (i) to specify the definition of the learner corpus with regard to pedagogical practice, (ii) to present an overview of the use of learner corpora and to point out the benefits and limits of their use, (iii) to map the corpus approach in teaching Czech as a foreign language, (iv) to provide an overview of analyses performed so far on *CzeSL*, and (v) to present NLP applications using the *CzeSL* data.

Research projects reported in this chapter are based both on *CzeSL* and on the native parts of the *Czech National Corpus*, as long as they are concerned with Czech as a foreign language. We start from a general view of the corpus-based

approach to language teaching and the widely used method of data-driven learning (DDL) applied to Czech. We continue with a comparison of specifics of the standard reference corpus and the learner corpus in language analysis to finally focus on the research based on the *CzeSL* data and on the NLP application.

10.1 Learner corpora from the perspective of language teachers

Similarly to standard reference corpora, learner corpora are commonly defined as systematic digitized sets of authentic texts produced by foreign language learners. This definition is largely based on the characteristics of a linguistic corpus, cf., e.g., the definition due to Čermák (2005):

Today, corpus data is characterized in relation to language as (1) typical, (2) actual, synchronous and faithful, (3) non-selective, (4) objective and realistic, (5) sufficient, (6) non-randomly acquired and (7) obtainable and obtained easily and quickly.

With regard to the interpretation of the data obtained from the learner corpus and their subsequent use in teaching practice, some aspects of this definition need to be explained and refined.

It is important for all types of corpora that they are built systematically, which ensures that texts in the corpus are selected on the basis of a number of well-defined, largely external criteria. Given that learner corpora collect interlanguage (see §2.1), their representativeness and balance requires a perspective different from standard reference corpora. Due to the progressive acquisition of the target language, interlanguage is a dynamic, constantly evolving phenomenon. Thus a learner corpus is not a representative set of homogeneous language documenting a certain period of time in language development, but rather a heterogeneous mass related to individual levels of acquisition. In *CzeSL*, this parameter is encoded in the metadata for each text as a specific proficiency level according to CEFR, the Common Framework of Reference for Languages, and complemented by details about the learner's exposure to L2 (see §4.4).

Another parameter that should be viewed differently in learner corpora than in reference corpora is their authenticity. The authenticity of language data collected for learner corpora is of a different nature. Strictly speaking, texts in learner corpora cannot be characterized as authentic, i. e., as naturally occurring or spontaneously produced in a specific communicative situation, in the same sense as authenticity is perceived in reference corpora. Authentic texts rarely reach learner corpora, as it is

very difficult to obtain such texts for any stage of the development of interlanguage. The language of L2 learners is usually tied to the school environment and, to some extent, its production is always controlled. A strong influence on the language produced has the textbook used by the learner, but the main factor is the type of text and the overall communication situation in which the text was obtained. For example, texts that are obtained as part of a certified test (as in the *MERLIN* learner corpus – see §2.3.7) carry a clear trace of the formal situation in which they are produced. The result of the exam, incurring a fee paid by the candidates, often has an impact on their career, such as admission to a school, getting a job, etc. For these reasons, learners approach the task of writing a text in the situation of a formal exam in a specific way. They try to stick to 'safe' expressions, minimizing the risk of an error. Given a topic, they produce fairly stereotypical texts. Conversely, in texts written as homework, learners feel more free to express their opinion and attitude to situations of their interest. In such texts they experiment much more and venture into a territory of grammatical structures and vocabulary that they have not quite mastered yet. For these reasons, the authenticity of texts included in a learner corpus varies widely. However, it is also possible to make use of this variability for pedagogical purposes. It is this heterogeneity of texts that can show the center and periphery of the range of learners' competence at different proficiency levels.

In the following sections we summarize options of using corpora for the research of production of a non-native language and comment on studies based on *CzeSL* and the native parts of the *Czech National Corpus*, including their application in language teaching.

10.2 The use of corpora in language research and teaching

Since the early 1990s, when the use of corpora (i. e., not only learner corpora but language corpora in general) began to expand, interest in their application has been steadily rising and both scientists and educators have been trying to exploit the data that corpora provide. Johns (1991), the pioneer of the use of corpora and the creator of a teaching method based on corpus data and their direct use, developed the DDL method for teaching English at the university level. This method heralded the change of the methodological paradigm in foreign language teaching, where the deductive approach is replaced by an exclusively inductive approach and the L2 learner is placed in a position similar to that of a child in the acquisition of L1 or that of a linguist, a language researcher. In the process of teaching, learners discover language on their own, formulate hypotheses about its use, about the functioning

of grammatical rules, lexical collocability, etc.

However, Johns also anticipated the fate of the use of corpora, i. e., that corpora are mainly used for academic and higher education purposes, while there is little evidence of their use for general language learning needs. In the Czech environment, the primary and secondary school teachers, when confronted with corpora, express their interest and positive opinion on their usefulness, but at the same time reject corpora in their own teaching as a too demanding and complex tool. It seems that it would take time before corpora become a practical pedagogical tool.

A similar attitude of teachers towards corpora is fairly common throughout Europe. This is why some methodological support for the use of corpora in teaching is emerging.¹ However, to boost the use of corpora in education, methodology is not enough. Corpus-based lexicons and grammars are equally important.² Although a wide choice of textbooks is available for Czech as a foreign language, especially at the beginners' level, there is currently no representative grammar of Czech as a foreign language or an explanatory dictionary for non-native speakers that foreign learners could use.

It is for these reasons that the use of corpora seems to be crucial for the needs of Czech teachers abroad, for whom contact with contemporary Czech is not always as easy as, for example, with English or Russian in the global context. Given that Czech is one of the less common languages, corpus could be an essential resource for its teaching.

10.2.1 Benefits of learner corpora

One of the advantages of learner corpora is the broad material base they offer. This is exactly what research into teaching and learning a foreign language needs: until recently, research in learner language was based mainly on experimental data (e. g., multiple choice tests). Experimental data can be used in analyses if the research focuses, e. g., on an abstract knowledge of a language phenomenon. For many purposes, however, it is important to find out what a student can produce spontaneously.³ Only in indirectly controlled spontaneous production can a fundamental contradiction emerge between the abstract knowledge of the language system and

¹See, e.g., Thomas (2006), Vališová (2009), Šindelářová and Škodová (2013), and Zasina (2019).

²Such as the frequency dictionary of Czech (Čermák and Křen 2004) or descriptive grammars (Cvrček 2010; Štícha 2013).

³The degree of spontaneity is affected mainly by the learner's awareness of which linguistic phenomena are elicited. Moreover, while writing a text the learner is free to adopt avoidance strategies by using easier expressions.

the actual linguistic performance of the learner. It may be risky to draw conclusions about a learner's spontaneous production from experimental data only.

Another indisputable positive of learner corpora is the fact that they can provide contextual rather than piecemeal evidence for various practically oriented research goals focused on individual phenomena. While experimental data allow for studying a limited number of phenomena of learner language at the same time, learner corpora allow monitoring of several (possibly interacting) topics at the same time. For example, the relative frequency of different types of errors can be monitored as influencing each other. Moreover, it is not entirely necessary to approach corpus data with a pre-formulated hypothesis; as a result, new aspects of learner language can be discovered by chance. Last but not least, learner corpora offer a wide textual anchoring of language phenomena; both the linguistic context and the metadata can be used to study pragmatic and discourse issues, including communication strategies.

By definition, a learner corpus is compiled on the systematic basis of precisely defined criteria. Thus it can be used to analyze the influence of individual criteria on the final form of the text and on the linguistic phenomena in the text. For example, any phenomenon can be analyzed in terms of the level of knowledge of the target language, in terms of the learner's first language, type of text, type of the environment in which it originated (natural, instructed), age, gender, duration of learning the language, L3 influence, etc.

10.2.2 Limitations of learner corpora

Although learner corpora provide a very comprehensive picture of learners and their linguistic production, there are still questions that currently available learner corpora cannot answer. One of the topics that would be difficult to investigate this way is how confident the learner is in the use of a particular language phenomenon in a specific context, i. e., whether the learner is able to use the phenomenon in its whole range (e. g., in Czech the genitive case both in the singular and plural number for all the three genders.)

Another limitation, which is not specific to learner corpora but applies throughout corpus linguistics, concerns the fact that it is impossible to investigate phenomena absent in the corpus. If a certain linguistic structure or element does not appear in the text, there is no way to find out whether the learner is aware of the phenomenon and able to use it. There are certain phenomena that cannot be found in corpora and need to be verified by elicitation in an experiment. Communicative functions are a case in point. In *CzeSL* we can find some of them, e. g., greetings, farewells, apologies, suggestions, but their range is very limited. To examine specific

communicative functions, a special elicitation prompt was required to ensure that they would be included in the texts.⁴ Another example from Czech is the range of meanings of the verb *jít* ‘go’. In comparison with the definition of this verb in a dictionary, only a negligible number of the listed meanings are attested in *CzeSL*. However, contact elicitation shows that learners actively master a wider range of meanings than they show in corpus texts (Škodová 2020).

Also, a more detailed investigation of the implicit characteristics of a particular learner is limited by the data and metadata present in the corpus. For example, it is impossible to investigate the motivation of using certain structures. Similarly, it is not possible to accurately analyze the role of various inputs to the learning process. This applies both to textbooks or other materials and to the impact of teaching methods. Such factors remain in the background of a learner corpus and are inaccessible to direct verification. They can only be examined in an experimental way, because the learner is not able to reflect on them and even for teachers such a reflection could be infeasible and subjective in principle.

The list of such phenomena should also include the role of interaction in language teaching, i. e., the stimulus for the communicative act. This phenomenon could be investigated in spoken learner corpora, as long as they are appropriately parameterized.

Other limitations are caused by the still underdeveloped state of the learner corpus field. What follows is just a partial list: (i) Learner corpora exist only for a small number of languages, in fact most of them focus on English, non-English corpora are limited to several types of texts. (ii) Narrative texts are almost non-existent in everyday production collected in the teaching process. Only fairly recently, narrative texts are beginning to be included in some learner corpora.⁵ (iii) The amounts of texts representing specific proficiency levels may differ, resulting in an unbalanced corpus. (iv) The assignment to the proficiency level according to CEFR may be approximate or subjective. (v) The learning input, i.e. the quality and scope of instructions, which subsequently lead to learner production, is very difficult to ascertain. (vi) For some languages, the number of respondents is very small.

Arguably, what follows from this list of limitations is the conclusion that the best approach to comprehensive research in second language acquisition is to combine corpus analysis with an experimental approach.

⁴See Škodová (2017) for details.

⁵A corpus of narrative texts of native and non-native speakers is under construction as a part of the AKCES project.

10.3 Corpus-based research and teaching of Czech as a foreign language

Although the teaching of Czech as a foreign language still has an upward quantitative trend, the interest in applying the corpus approach has so far been largely confined to the actual building of the learner corpus. Research based on the corpus data and their use in the classroom is still a rare sight.

This is why we start our overview with approaches using the standard reference corpora included in the *Czech National Corpus* to inform methodologies for teaching Czech as a foreign language. Only then we show research based on *CzeSL*. Due to their low number, the studies are not grouped by the topic, but presented in a chronological order.

10.3.1 The *Czech National Corpus* in the service of Czech as a foreign language

The *Czech National Corpus* (*CNC*) is used both for direct and indirect teaching of L2 Czech and in the study of the production of non-native speakers of Czech leading to methodological recommendations, but also to the emergence of new educational applications.

Czech authors have made several contributions to the corpus-based approach to teaching. Using corpus data, Vališová (2009) was the first to open a new perspective on Czech conjugation by proposing an alternative description, based on the frequency of the individual endings. She compared the results of a frequency analysis of verbs with descriptions of conjugation in grammars of Czech and Czech textbooks for foreigners. Based on this comparison, she developed an alternative classification of Czech verbs into classes and patterns that would suit the teaching of foreigners. She also developed a research probe and types of exercises within DDL for Czech, which – as an inflectional language – cannot always use the same methods as English.

The text of Osolsobě (2010), describing the use of the Czech reference corpus for DDL, is also focused on the corpus-based teaching of Czech morphology. Like Vališová, she sees in the corpus approach the chance for the learner to become independent in obtaining information about the functioning of the live language, but also a better motivation for researching the language system and manipulating language data.

This is also the topic of Lukšija (2009), who analyzes the use of the Czech prepositions *do* ‘to’ + genitive vs. *na* ‘on’ + accusative, competing to express direction

(dynamic location).⁶

Lukšija (2011) continues with her focus on the methodology of the corpus-based teaching Czech as a foreign language, showing the possibilities of using corpora to present the Czech declension. Based on the corpus data, she presents alternative declension tables of pronouns, adjectives and nouns, taking into account the frequency of their occurrences. Then she shows the methodology of using the corpus in compiling exercises for a given morphological phenomenon.

Morphology is analyzed also by Hudousková (2014). Based on data from the *CNC*, she is concerned with the use of competing forms of personal and possessive pronouns by native speakers of Czech. She focuses on the distribution and frequency of use of individual forms, from which she draws conclusions for the presentation of the paradigms in Czech for foreigners.

A methodologically important text is an article by Vališová (2016), describing the *CNC* tools *SyD* and *KonText* and options for using them directly in teaching as complements of textbooks and dictionaries. She also discusses the indirect use of *CNC*, namely the types of exercises built on the *CNC* data and their suitability for teaching Czech as a foreign language, but also how demanding their preparation is.

The most extensive work, based on both *CNC* and *CzeSL*, is the dissertation of Zasina (2019). The thesis employs both resources in a combination of research-oriented and methodology/application-oriented approach to Czech as a foreign language. The link between these two perspectives is the corpus. On the one hand the corpus allows researching the language production of non-native speakers in Czech as the target language. On the other hand, based on the results of the research, it offers linguistic data for compiling compensatory texts and exercises to support learners or learners of a certain type in coping with specific difficulties in learning the target language. The *CzeSL* corpus is used to identify linguistic phenomena with a high error rate, while the *CNC* reference corpus of contemporary Czech is used for compiling compensatory exercises. Zasina offers a systematic methodological procedure for teachers of Czech as a foreign language to create their own exercises, based on the evidence of problematic areas of language acquisition, which should be supported by complementary and expanding exercises using a representative sample of the Czech language and not based only on the teacher's idiolect as the primary source of language data.

⁶See Škodová (n.d.) for an analytical study on this topic based on learner data.

10.3.2 Analyses based on learner corpus data

Hudousková (2013) published one of the first studies based directly on the *CzeSL* data. The author combines a survey of the use of the pronoun *který* ‘who/which/that’ in *CzeSL* with an analysis of how the pronoun is presented in textbooks of Czech as a foreign language. She concludes that the use and error rate in the texts indicates an inadequate presentation of this phenomenon in textbooks. This contribution demonstrates how important the analysis of corpus data can be for the methodology of teaching Czech as a foreign language.

A comprehensive work based on the analysis of *CzeSL* is the diploma thesis of Vokáčová (2016). She examines the influence of frequency characteristics of Czech nouns on their acquisition by non-native speakers. She shows that non-native speakers tend to be guided by grammatical profiles of the nouns. The production of nouns by non-native speakers corresponds to the frequency characteristics of nouns and shows a low proportion of morphological errors in their most frequent forms. Vokáčová explains the cases in which the error rate deviates from this model by the type frequency – the simultaneous effect of productivity of certain declension patterns – and by the higher relevance (for non-native speakers) of the nominative, the basic form.

Another comprehensive work is the dissertation of Pečený (2017). Because of his focus on language testing, the study is based on the *MERLIN* learner corpus – all texts in *MERLIN* are exactly classified at the CEFR scale, because they were collected from certified exams. Pečený describes the repertoire of connectors used by non-native speakers of Czech and examines tendencies of their use. Using correspondence analysis, he captures the relationship of connectors to the individual CEFR levels. The thesis also includes a detailed quantitative-qualitative error analysis of the use of connectors.

One of the results attesting the long-lasting interest of the author in the analysis of *CzeSL* data is the study by Škodová (2018), analyzing the use of the verb *jít* ‘go’, mainly in the whole range of its collocations and grammatical properties, i. e., also in meanings other than the primary meaning of movement proper.

Zasina and Škodová (2020) analyze the use of prefixes for the verbs movement *jít* ‘walk, go by feet’ and *jet* ‘ride, go by some means of transport’. The prefixation of these verbs is connected both with the directionality of the process and with the phases of its course. The study points out the principles of overuse and underuse of particular verbal prefixes.

Škodová (2020) presents a comprehensive analysis of the use of the verb *jít* ‘go’ in texts of non-native speakers. She focuses mainly on the polysemy of the verb, one of the most frequent in Czech, and shows how the use of the semes changes

depending on the achieved proficiency level.

Škodová (n.d.) analyzes the competition of genitive and local verbal complements based on the material of the learner corpus. The constructions are examined in relation to dynamic and static verbs. The ability to understand directionality as such plays a crucial role in distinguishing between the potential of genitive and local valency. The fundamental semantic component of location predicates is the spatial relation, very often expressed by prepositions.

Although there are still not many studies based on *CzeSL* data, their number is bound to grow, because seminars presenting and using *CzeSL* have been included in the university programs. The programs preparing teachers of Czech for both non-native and native learners now include regular courses on the methodology of teaching Czech as a foreign language. In these courses, learner corpora are presented together with possibilities they offer, and the *CzeSL* corpus is introduced in a hands-on tutorial. For doctoral students, a lecture in the doctoral seminar features the learner corpus topic. The number of bachelor's, master's and doctoral theses using the *CzeSL* data is thus increasing.

The program *Czech as a foreign language* has become available recently as a three-year bachelor course at Technical University in Liberec, and as a two-year master course at Charles University in Prague and Masaryk University in Brno. Texts collected for *CzeSL* are already in use in the training of teachers at all the three universities and also at Palacký University in Olomouc to give them an idea about the traits of the learner language in relation to the author's L1 and proficiency. This should help them to change perspective from viewing the language as an abstract system to approaching Czech as a sum of components acquired by learners at a specific stage of the development of their interlanguage.

A specific problem is the issue of educating children with a native language other than Czech, whose presence at Czech primary schools is a recent phenomenon. Until recently, primary school teachers received no training in teaching Czech as a foreign language, resorting to an individual and intuitive approach. By its inclusion in *AKCES*, *CzeSL* becomes a resource for research and design of teaching materials assisting teachers of young non-native speakers at different stages of the acquisition of Czech. At the same time, *CzeSL* should provide representative data that would help initiate and develop systematic and comprehensive research of Czech as a foreign language – there are no monographs available dealing with this topic so far.

10.4 Applications in natural language processing

Texts written by non-native and native learners of Czech, collected and annotated within the *CzeSL* project, and some other texts from the *AKCES* project have been used in NLP applications of at least three types: (i) scoring of texts written by non-native speakers according to the CEFR proficiency scale, (ii) text correction, (iii) estimating the native language of the text author.

10.4.1 Text scoring

EVALD (Evaluator of Discourse; Novák, Rysová, Mírovský, et al. 2017; Novák, Rysová, Rysová, et al. 2017; Novák et al. 2019) is based on *CzeSL* and *MERLIN*. *EVALD* scores essays written by non-native speakers of Czech in accordance with the CEFR scale, distinguishing the six grades (A1–C2). Trying to imitate assessments made by humans, the software was trained on 945 original essays from *CzeSL* and *MERLIN*, estimating proficiency level of the text according to features derived from various linguistic domains: spelling, morphology, vocabulary, syntax, and text structure (in terms of coreference and discourse relations). In addition, the system also provides some assessment of the weak and strong points of the text.

*EVALD*⁷ is based on *CzeSL* and *MERLIN*. *EVALD* scores essays written by non-native speakers of Czech in accordance with the CEFR scale, distinguishing the six grades (A1–C2). Trying to imitate assessments made by humans, the software was trained on 945 original essays from *CzeSL* and *MERLIN*, estimating proficiency level of the text according to features derived from various linguistic domains: spelling, morphology, vocabulary, syntax, and text structure (in terms of coreference and discourse relations). In addition, the system also provides some assessment of the weak and strong points of the text.

In spelling, *EVALD* reflects mainly typing errors and punctuation marks. At the level of morphology, it monitors features such as distribution of grammatical cases, distribution of parts of speech, use of verbal voice, mood, aspect, tense etc. In terms of vocabulary, *EVALD* takes into account a variety of used lemmas (richness of vocabulary), the number of unrecognized lemmas or word length. Concerning syntax, it reflects sentence length, distribution of main and subordinate clauses, nonverbal clauses or structural complexity of the sentence (in terms of the tree height and the number of nodes). Evaluation of text structure covers coreference (including the number of coreference relations, length of coreference chains or variety of lemmas used in the coreference chains) and discourse relations (covering variety of

⁷The *Software Applications for Automatic Evaluation of Discourse in Czech*, see Novák, Rysová, Mírovský, et al. (2017), Novák, Rysová, Rysová, et al. (2017), and Novák et al. (2019).

discourse connectives, distribution of coordinating and subordinating connectives, variety of discourse relations, distribution of inter- vs. intra-sentential relations, distribution of semantico-pragmatic relations, etc.).

This application was used in an experiment (Škodová, Rysová, and Rysová 2019), evaluating the agreement on the CEFR level assessment of texts of non-native speakers. Results of human evaluators were compared with each other and with those produced by the system. The result of the experiment demonstrated the high variance in the assessments of human evaluators and the incompatibility in the application of descriptors to individual proficiency levels. An automatic assessment tool trained on learner corpus data proves to be more compact in evaluating the individual aspects of learner texts.

10.4.2 Text correction

Learner corpora can be useful for training NLP tools for checking and/or correcting texts written by both native and non-native speakers. Richter (2010) designed and implemented *Korektor*, a general language-independent stochastic tool, combining the functionality of correcting the spelling of unknown word forms with the option of correcting real-word errors, incorrect only in context, such as *at the end of the *weak* → *at the end of the week*.

The models used by *Korektor* are built from three resources: a lexicon with a morphology module, a corpus of correct texts, and a corpus of texts with errors and corrections.⁸ *Korektor* can be used in several modes. To correct each potentially incorrect word form it can produce a single most likely correction, or generate one or more most likely suggestions. *Korektor* can also generate or strip diacritics.

In the initial version (Richter 2010), the error corpus was substituted by a text transcribed from its audio version.⁹ In a later version (Richter, Straňák, and Rosen 2012), the error model was built from additional resources: *Chyby* (Pala, Rychlý, and Smrž 2003) and misspellings extracted from *WebColl* (Marek, Pecina, and Spousta 2007). *Korektor* was then evaluated using the doubly hand-annotated part of *CzeSL-man* (67 texts, 9,372 tokens, see §8.3). About 10% of the tokens were not

⁸See <https://ufal.mff.cuni.cz/korektor> (Richter 2013; Straka and Richter 2015). The site offers *Korektor* as a command line utility, a publicly available web service with an API and a web service with an HTML front end. There are also instructions on a number of customizable options and on building the model, consisting from the lexicon+morphology module, the language module and the error module from data in a specified format. Several ready-made models built for Czech are available. However, the currently available implementation *Korektor* does not handle incorrectly split or joined words.

⁹The text was Jaroslav Hašek's novel *Osudy dobrého vojáka Švejka* 'The Good Soldier Švejk'. The audio is available from <https://www.rozhlas.cz/ctenarskydenik>.

recognized by a tagger (Spoustová et al. 2007), 13% of the tokens were corrected in the same way by both annotators at T1 and 16% at T2. In the comparison of the results of *Korektor* with those of the tagger, *Korektor* scored 86% in terms of F-measure for the correct/incorrect status of each form. In the comparison with forms annotated at T1 and T2, provided both annotators were in agreement, *Korektor* scored 72% for T1 and 53% for T2.

The results supported the idea to integrate *Korektor* into the learner corpus annotation workflow, to provide suggestions to the annotator or to perform fully automatic annotation. In fact, the normalization (correction) task in the error annotation of *CzeSL-SGT* (see §8.2) is done exclusively by *Korektor*. More recently, the web service of *Korektor* has become a one-click option for the annotator using *TEITOK* (see §9.2.3) to perform or suggest corrections in a text.

Ramasamy, Rosen, and Straňák (2015) use the *WebColl* corpus to train an error model of native Czech, alongside two error models of (almost) non-native Czech, trained separately on T1 and T2 corrections extracted from *CzeSL-man v0*, i. e., the release including the Romani ethnolect.

The three models were tested on three data sets: the audio transcripts, introduced above, as a native text and the T1 and T2 versions of *CzeSL-man* as non-native texts. Interestingly, the best results on the native texts were achieved by the non-native models (95.9% for error detection and 95.2% for error correction in terms of F-measure), while the best results for error correction of non-native texts were achieved by the native model (75.4% for T1 and 68.8% for T2). For error detection in non-native texts the results matched expectations: 82.2% on T1 texts for the T1 model and 76.1% on T2 texts for the T2 model.

The authors attribute the unexpected outcome to the untuned state of the non-native models and the variety of learner texts, written by learners with various proficiency levels and first languages and also by Czech pupils with Romani background. The paper also includes a detailed analysis according to error types of some T2 texts as corrected by *Korektor*, concluding that real-word errors due to missed agreement or government are the most problematic. Apart from the finding that both native and non-native error models perform well on spelling-only errors, the main conclusion is that non-native models can outperform native models even on native texts. Last but not least, the experiment showed an improvement over the results reported in Richter, Straňák, and Rosen (2012).

Náplava (2017) designs and implements several neural network models for several language checking/correction tasks. The models are trained on the existing and two additional resources: a grammatical error correction dataset based on *CzeSL*¹⁰

¹⁰See *CzeSL-GEC* in §8.7.

and an automatically created spelling correction dataset. The models significantly outperform existing systems on the diacritization task and achieve the best results for two out of the three datasets in the spelling and grammar correction tasks.

Náplava and Straka (2019) achieve even better results, by using a neural machine translation system trained on texts treated as parallel corpora. The hand-annotated corpus *CzeSL-GEC* is extended by additional *CzeSL* texts, which were manually normalized (but not error-labeled) more recently, and released in the M2 format as *AKCES-GEC*.¹¹ The best results in terms of F-measure are as high as 80.2% for all *AKCES-GEC* texts. This figure is different for the three observed groups of learners: 81.4% for Slavic learners, 76.5% for learners with other L1 and 83.0% for Romani pupils. The differences are attributed to the inverse proportion of errors in the corresponding subcorpora. The authors also report better performance on T1 errors than on T2 errors, which can be explained by a higher frequency of T1 errors.

10.4.3 Natural language identification

The knowledge of author's native language can be useful for various NLP applications. For example, it allows tuning NLP tools to the typical errors of authors with a particular L1, or a language tutoring systems can refer to particular aspects of the learner's L1 in its feedback (cf. Amaral and Meurers 2008). In addition, the task itself can provide insight on the nature of language transfer (Jarvis 2012).

It has been shown (Bykh and Meurers 2012; Hladká, Holub, and Kříž 2013; Tetreault et al. 2012) that non-native text contains enough information to identify the native language of the author with a reasonable accuracy.

Aharodnik et al. (2013) and Tydlitátová (2016) used *CzeSL* data to validate that similar results can be obtained for Czech, a highly inflectional language. Aharodnik et al. (2013) classify the authors as native speakers of Indo-European or non-Indo-European language, while Tydlitátová (2016) distinguishes Slavic and non-Slavic backgrounds.

Aharodnik et al. (2013) intentionally avoid content-based features (i. e., features directly based on the source text, including word and character n-grams). Instead, they use POS n-grams and/or manually provided error-tags as features. Tydlitátová (2016) did not eschew content features but limited herself to information that can be obtained automatically: character, word and POS n-grams, function words, average

¹¹See *AKCES-GEC* in §8.7.

sentence and word length, and automatically derived error-types.¹²

The results of these projects showed that:

1. Character n-grams and/or POS n-grams are a good indication of author's native background.
2. Word n-grams worked well but showed a topical bias (e. g., the occurrence of word *ruský* 'Russian' is a good indication of an essay written by a Russian speaker). Therefore these features are not very useful outside of artificial benchmarks.
3. Manual classification of errors worked very well (Aharodnik et al. 2013), while automatically obtained error-tags are much less useful (Tydlitátová 2016).

¹²Formal error tags derived from *Korektor*-based corrections as available in *CzeSL-SGT*, see §8.2.

Chapter 11

Lessons learned and perspectives

We start with the nicer parts of the long journey towards the present shape of the corpus (see §11.1) and continue by a list of some pitfalls we fell into (see §11.2). Finally, we draw a sketch of how to proceed further (see §11.3).

11.1 What we would do the same way again

The result is worth the trouble

Experience from teaching Czech as a foreign language clearly indicates the need for a rich source of data on the language of learners, one which would help to design an intuitive presentation of the Czech language for non-native speakers, accompanied by exercises and tests. A learner corpus is the answer also because the typological properties of Czech as a highly inflectional language make the use of experience from other, better positioned languages at least questionable. In this sense, Czech may serve as a testbed for the development of methods and tools targeting inflectional languages.¹

Several approaches, various uses

We have designed and implemented several approaches to the concept, collection, annotation and exploitation of a learner corpus of Czech, resulting in several releases,

¹Some aspects in the design and compilation of *CroLTeC* (see §2.3.4) and *RLC* (see §2.3.8) were influenced by *CzeSL* as both a positive and a negative example. Foundations of a learner corpus of Polish, inspired by *CzeSL in TEITOK* (see §8.8) are described by Kaczmarska and Zasina (2020).

available for on-line queries and downloads. The corpus data have been found useful as a resource for building proofing tools and other NLP applications and, so far at least to some extent, for the practice of teaching Czech as a foreign language and research in its acquisition. Furthermore, some of the methods and tools developed within this project have been re-used in other projects.

Wide user focus reflected in the annotation

A learner corpus may be intended for a group of users with specific research or practical needs,² or for a wide audience of language acquisition experts, researchers or practitioners. *CzeSL* has gone the latter path, necessitating some compromise or generalizing solutions which may not quite fit a specific goal of the user. The wider focus is reflected especially in the approach to error annotation, which employs categories adopted from established grammar-based descriptions of native language rather than from a specific view of the learners' interlanguage. We believe that this type of annotation can serve as a common ground for uses of various kinds with the option to provide additional, perhaps more targeted annotation. Moreover, its compatibility with the annotation used in the native Czech parts of the *CNC* supports comparisons of native and non-native Czech based on search results or statistics.

Several complementary types of error annotation

Annotation in general, and error annotation in particular, takes up a substantial share of space in this book, because it represents some of the essential analysis steps of learner texts, essential in most types of corpus use. Recognizing the variety of users' expectations and the multitude of learner text aspects contending to be shown in the annotation, we have designed or modified several annotation schemes and error taxonomies. They are applied to complement each other rather than as alternatives.

The schemes differ also in how they resolve the frequent inherent impossibility to decide about the causes of observed deviations from the standard language. In the 2T scheme the annotator is instructed to pick the "most sophisticated" error tag (see §5.4.2.2). On the other hand, in the MD scheme, two or more tags related to different domains, such as spelling, morphonology, morphology or syntax, can be used simultaneously to provide alternative descriptions of an error if the cause is not clear.

²For Czech, *MERLIN* (see §2.3.7) may be an example of a learner corpus of this type, designed to provide texts illustrating the CEFR proficiency levels.

Benefits of grammar-based error annotation

The support of varied types of use is not the only reason why we opted for an annotation scheme and error taxonomy based on grammatical deviations from the standard, without a specific focus. Other reasons are due to concerns about annotation as a process. The grammar-based strategy fits well with the typological properties of Czech and the fairly common homonymy and synonymy of affixes. A grammar-based, i. e., formally well-defined, taxonomy has a desirable effect of maintaining better consistency of manual annotation. At the same time, it allows for extensions into new domains of annotated phenomena, and into more efficient annotation processes, such as automatic assignment of more detailed error categories, automatic morphological and syntactic analysis or (semi-)automatic correction and error tagging.

Automatic annotation

Nearly all linguistic annotation tasks have been referred to taggers and parsers, even tasks involving source texts, for which the tools were not designed or trained. It turns out that a source learner text tagged and lemmatized this way is definitely more useful than the same text without any linguistic annotation, even if the bonus is merely the clear identification of non-words.

For error annotation, proofing tools providing automatic corrections as well as purpose-built error identifiers were used in pre-processing and post-processing to assist annotators, or in fully automatic annotation of larger volumes of texts. While a combination of manual and automatic annotation is appreciated by the annotators and corpus users, results of a fully automatic process are considerably less reliable and sophisticated than manual error annotation, but still useful for some purposes, as evidenced by research based on the automatically annotated *CzeSL-SGT* corpus (e. g., Hudoušková 2013, 2014; Novák, Rysová, Rysová, et al. 2017; Novák et al. 2019) and the history of user interactions with this corpus via the search interface. A fully automatic annotation is obviously justified as an alternative to manual annotation when the demand for large data is higher than concerns about the error rate.

The importance of annotators' feedback

The annotation process brings plentiful feedback, reflected in discussions in the web forum, training sessions for the annotators and in the annotation manual. The feedback helped to improve instructions to deal with thorny issues such as the uncertainty about the author's intended meaning and its expression, the inference

errors, the proper amount of interference with the original, or the occurrence of colloquial language. In all of this, annotators should handle similar phenomena in the same way.

The IAA results show that the rules for manual tagging of errors in spelling, morphonology, morphology and morphosyntax, such as `incorStem`, `incorInfl`, `agr` and `dep`, are assigned fairly consistently. However, we were unable to obtain a similarly robust annotation of semantic errors, which are much more dependent on subjective judgment. It is even unclear whether it is desirable to aim at a standard for their annotation.

11.2 Blind alleys and second thoughts

Sloppy planning

In order to reach its goals and become useful, a learner corpus project should be conceived carefully, considering many factors and avoiding blind spots in the plan, from text collection to user access. A change in the corpus design may leave permanent traces. For example, *CzeSL-plain* and its hand-annotated part *CzeSL-man v0* include a substantial share of the Romani ethnolect, actually produced by native speakers of a dialect of Czech, rather than by non-native speakers of Czech. This is due to the original strategy of grouping texts by the way they are processed. This has been changed in later releases, where texts produced by non-native and native learners (the latter including speakers of the Romani ethnolect of Czech) are parts of distinct corpora.

Another example is the fact that many texts have not been processed properly. In the early days of the project, most efforts were focused on collecting, transcribing and hand-annotating, yet not all transcribed texts were annotated and checked, and no doubly annotated texts were adjudicated, as originally planned. Later, resources for such tasks dried up, which invited more interest in automating the annotation tasks.

Missing or inaccurate metadata

Neither *CzeSL-plain* nor *CzeSL-man v0* include the full set of metadata, which were not available in the appropriate form and content at the time the two corpora were prepared and released. In *CzeSL-plain*, the texts are categorized into three groups: as essays, written either by non-native learners, or by speakers of the Roma ethnolect of Czech, and as theses written by non-native students. In the searchable release of *CzeSL-man v0*, even this basic distinction is not available.

An issue of a different sort is the unreliability of an important metadata item. The CEFR level is specified according to the teacher’s assessment, because many texts in *CzeSL* do not come from test situations. As a result, the information about proficiency level is a major weakness of the corpus. The CEFR classification of texts is based on a holistic evaluation made by teachers, collectors or annotators, or according to the level of the whole class. Collectors and annotators were instructed and trained about the proficiency levels, using guidelines and references. Many of them were experienced teachers of Czech as a foreign language.

Yet even experienced educators and methodologists may arrive at different ratings for a single learner, while learners in a single class can differ substantially in their levels. Moreover, the CEFR level evaluation standard for Czech was not available at the time the texts were collected, transcribed and annotated. Without an independent objective metrics it was difficult to provide a consistent classification, especially across various L1s.

Ideally, the texts should be re-evaluated, as was done, for example, in the *MER-LIN* project, but it would be a labor-intensive project in itself, if done manually.

Multiple taggers and the multidimensional tagset

We did not pursue all ideas about automatic annotation until their implementation in the annotation toolchain. One of the most interesting experiments was an attempt to apply different POS tagging methods to the source text, as in Díaz-Negrillo et al. (2010). We expected different results for faulty forms across the taggers and planned to implement a method proposing a hypothesis about the error type by comparing these results. However, the results of multiple taggers, based on different tagging strategies, lead to a usable interpretations of faulty forms only in a limited number of cases.

We also tried to apply the concept of multidimensional word classes, at first in combination with the application of multiple taggers, and later – independently from any approach to the annotation process – as a categorization of morphological and morphosyntactic errors in Czech. According to this initial and later abandoned blueprint, the dimensions coincided with the lexical, morphological and syntactic properties of a word form. As in Díaz-Negrillo et al. (2010), the assumption was that phenomena of non-standard language can be modelled as mismatches between the three dimensions. For example, in *Petr viděl *lev* ‘Petr saw a lion’ instead of *Petr viděl **lva**, lev* ‘lion’ is morphologically nominative, but syntactically accusative (*viděl* ‘saw’ requires its object to be in the accusative case). In *Eva *bude *napsat dopis* ‘Eva will write a letter’ instead of *Eva **napiše** dopis* or *Eva **bude psát** dopis*, the ‘lexical’ aspect of the content verb *napsat* is perfective, while the auxiliary verb *bude*

has a ‘syntactic’ requirement for an imperfective form *psát*. In *Whitney Houston zpívala *krásný* ‘Whitney Houston sang beautiful’ instead of *Whitney Houston zpívala krásně*, the author used an adjective *krásný* ‘beautiful’ rather than the adverbial *krásně* ‘beautifully’. The word can be annotated as an adjective in the morphological dimension and as an adverb in the syntactic dimension.

However, the morphological idiosyncrasies of non-native Czech call for additional error categories capturing morphemic rather than morphosyntactic phenomena. The wrong form *leva* in (16) could still be modelled as a mismatch across the different dimensions of the annotation, in this case between the proper “lexical” paradigm and the “inflectional” paradigm assumed by the learner. Mismatches between different aspects of the analysis of a form should then coincide with a taxonomy of errors.

Despite its theoretical appeal and an affinity with standard linguistic concepts, this approach to classifying learner errors in Czech morphology turned out to be imposing a somewhat artificial paradigm upon the empirical facts.

The crucial difference between the original concept of multidimensional word classes and the final design of the multidimensional error annotation scheme (see §5.6) is in the interpretation of multidimensionality. According to the original proposal, an error was characterized by the conjunction of dimensions. According to the final design, the dimensions (or linguistic domains) are treated as alternative explanations of the error, i. e., in disjunction.

Unbalanced representation of text types and learner categories

Some balance or at least representative proportions of text types (argumentative essays, descriptions) and learner categories (L1, CEFR level) are necessary or at least useful. Tables 8.5–8.8 show an opposite, opportunistic approach, driven by practical constraints, often justified by the unavailability of texts of a specific category. To some extent, the imbalance has been remedied in more recent text collection and annotation rounds (see §8.4 and §8.8).

Transcription

To avoid the need of cleaning transcripts with improperly used mark-up, an editing tool including strict format controls is preferable to a free-text editor.

Tokenization

Designing tokenization rules specific to learner texts and different from those used in tools for native texts is not worth it. It makes it much harder to use existing NLP tools as they typically assume certain tokenization.

Annotation scheme vs. the ease of using the corpus

A scheme ideally suited to the data may turn into a problem later, if the consequences for the annotation process and the use of the corpus are not foreseen. Standard concordancers may require substantial tweaking of the data, while a custom-built tool may lack features of the tools developed for a long time. At the same time, most users of this type of corpora definitely need a friendly interface.

The 2T annotation scheme, designed to fit the needs of error annotation of L2 Czech, requires a specific corpus search tool or lossy conversion into a more common format. *SeLaQ*, as the custom-built search tool, is able to process and search the 2T data without conversion and information loss, but cannot display the tiers in parallel, ignores metadata and lacks many features of a more mature corpus search tool. On the other hand, the *KonText* and *TEITOK* search tools, even though they require conversion of the 2T format without retaining all details of the error annotation, offer many more user options. In fact, the *CzeSL-man v2* and *CzeSL in TEITOK* corpora go a long way towards handling most of the 2T error annotation, even if some of the properties and information present in the 2T scheme get lost in the conversion to the format used by the corpus search tool.

Too many data formats

The *CzeSL* texts are available in several annotation schemes and various data formats. Formats of the on-line searchable releases correspond to the search tools and search interfaces while formats of the downloadable data sets to standards or preferences of potential users or applications. Some of these formats exist due to purely pragmatic reasons. As explained above in [Annotation scheme vs. the ease of using the corpus](#), a format used for transcription and annotation may not be suitable for searching using a standard tool with necessary features and must be converted to a different format. This is the case of the *CzeSL-man v0* corpus, which has been converted to *CzeSL-man v1* and *CzeSL-man v2*. The format of *CzeSL-plain* and *CzeSL-SGT* corpus, which were not annotated manually, was determined by the annotation tools and the search tool.

Other formats were introduced to allow for annotation which was not originally previewed. This is the case of the 2T annotation scheme. Together with the annotation editor it was designed to suit the error annotation of Czech with a focus on syntax and morphosyntax. The scheme was not meant to handle morphs and other segments smaller than a word. Also it was not intended to accommodate linguistic annotation of syntactic structure. These are the reasons why the formats used in the *CzeSL-MD* and *CzeSL-UD* corpora are different from the format used in the *CzeSL-man v0* corpus.

A format suitable for a search tool is usually less flexible than the annotation format. The 2T annotation cannot be represented in the vertical format, used by the standard corpus search tools, without some loss of information. The loss is higher in *CzeSL-man v1*, representing only the TH and error tags at T2, and the corresponding tokens at T0, unless cross-tier links other than 1:1 are involved. The rest is discarded, including source tokens which cannot be associated 1:1 with T2 tokens. In *CzeSL-man v2*, the loss is much lower: only corrections involving long-distance word order changes and some complex multiple cross-tier links are lost. The format used in *CzeSL-MD* is even harder to reconcile with the token-based vertical format, because segments smaller than words can be annotated.

As a general solution, for multi-tier learner corpora annotated in various independent or interacting ways, a consistently applied stand-off annotation format is – at least in theory – the best solution.³ Our approach is similar. In order to integrate several corpus releases including the same texts annotated in different ways in different formats for the purpose of representing and searching the annotation in a consistent and user-friendly way, we convert existing non-token-based annotation into stand-off annotation applied to the texts. However, these texts are tokenized and any annotation which can be token based, is converted from the various sources and represented as token attributes. The annotation follows the TEI guidelines and the result is searchable by *TEITOK*.

There are several reasons behind our choice. The solution allows for compatibility with the standard token-based search tools: the same corpus can be searched in one tool and the concordance represented in another tool. Moreover, the corpus, including its metadata, can be extended, modified and annotated within the same tool. The texts can also be viewed as facsimiles or at different annotation tiers. On the other hand, even though the conversions from various annotation formats can reveal some inconsistencies, they also require some human assistance and are not completely error-free. Obviously, a format and tools compatible with the desirable

³The *ANNIS* search tool site at <https://corpus-tools.org/annis/> includes also suggestions for annotation editors.

annotation and search options would be the ideal solution.

11.3 Outlook

- Compilation of a balanced hand-annotated subset of *CzeSL*
- Integration of all *CzeSL* texts with all available annotation in a single search and editing tool (*TEITOK*)
- Periodical release of the updated corpus in *KonText*
- Incremental inclusion of new texts, including annotation
- Continuous proofreading of annotated content
- Development of automatic tools for linguistic and error annotation of learner language
- Manual annotation of learner texts to serve as training data for the tools

Chapter 12

Acknowledgements

The *CzeSL* corpus could not be built without the efforts of many students – the devoted and careful collectors, transcribers and annotators. Their feedback was invaluable in making the guidelines more accurate and easier to follow.

We are also grateful to those who guided and coordinated the collection, transcription and annotation, including the collection of metadata for the original texts and proofreading the results: Zuzanna Bedřichová, Milena Hnátková, Kateřina Lundáková, Piotr Pierścieniak, Dagmar Toufarová, Kateřina Šormová and others.

Many thanks are due to Hana Skoumalová (Institute of Theoretical and Computational Linguistics), who was the patient advisor for all sorts of technical problems; Maarten Janssen (Institute of Formal and Applied Linguistics), the author of *TEITOK*, whose advice and generous help was crucial for making *CzeSL* available in the versatile and useful tool; Tomáš Machálek and Pavel Procházka (Institute of the Czech National Corpus) for their share in releasing the *CzeSL* corpora in *KonText*; Vojtěch Kovář and Vlasta Ohlídalová (Lexical Computing) for help with *Sketch Engine*; Olivia Goodman (Cambridge University Press) for help with *Cambridge Learner Corpus*. We are also grateful to Jan Štěpánek, the author of *SeLaQ*, and to everyone who helped to build the tools needed for the linguistic annotation of native Czech – the taggers and parsers (e. g., Votrubec 2006; Straková, Straka, and Hajič 2014, and also for the error annotation of non-native Czech – the *Korektor* tool (Richter 2010; Richter, Straňák, and Rosen 2012; Náplava and Straka 2019).

We also wish to thank others who were members of the team during some important stages, namely Milena Hnátková, Vladimír Petkevič, and Hana Skoumalová, for their numerous stimulating ideas, acute insight and important feedback. We are especially grateful to Karel Šebesta, for all of the above and for initiating and

guiding this enterprise in the wider context of acquisition corpora, including corpora of native learners and corpora of spoken Czech, produced by both native and non-native learners.

Last but not least, we wish to thank Elena Volodina and Detmar Meurers, for carefully reviewing the book, helping us to improve it with their expert knowledge and rich experience in many points and aspects. Any remaining problems are entirely our fault.

A project of this scale would not be possible without institutional support. Throughout the years we have received funding from various sources, gratefully acknowledged in the list below.

- The project was supported in 2009–2012 within a joint project of Technical University Liberec and Charles University Prague with the *CzeSL-plain* corpus as a main result: the operational program Education for Competitiveness, funded by the European Structural Funds (ESF) and the Czech government – Innovation of Education in the Field of Czech as a Second Language (project no. CZ.1.07/2.2.00/07.0259).

The institutions involved in the creation of the corpus include: Technical University of Liberec as the beneficiary of the support, Charles University in Prague and The Association of Teachers of Czech as a Foreign Language as partners, and a number of elementary schools and high schools, civic associations, NGOs and other institutions as well as individual collaborators.

- The project has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).
- The *feat* annotation tool was partially funded by grant no. P406/10/P328 of the Grant Agency of the Czech Republic.
- Data format development was partially funded by grant no. P406/2010/0875 of the Grant Agency of the Czech Republic.
- The technical aspects of publication of several corpora and their on-line availability at the site of the Czech National Corpus have been partially supported by the Czech Ministry of Education, Youth and Sports, Large Research, as a part of the program Development and Innovation Infrastructures, ‘The Czech National Corpus’ project no. LM2011023 and the follow-up project no. LM2018137.

- Acquisitions, transcriptions, the provision of metadata and other work related to the preparation of some corpora were supported from PRVOUK, the research funding program at Charles University: P10 – Linguistics, Acquisition and Development of Linguistic and Communicative Competence in Selected Communities of the Czech Republic.
- Some of the recent activities, especially manual annotation in the *TEITOK* tool, has been supported by the KREAS project of the Operational Programme Research, Development and Education, a part of the Structural and Investment Funds of the European Union (<https://kreas.ff.cuni.cz/en/>).
- The recent developments, especially in the linguistic and error annotation (the design of the UD, MD and implicit schemes, testing of the guidelines based on manual annotation, annotation of *CzeSL-TH* and integration of the results) have been funded by the Grant Agency of the Czech Republic, grant 16-10185S (Non-native Czech from the Theoretical and Computational Perspective). The agency deserves our gratitude also for helping us publish this book.

Appendix A

Notes about examples

- Nearly all examples are based on authentic data from the *CzeSL* texts. If so, they are followed by a reference to the text ID, the code for L1 and the CEFR level of the author. A wider context of the example can be found in some of the searchable releases, e. g., in *CzeSL-SGT*. Some examples in Appendix B are from the SYN-series of the Czech National Corpus: SYN2005 (*Czech National Corpus – SYN2005 2005*) and SYN2006PUB (*Czech National Corpus – SYN2006PUB 2006*).
- Most examples include source text (as written by the learner) and its correction (a target hypothesis – TH). Examples like this in running text are represented with an arrow separating the source and the TH: *incorrect* → *corrected*. In such examples, a single TH is given, rather than successive or alternative THs. Ill-formed strings and their corrections are often highlighted by boldface.
- In numbered examples where the concept of annotation tiers is relevant we use the label “T0:” for the source forms and “T2:” for the target forms (the TH). When appropriate, we also use the label “T1:” for the intermediate TH. When the error annotation scheme does not use the concept of annotation tiers, we use “S:” for the source text and “T:” for the TH instead of T0 a T2 (3).
- If possible, glosses follow conventions of the Leipzig Glossing Rules.¹

¹<https://www.eva.mpg.de/lingua/resources/glossing-rules.php>

- When a particular inflectional form is not relevant for the discussion, the glosses are simplified by using base forms, e. g., nominative singular for nouns, masculine nominative singular for adjectives, infinitive for verbs.
- The gloss of an ambiguous word form shows only the contextually relevant labels rather than all possible interpretations. In (16) on page 109, the form *lva* in the TH sentence *vidím lva* ‘I see a lion’ is glossed only as ‘lion.ACC’, even though the same form can also be used in *bojím se lva* ‘I am afraid of the lion’, where the verb requires the genitive case.
- The asterisks in glosses of the source forms are used to mark an ill-chosen inflectional suffix, interpreted in terms of a morphosyntactic category (. *DAT in (17)).
- The parentheses in glosses of the source forms are used to mark an interpretation of the whole form or its inflectional suffix which is not correct but can be understood as such. In *míluje* → *míluje* ‘(loves)’ 34 on page 159 the parenthesized gloss of *míluje* indicates that it is not a correct form but can be understood as ‘loves’. In *leva* → *lva* ‘lion.(ACC)’ (16) or *levy* → *lvům* ‘lions.(*ACC)’ 18 on page 110 the category label is parenthesized to signal that the analysis provided by the gloss is based on an intended word form rather than on the form actually used.
- In phonetic transcriptions, we do not always adhere to the International Phonetic Alphabet. Instead, we use Czech spelling that would yield the desired pronunciation.

Appendix B

The Czech language

The Czech language is one of the West Slavic languages. It is spoken by slightly more than 10 million speakers, mostly in the Czech Republic. Here we discuss properties of morphology and syntax of the language relevant to our work. For a more detailed discussion, see for example (Karlík, Nekula, and Rusínová 1996; Petr 1987). Unfortunately, there is no detailed grammar of Czech in English. Overviews can be found in (Naughton 2005; Short 1993; Janda and Townsend 2002; Fronek 2007; Harkins 1953).

For historical reasons, there are two variants of Czech: a prescriptive variant, called Official, Literary, or Standard Czech (*spisovná čeština*, hence SCz), and a variant used by most speakers in everyday spoken communication especially in the Western parts of the country, called Common or Colloquial Czech (*obecná or hovorová čeština*, hence CCz). While CCz has developed into an interdialect, incorporating most other dialects of Czech except for most Moravian dialects in the Eastern part of the country, SCz is based on a 19th-century resurrection of 16th-century Czech. Sometimes it is claimed, with some exaggeration, that it is the first foreign language Czechs learn, and the linguistic situation in the Czech Republic is viewed as a case of diglossia. The differences between SCz and CCz are mainly in morphology and lexicon. The two variants are influencing each other, resulting in a significant amount of irregularity, especially in morphology.¹

¹There is no consensus about the interpretation of the linguistic situation in the Czech Republic, which could only be sketched here with a gross simplification. See, e.g., Bermel (2000), Sgall et al. (1992), and Sgall and Hronek (1992).

B.1 Morphology

Like most other Slavic languages, Czech is richly inflected. Czech morphology is important in determining the grammatical functions of phrases. As [Table B.1](#) shows, inflectional morphs are highly ambiguous. There are three genders: neuter, feminine and masculine. The masculine gender further distinguishes the subcategory of animacy. This view is often simplified by treating masculine animate and masculine inanimate categories as separate genders. In addition to singular and plural, some dual number forms survive in body parts nouns and modifiers agreeing with them.² There are seven cases: nominative, genitive, dative, accusative, vocative, locative, instrumental. Only nouns, only in singular, and only about half of the paradigms have a special form for vocative, otherwise the vocative form is the same as nominative.

B.1.1 Nouns

Traditionally, Czech grammars distinguish 13 basic noun paradigms: 4 neuter, 3 feminine, 4 masculine animate and 2 masculine inanimate; plus there are nouns with adjectival declension (other 2 paradigms). In addition, there many subparadigms and subsubparadigms. All of this involves a great amount of variation and irregularity. As an illustration, [Table B.2](#) shows the declension patterns of several nouns.

B.1.2 Adjectives

Adjectives follow two declension paradigms: *hard* and *soft*. Both of them are highly ambiguous, populating the 60 possible combinations of categories of non-negated positive grade adjectives (4 genders \times (2 numbers \times 7 cases + 1 dual form)) with only 12 forms (hard declension) or 8 forms (soft declension). In CCz it is even less: 10 forms (hard) and 8 forms (soft). See [Table B.3](#) for the hard paradigm and [Table B.4](#) for the soft one.

Negation and comparison forms are expressed morphologically. Negation by the prefix *ne-*, comparative by the suffix *-(e)jší-* and superlative by adding the prefix

²There is no dual in CCz. The CCz plural forms are the same as the SCz dual forms. For example, SCz: *velkýma rukama* ‘big.FEM.DL.INS hands.FEM.DL.INS’ vs. *velkými lžícemi* ‘big.FEM.PL.INS spoons.FEM.PL.INS’ (there is no ‘hands.FEM.PL.INS’ or ‘spoons.FEM.DL.INS’); CCz: *velkejma rukama* ‘big.FEM.PL.INS hands.FEM.PL.INS’ vs. *velkejma lžícema* ‘big.FEM.PL.INS spoons.FEM.PL.INS’ (according to *ORAL 2006* (Kopřivová and Waclawičová 2006), *-ejma* ending is the most frequent accounting for 82% of 263 tokens, *-ými* for 8% *-ýma* for 10%, and **ejmi* has no occurrences).

Form	Lemma	Gloss	Category
<i>měst-a</i>	MĚSTO	town	NOUN NEUT SG GEN NOUN NEUT PL NOM (VOC) NOUN NEUT PL ACC
<i>tém-a</i>	TÉMA	theme	NOUN NEUT SG NOM (VOC) NOUN NEUT SG ACC
<i>žen-a</i>	ŽENA	woman	NOUN FEM SG NOM
<i>pán-a</i>	PÁN	man	NOUN MASC-ANIM SG GEN NOUN MASC-ANIM SG ACC
<i>ostrov-a</i>	OSTROV	island	NOUN MASC-INANIM SG GEN
<i>předsed-a</i>	PŘEDSEDA	president	NOUN MASC-ANIM SG NOM
<i>vidě-l-a</i>	VIDĚT	see	VERB PAST PARTICIPLE FEM SG VERB PAST PARTICIPLE NEUT PL
<i>vidě-n-a</i>			VERB PASSIVE PARTICIPLE FEM SG VERB PASSIVE PARTICIPLE NEUT PL
<i>vid-a</i>			VERB TRANSGRESSIVE MASC SG
<i>dv-a</i>	DVA	two	NUMERAL MASC SG NOM NUMERAL MASC SG ACC

Table B.1: Homonymy of the *-a* ending.

	NEUT 'Monday'	FEM 'song'	FEM 'fly'	M.ANIM 'Jirka'	M.ANIM 'brother'	M.INAN 'castle'
NOM SG	<i>pondělí</i>	<i>píseň</i>	<i>moucha</i>	<i>Jirka</i>	<i>bratr</i>	<i>hrad</i>
GEN SG	<i>pondělí</i>	<i>písně</i>	<i>mouchy</i>	<i>Jirky</i>	<i>bratra</i>	<i>hradu</i>
DAT SG	<i>pondělí</i>	<i>písni</i>	<i>mouše</i>	<i>Jirkovi</i>	<i>bratru/ovi</i>	<i>hradu</i>
ACC SG	<i>pondělí</i>	<i>píseň</i>	<i>mouchu</i>	<i>Jirku</i>	<i>bratra</i>	<i>hrad</i>
VOC SG	<i>pondělí</i>	<i>písni</i>	<i>moucho</i>	<i>Jirko</i>	<i>bratře</i>	<i>hrade</i>
LOC SG	<i>pondělí</i>	<i>písni</i>	<i>mouše</i>	<i>Jirkovi</i>	<i>bratru/ovi</i>	<i>hradu</i>
INS SG	<i>pondělím</i>	<i>písni</i>	<i>mouchou</i>	<i>Jirkou</i>	<i>bratrem</i>	<i>hradem</i>
NOM PL	<i>pondělí</i>	<i>písně</i>	<i>mouchy</i>	<i>Jirkové</i>	<i>bratři/ové</i>	<i>hrady</i>
GEN PL	<i>pondělí</i>	<i>písni</i>	<i>much</i>	<i>Jirků</i>	<i>bratrů</i>	<i>hradů</i>
DAT PL	<i>pondělí</i>	<i>písním</i>	<i>mouchám</i>	<i>Jirkům</i>	<i>bratrům</i>	<i>hradům</i>
ACC PL	<i>pondělí</i>	<i>písně</i>	<i>mouchy</i>	<i>Jirky</i>	<i>bratry</i>	<i>hrady</i>
VOC PL	<i>pondělí</i>	<i>písně</i>	<i>mouchy</i>	<i>Jirkové</i>	<i>bratři</i>	<i>hrady</i>
LOC PL	<i>pondělích</i>	<i>písních</i>	<i>mouchách</i>	<i>Jircích*</i>	<i>bratřích*</i>	<i>hradech</i>
INS PL	<i>pondělími*</i>	<i>písněmi*</i>	<i>mouchami*</i>	<i>Jirky*</i>	<i>bratry*</i>	<i>hrady*</i>

Table B.2: Examples of declined nouns; forms marked by * are only used in SCz

	SCz				CCz			
	M.ANIM	M.INAN	NEUT	FEM	M.ANIM	M.INAN	NEUT	FEM
NOM SG		<i>mladý</i>	<i>mladé</i>	<i>mladá</i>		<i>mladej</i>	<i>mladý</i>	<i>mladá</i>
GEN SG		<i>mladého</i>		<i>mladé</i>		<i>mladýho</i>		<i>mladý</i>
DAT SG		<i>mladému</i>		<i>mladé</i>		<i>mladýmu</i>		<i>mladý</i>
ACC SG	<i>mladého</i>	<i>mladý</i>	<i>mladé</i>	<i>mladou</i>	<i>mladýho</i>	<i>mladej</i>	<i>mladý</i>	<i>mladou</i>
VOC SG		<i>mladý</i>	<i>mladé</i>	<i>mladá</i>		<i>mladej</i>	<i>mladý</i>	<i>mladá</i>
LOC SG		<i>mladém</i>		<i>mladé</i>		<i>mladým</i>		<i>mladý</i>
INS SG		<i>mladým</i>		<i>mladou</i>		<i>mladým</i>		<i>mladou</i>
NOM PL	<i>mladí</i>	<i>mladé</i>	<i>mladá</i>	<i>mladé</i>		<i>mladý*</i>		
GEN PL		<i>mladých</i>				<i>mladých</i>		
DAT PL		<i>mladým</i>				<i>mladým</i>		
ACC PL		<i>mladé</i>	<i>mladá</i>	<i>mladé</i>		<i>mladý*</i>		
VOC PL	<i>mladí</i>	<i>mladý</i>	<i>mladá</i>	<i>mladé</i>		<i>mladý*</i>		
LOC PL		<i>mladých</i>				<i>mladých</i>		
INS PL		<i>mladými</i>				<i>mladými</i>		
INS DL		<i>mladýma</i>						

Table B.3: Hard adjectival paradigm; forms marked by * can also be *mladé* in the neuter, and to some extent also in the feminine gender

nej- to the comparative. The comparative and superlative forms are declined as soft adjectives.

B.1.3 Pronouns

Some pronouns have nominal declension, some have adjectival declension and some have their own (e. g., *já* ‘I’). Selected forms of personal pronouns are listed in Table B.5.

B.1.4 Numerals

Only *jeden* ‘1’, *dva* ‘2’, *tři* ‘3’, and *čtyři* ‘4’ fully decline, all of them distinguishing case and *jeden* and *dva* also gender. The inflection of the other cardinal numerals is limited to distinguishing oblique and non-oblique forms. Numerals expressing hundreds and thousands have in certain categories a choice between an undeclined numeral form or a declined noun form (*sto dvaceti*, *sta dvaceti* ‘120.GEN’). Ordinal complex numerals have all parts in the ordinal form and fully declining (*dvacátý*

	M.ANIM	M.INAN	NEUT	FEM
NOM SG		<i>jarní</i>		
GEN SG		<i>jarního</i>		<i>jarní</i>
DAT SG		<i>jarnímu</i>		<i>jarní</i>
ACC SG	<i>jarního</i>		<i>jarní</i>	
VOC SG		<i>jarní</i>		
LOC SG		<i>jarním</i>		<i>jarní</i>
INS SG		<i>jarním</i>		<i>jarní</i>
NOM PL		<i>jarní</i>		
GEN PL		<i>jarních</i>		
DAT PL		<i>jarním</i>		
ACC PL		<i>jarní</i>		
VOC PL		<i>jarní</i>		
LOC PL		<i>jarních</i>		
INS PL		<i>jarními</i>		
INS DL		<i>jarníma</i>		

Table B.4: Soft adjectival paradigm

pátý ‘25th’).³ Similarly as in German, two-digit numerals may have an inverted one-word form (*pěťadvacet* ‘25’, lit.: five-and-twenty, *pěťadvacátý* ‘25th’).

B.1.5 Verbs

As in all Slavic languages, verbs distinguish aspect – perfective and imperfective. Aspect is usually marked by prefixes, sometimes suffixes or by suppletion. Change of aspect is usually accompanied by a change, often subtle, in lexical meaning. For example, *psát* ‘write.IMP’, *napsat* ‘write.PERF’, *dopsat* ‘finish writing.PERF’, *sepsat* ‘write up.PERF’, *sepisovat* ‘write up.IMP’, etc. For more information on Czech aspect see, e. g., Filip (1999).

There are three tenses: present, past and future. Present tense of imperfective verbs and future tense of perfective verbs is marked inflectionally, distinguishing number and person. Perfective verbs do not have present tense. The conjugations of perfective future and imperfective present are the same; sometimes they are subsumed under the morphological present tense. Past tense and imperfective future is

³Again, this is the case of SCz, complex numerals in CCz usually have only their tens and units in ordinal forms.

	NOM	DAT			GEN/ACC		
		weak	either	strong	weak	either	strong
1SG	<i>já</i>	<i>mí</i>	<i>mně</i> [mpɛ]			<i>mě</i> [mpɛ]	<i>mne*</i>
2SG	<i>ty</i>	<i>ti</i> [ci]		<i>tobě</i> [tobjɛ]	<i>tě</i> [cɛ]		<i>tebe</i>
3SG M	<i>on</i>	<i>mu</i>		<i>jemu</i>	<i>ho</i> <i>jej*</i>		<i>jeho</i>
3SG N	<i>ono</i>				<i>ho</i> <i>jej*</i> <i>je.ACC</i>		
3SG F	<i>ona</i>		<i>jí</i> [ji:]			<i>jí</i> [ji]	
1PL	<i>my</i>		<i>nám</i>			<i>nás</i>	
2PL	<i>vy</i>		<i>vám</i>			<i>vás</i>	
3PL M	<i>oni</i>		<i>jím</i>			<i>jich.GEN</i>	
3PL N	<i>ona</i>					/	
3PL F	<i>ony</i>					<i>je.ACC</i>	

Table B.5: Personal pronouns: selected forms; * – rare; *je.ACC* – only in accusative, *jich.GEN* – only in genitive

expressed periphrastically.⁴ Sample conjugations are in Table B.6. In CCz, gender distinction in plural past participles is lost, all being pronounced as SCz feminine plural form. Also Common Czech uses adjectives instead of passive participles. Pluperfect is rare and an aorist is absent in modern Czech.

Five main conjugational types are recognized. Each class has several, quite similar, paradigms (6, 3, 2, 3, 1; 15 in total). Certain categories are expressed analytically; various forms of the verb *být* serve as the auxiliary. Some of the auxiliary forms are constant or inconstant clitics (see §B.3).

⁴Except for 3rd person past tense, where there is no auxiliary. However, some linguists (Veselovská 1995) assume a phonologically null auxiliary.

	‘to be’	‘lubricate’	‘say please’	‘do/make’	‘do/make’
		IMPF	IMPF	IMPF	PERF
INF	<i>být</i>	<i>mazat</i>	<i>prosit</i>	<i>dělat</i>	<i>udělat</i>
PRESENT					
1.SG	<i>jsem</i>	<i>mažu</i>	<i>prosím</i>	<i>dělám</i>	<i>udělám</i>
2.SG	<i>jsi</i>	<i>mažeš</i>	<i>prosíš</i>	<i>děláš</i>	<i>uděláš</i>
3.SG	<i>je</i>	<i>maže</i>	<i>prosí</i>	<i>dělám</i>	<i>udělám</i>
1.PL	<i>jsme</i>	<i>mažeme</i>	<i>prosíme</i>	<i>děláme</i>	<i>uděláme</i>
2.PL	<i>jste</i>	<i>mažete</i>	<i>prosíte</i>	<i>děláte</i>	<i>uděláte</i>
3.PL	<i>jsou</i>	<i>mažou</i>	<i>prosí</i>	<i>dělají</i>	<i>udělají</i>
PAST PRTCP					
MASC SG	<i>byl</i>	<i>mazal</i>	<i>prosil</i>	<i>dělal</i>	<i>udělal</i>
FEM SG	<i>byla</i>	<i>mazala</i>	<i>prosila</i>	<i>dělala</i>	<i>udělala</i>
NEUT SG	<i>bylo</i>	<i>mazalo</i>	<i>prosilo</i>	<i>dělalo</i>	<i>udělalo</i>
M.ANIM PL	<i>byli</i>	<i>mazali</i>	<i>prosili</i>	<i>dělali</i>	<i>udělali</i>
FEM/M.INAN PL	<i>byly</i>	<i>mazaly</i>	<i>prosily</i>	<i>dělaly</i>	<i>udělaly</i>
NEUT PL	<i>byla</i>	<i>mazala</i>	<i>prosila</i>	<i>dělala</i>	<i>udělala</i>
PASS PRTCP					
MASC SG	–	<i>mazán</i>	<i>prosen</i>	<i>dělán</i>	<i>udělán</i>
FEM SG	–	<i>mazána</i>	<i>prosená</i>	<i>dělána</i>	<i>udělána</i>
NEUT SG	–	<i>mazáno</i>	<i>proseno</i>	<i>děláno</i>	<i>uděláno</i>
M.ANIM PL	–	<i>mazáni</i>	<i>proseni</i>	<i>děláni</i>	<i>uděláni</i>
FEM/M.INAN PL	–	<i>mazány</i>	<i>proseny</i>	<i>dělány</i>	<i>udělány</i>
NEUT PL	–	<i>mazána</i>	<i>prosená</i>	<i>dělána</i>	<i>udělána</i>
IMPERATIVE					
2.SG	<i>buď</i>	<i>maž</i>	<i>pros</i>	<i>dělej</i>	<i>udělej</i>
1.PL	<i>buďme</i>	<i>mažme</i>	<i>proste</i>	<i>dělejme</i>	<i>udělejme</i>
2.PL	<i>buďte</i>	<i>mažte</i>	<i>prosme</i>	<i>dělejte</i>	<i>udělejte</i>

Table B.6: Sample verbal paradigms (SCz)

B.2 Syntax

B.2.1 Agreement

In Czech, there is agreement between subject and predicate and agreement within the NP. Below, we provide a basic overview; for a detailed description of Czech agreement see (Avgustinova et al. 1995).

B.2.1.1 Subject-predicate agreement

Two types of agreement with subject can be distinguished:

- Subject – finite verb agreement

The finite verb agrees with the subject in person and number.

- (44) *Střední Evropa je/*jsem/*jsou ve vzduchoprázdnu.*
 Central Europe is.3SG/am/are.3PL in vacuum
 ‘Central Europe is in vacuum.’ (CNC – SYN2006PUB)

- Subject – participles/predicative adjectives agreement

Predicative adjectives and participles in periphrastic constructions agree in number and gender with subject. In (45), the dropped 2nd person singular (and masculine since referring to *Oto*) subject agrees with the participle *byl* and adjective *zavřený* in number and gender. Similarly *služba* in (46) agrees with *povinná* in number and gender.

- (45) *Oto, za co jsi byl/*byla/*byli*
 Ota.M.SG, for what aux.2SG was.M.SG/was.F.SG/was.M.PL
*zavřený/*zavřená/*zavření?*
 jailed.M.SG/jailed.F.SG/jailed.M.PL
 ‘Ota, what were you jailed for?’ (CNC – SYN2006PUB)
- (46) *Vojenská služba je ve Švédsku povinná.*
 Military.FEM.SG service.FEM.SG is in Sweden obligatory.FEM.SG
 ‘Military service is obligatory in Sweden.’ (CNC – SYN2006PUB)

Only SCz distinguishes gender for plural participles (see Table B.6). In spelling, there are three forms: *chrápali* [-li] ‘snored.M.PL’, *chrápaly* [-li] ‘snored.F/I.PL’, *chrápala* [-la] ‘snored.N.PL’ (note that *chrápali* and *chrápaly* have

the same pronunciation). CCz uses the [-li] form for all genders in plural (spelling is unclear). Plural adjectives pattern similarly (see §B.1.2).

Non-nominative subjects In case of non-nominative subjects (certain numeric expression (47a),⁵ (47b), etc.) and constructions that are traditionally analyzed as subject-less (47c or 47d), the predicate is in 3rd person singular neuter form.

- (47) a. *Pět/Mnoho lodí* *zmizelo.*
 five/many ships.FEM.PL.GEN disappeared.NEUT.SG
 ‘Five/Many ships disappeared’
- b. *Otevřít soubor je jednoduché.*
 open.INF file.INAM.SG is.3SG simple.NEUT.SG
 ‘To open a file is easy.’
- c. *Prší/Pršelo.*
 rains.3SG/rained.NEUT.SG
 ‘It is/was raining.’
- d. *Je mi příjemně.*
 is.3SG me.DAT fine.ADVERB
 ‘I am feeling fine.’

Coordinated subjects Agreement with coordinated subjects is rather complex. The gender of the predicate is the minimal gender of participants of coordination, computed under the following order: *masc.animate* < {*masc.inanimate*, *feminine*} < *neuter*. This covers also the trivial case when the gender of all participants is the same. However, there is an exception: if all participants have neuter gender and at least one is in singular then the gender of the predicate is feminine. This complexity is absent in CCz because colloquial plural participles and adjectives do not distinguish gender. There is a similar hierarchy for determining person of subject with heterogenous persons. Under certain conditions (especially when the predicate precedes the subject, or the subject consists of abstract nouns), the predicate can agree only with the member of the coordinated subject it is closest – as (48c) and (48d) show.

⁵In similar phrases, the noun in genitive is traditionally seen as the head. We could also assume the numeral to be the head. In such a case, it would be natural to assume the numeral is in the default form (neuter singular).

- (48) a. Two concrete nouns:

Byl jsem rád, že máma s tátou byli/byla
 was.M.SG aux.1SG happy, that mom with dad were.M.PL/were.F.SG
v pořádku.
 fine

‘I was happy that mom and dad were fine.’ (CNC – SYN2006PUB)

- b. *Hitler a Německo už měli hotové plány na*
 Hitler.M.SG and Germany.N.SG already had.M.PL finished.ACC plans.ACC for
znovuzískání Horního Slezska ..
 reclaiming Upper Silesia ..

‘Hitler and Germany already had finished plans for reclaiming Upper Silesia ...’
 (CNC – SYN2005)

- c. Two abstract nouns:

Přesnost a srozumitelnost je příznačná / jsou
 Accuracy.FEM.SG and comprehensibility.FEM.SG is typical.FEM.SG / are
příznačné pro jeho výklady.
 typical.FEM.PL for his explanations.

‘Accuracy and comprehensibility are typical for his explanations.’ (Karlík, Nekula, and Rusínová 1996)

- d. Verb preceding subject:

Včera přišla / přišli máma a táta domů brzo.
 Yesterday came.FEM.SG / came.FEM.PL mom and dad home early.

‘Yesterday came mom and dad home early.’

B.2.1.2 Agreement within the NP

So called *agreeing attributes* agree with the noun in gender, number and case. This includes

- Normal adjectives such as *starý* ‘old’: for example, in (46) the adjective *vojenská* ‘military.FEM.SG’ agrees with the noun *služba* ‘service.FEM.SG’.

- Possessive adjectives such as *otcův* ‘father’s’⁶
- Relative clauses: the relative pronoun agrees with the modified noun only in gender and number; its case is dependent on its function in the relative clause. In CCz, relative clauses are often introduced by a universal nondeclined relative pronoun *co*. The pronoun *jenž* ‘that’ is also often not declined.
- Ordinal numerals
- Possessive pronouns and various determiners

Note that there are some limited exceptions. For historical reasons, attributes modifying accusative or nominative pronouns like *nic* ‘nothing’ or *něco* ‘something’ are in genitive as in (49).

- (49) *Nikdo nechtěl říci nic konkrétního.*
 Nobody not-wanted say.INF nothing.NEUT.SG.ACC concrete.NEUT.SG.GEN
 ‘Nobody wanted to say anything concrete’

In nominative or vocative, the gender can be feminine even when the noun is not, this gives the phrase an expressive flavor as in (50).

- (50) *Kluku líná!*
 Boy.MASC.SG.VOC lazy.FEM.SG.VOC
 ‘You lazy boy!’

B.2.2 Numeral expressions

Numerals expressions with *jeden* ‘1’, *dva* ‘2’, *tři* ‘3’, *čtyři* ‘4’, *oba* ‘both’ behave in a “normal” way: a numeral agrees with its noun in case; *jeden*, *dva* and *oba* also in gender. However, numerals *pět* and above in nominative or accusative positions are followed by nouns in genitive plural (47a). Otherwise (other numerals or other cases), the noun is in the same case as the whole phrase.

⁶However, in the dialects of Southern and Western Bohemia, possessive adjectives do not decline. The form ending in *-ovo* (for masculine possessors) or *-ino* for feminine possessors is used regardless of case, number and gender of the possessed noun. In other dialects, this form is used only for accusative singular. However, the dialects of Southern and Western Bohemia also often use prenominal genitive to express possession instead, especially when the possessive adjective would involve a phonological change: *s Hanky kolem* ‘with Hanka.GEN bike.MASC.INAN.SG.INS’ or *s Hančino kolem* ‘with Hanka’s bike.MASC.INAN.SG.INS’ for SCz *s Hančíným kolem* ‘with Hanka’s.MASC.INAN.SG.INS bike.MASC.INAN.SG.INS’.

B.2.3 Negation

Sentence negation in Czech is formed by the prefix *ne-* attached to the verb.⁷ As in the other Slavic languages, multiple negation is the rule, negative subject or object pronouns, adjectival pronouns and adverbs combine with negative verbs.

- (51) *Nikdy nikomu nic neslibuj.*
 never nobody.DAT nothing.ACC not-promise.IMPER.2SG
 ‘Never promise anything to anybody.’

B.3 Word order and clitics

Czech has exceptionally free word order. Unlike English, where word order is mostly fixed and is mainly used to express grammatical functions, word order in Czech is used to express information structure.

Similarly to most other Slavic languages, Czech contains second-position clitics. Syntactically, they are enclitics, following their host, a certain clause-initial unit, usually the first constituent. Their relative ordering is very restricted. In (52), *bych* ‘would.1SG’, *mu* ‘him.DAT’ and *to* ‘it.ACC’ are all clitics. While the other words can be rearranged fairly freely and still form a grammatical sentence, the clitics have to follow the first constituent and appear in this particular order.

- (52) *Příští sobotu bych mu to mohl dát.*
 Next Saturday would.1SG him.DAT it.ACC could give.INF
 ‘Next Saturday, I could give it to him.’

The set of clitics includes:

- some auxiliaries: for example, the conditional auxiliary is a clitic, but the future auxiliary is not; passive auxiliary can but need not be a clitic.
- non-strong personal pronouns (see §B.5) and weak reflexives
- certain other short words: *to* ‘it’, *tu* ‘here’, etc.

For more details, see, for example, Avgustinova and Oliva (1995), Rosen (2001), and Hana (2007).

⁷Traditionally, *ne-* is classified as prefix, although it is rather a proclitic.

B.4 Romani ethnolect of Czech

In addition to non-native Czech, there is also the Romani ethnolect of Czech, used by some speakers with Romani background. Although their Czech is nearly always their first language, it differs from SCz (Bořkovcová 2007; Šotolová 2008). Despite the differences in comparison to the acquisition of Czech by foreigners, some linguistic issues involved in educating young speakers of the Romani ethnolect in Czech schools (hence Roma pupils) suggest that using a methodology developed in this project for non-native speakers and adapted for Roma pupils is justified.

Sometimes it is difficult to decide whether Czech is the first or second language of the Roma pupils. Bedřichová et al. (2011) assume that the social, cultural and linguistic differences between the non-Roma majority and some Roma communities may imply specific language development of Roma children and show some traits of L2 acquisition. Because their linguistic integration represents a significant issue in the country's education system, this part of the *CzeSL* corpus became a separate component of *AKCES*.

Bibliography

- Abuhakema, Ghazi, Anna Feldman, and Eileen Fitzpatrick. 2009. "ARIDA: An Arabic Interlanguage Database and Its Applications: A Pilot Study." *Journal of the National Council of Less Commonly Taught Languages (NCOLCTL)* 7:161–184. <https://www.aai.org/Papers/FLAIRS/2008/FLAIRS08-055.pdf>.
- Aharodnik, Katsiaryna, Marco Chang, Anna Feldman, and Jirka Hana. 2013. "Automatic Identification of Learners' Language Background based on their Writing in Czech." In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJNCLP 2013), Nagoya, Japan, October 2013*, 1428–1436. <https://www.aclweb.org/anthology/I13-1200/>.
- Altenberg, Bengt, and Marie Tapper. 1998. "The use of adverbial connectors in advanced Swedish learner's written English." In *Learner English on Computer*, edited by Sylviane Granger, 80–93. London: Longman.
- Amaral, Luiz, and Detmar Meurers. 2008. "From Recording Linguistic Competence to Supporting Inferences about Language Acquisition in Context: Extending the Conceptualization of Student Models for Intelligent Computer-Assisted Language Learning." *Computer-Assisted Language Learning* 21 (4): 323–338. <http://www.sfs.uni-tuebingen.de/~dm/papers/Amaral.Meurers-08pub.pdf>.
- Avgustinova, Tania, Alla. Bémová, Eva Hajičová, Karel Oliva, Jarmila Panevová, Vladimír Petkevič, Petr Sgall, and Hana Skoumalová. 1995. *Linguistic problems of Czech. Project Peco 2924*. Technical report. Prague: Charles University.
- Avgustinova, Tania, and Karel Oliva. 1995. *The Position of Sentential Clitics in the Czech Clause*. CLAUS Report 68. Universität des Saarlandes, December.

- Bedřichová, Zuzanna, Karel Šebesta, Svatava Škodová, and Kateřina Šormová. 2011. "Podoba a využití korpusu jinojazyčných a romských mluvčích češtiny: CZESL a ROMi." In *Korpusová lingvistika Praha 2011: 2 – Výzkum a výstavba korpusů*, edited by František Čermák, 15:93–104. Studie z korpusové lingvistiky. Praha: Ústav Českého národního korpusu, Nakladatelství Lidové noviny.
- Bermel, Neil. 2000. *Register variation and language standards in Czech*. LINCOS Studies in Slavic Linguistics 13. München: Lincom GmbH.
- Bley-Vroman, Robert. 1983. "The comparative fallacy in interlanguage studies: the case of systematicity." *Language Learning – A Journal of Research in Language Studies* 33 (1). <https://doi.org/10.1111/j.1467-1770.1983.tb00983.x>.
- Borin, Lars, Markus Forsberg, and Johan Roxendal. 2012. "Korp — the corpus infrastructure of Språkbanken." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 474–478. Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/248_Paper.pdf.
- Boržkovicová, Máša. 2007. *Romský etnolekt češtiny*. Praha: Signeta.
- Boušková, Petra, Václav Jonáš, Yana Leontiyeva, Radim Křištof, Dana Lindová, Jarmila Marešová, Michaela Maršíková, et al. 2019. *Foreigners in the Czech Republic*. Yearbook 19. Praha: Czech Statistical Office, December. <https://www.czso.cz/documents/10180/91605941/29002719.pdf>.
- Boyd, Adriane, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. "The MERLIN corpus: Learner Language and the CEFRL." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 1281–1288. Reykjavik, Iceland: European Language Resources Association (ELRA). <https://www.aclweb.org/anthology/L14-1488/>.
- Bykh, Serhiy, and Detmar Meurers. 2012. "Native Language Identification using Recurring *n*-grams – Investigating Abstraction and Domain Dependence." In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, 425–440. Mumbai, India: The COLING 2012 Organizing Committee. <http://www.aclweb.org/anthology/C12-1027>.

- Čermák, František. 2005. "Korpus, informace a lingvistika." In *Přednášky z XLVIII. běhu Letní školy slovanských studií*, edited by Jiří Hasil, 15–24. Praha: Filozofická fakulta Univerzity Karlovy.
- Čermák, František, and Michal Křen. 2004. *Frekvenční slovník češtiny*. Praha: Nakladatelství Lidové noviny.
- Chiarcos, Christian. 2012. "Interoperability of Corpora and Annotations." In *Linked Data in Linguistics – Representing and Connecting Language Data and Language Metadata*, edited by Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, 161–179. Springer.
https://doi.org/10.1007/978-3-642-28249-2_16.
- Christ, Oliver. 1994. "A modular and flexible architecture for an integrated corpus query system." In *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research*, 23–32. Budapest, Hungary.
<http://cwb.sourceforge.net/files/Christ1994.pdf>.
- Christ, Oliver, Bruno M. Schulze, Anja Hofmann, and Esther König. 1999. *The IMS Corpus Workbench: Corpus Query Processor (CQP) – User's Manual*. Technical report. University of Stuttgart.
<http://corpora.dslo.unibo.it/TCORIS/cqpman.pdf>.
- Czech National Corpus – SYN2000*. 2000. Praha. <http://korpus.cz>.
- Czech National Corpus – SYN2005*. 2005. Prague. <http://korpus.cz>.
- Czech National Corpus – SYN2006PUB*. 2006. Prague. <http://korpus.cz>.
- Cohen, Jacob. 1960. "A coefficient of agreement for nominal scales." *Educational and Psychological Measurement* 20 (1): 37–46.
<https://doi.org/10.1177/001316446002000104>.
- Corder, S. P. 1967. "The Significance of Learner's Errors." *International Review of Applied Linguistics in Language Teaching* (Berlin, Boston) 5 (1–4): 161–170.
<https://doi.org/10.1515/iral.1967.5.1-4.161>.
- Cvrček, Václav. 2010. *Mluvnice současné češtiny*. Vol. I. Praha: Karolinum.
- Dahlmeier, Daniel, and Hwee Tou Ng. 2012. "Better Evaluation for Grammatical Error Correction." In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 568–572. Montréal, Canada: Association for Computational Linguistics. <https://www.aclweb.org/anthology/N12-1067>.

- Díaz-Negrillo, Ana, and Jesús Fernández-Domínguez. 2006. "Error Tagging Systems for Learner Corpora." *Resla* 19:83–102.
<http://icame.uib.no/ij31/ij31-page197-204.pdf>.
- Díaz-Negrillo, Ana, Detmar Meurers, Salvador Valera, and Holger Wunsch. 2010. "Towards interlanguage POS annotation for effective learner corpora in SLA and FLT." *Language Forum* 36 (1–2): 139–154.
<http://www.sfs.uni-tuebingen.de/~dm/papers/diaz-negrillo-et-al-09.pdf>.
- Dickinson, Markus, Ross Israel, and Sun-Hee Lee. 2010. "Building a Korean Web Corpus for Analyzing Learner Language." In *Proceedings of the 6th Workshop on the Web as Corpus (WAC-6)*. Los Angeles.
<https://www.aclweb.org/anthology/W10-1502/>.
- Dickinson, Markus, and Marwa Ragheb. 2013. *Annotation for Learner English Guidelines, v. 0.1 (June 2013)*.
<http://cl.indiana.edu/~salle/resources/salle-guidelines0.1.pdf>.
- Dušková, Libuše. 1969. "On Sources of Errors in Foreign Language Learning." *International Review of Applied Linguistics in Language Teaching (IRAL)* 7 (1): 11–36. <https://doi.org/10.1515/iral.1969.7.1.11>.
- Ellis, Nick C. 2017. "Cognition, Corpora, and Computing: Triangulating Research in Usage-Based Language Learning." *Language Learning* 67:40–65.
<https://doi.org/10.1111/lang.12215>.
- Ellis, Rod. 1994. *The Study of Second Language Acquisition*. Oxford University Press.
- . 2003. *Second Language Acquisition*. Oxford University Press.
- Ellis, Rod, and Gary Barkhuizen. 2005. *Analysing learner language*. Oxford University Press.
- Evert, Stefan, and Andrew Hardie. 2011. "Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium." In *Proceedings of the Corpus Linguistics 2011 Conference*. Birmingham, UK.
<http://www.stefan-evert.de/PUB/EvertHardie2011.pdf>.
- Filip, Hana. 1999. *Aspect, Eventuality Types and Nominal Reference*. New York: Garland Publishing, Taylor & Francis Group, Routledge.

- Fitzpatrick, Eileen, and Steve Seegmiller. 2001. "The Montclair Electronic Language Learner Database." In *Proceedings of the International Conference on Computing and Information Technologies (ICCIT)*, edited by George Antoniou and Dorothy Deremer. Montclair State University, NJ, USA. https://doi.org/10.1142/9789812810885_0046.
- . 2004. "The Montclair electronic language database project." In *Applied Corpus Linguistics: A Multidimensional Perspective*, edited by Ulla Connor and Thomas Albin Upton, 223–238. Rodopi. https://www.researchgate.net/publication/233602196_The_Montclair_Electronic_Language_Database_Project.
- Flor, Michael, and Yoko Futagi. 2012. "On using context for automatic correction of non-word misspellings in student essays." In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 105–115. Montréal, Canada: Association for Computational Linguistics, June. <http://www.aclweb.org/anthology/W12-2012>.
- Fronek, Josef. 2007. *Velký anglicko-český / česko-anglický slovník*. Praha: Leda.
- Granger, Sylviane, ed. 1998a. *Learner English on Computer*. London and New York: Addison Wesley Longman.
- . 1998b. "The computer learner corpus: a versatile new source of data for SLA research." In *Learner English on Computer*, edited by Sylviane Granger, 3–19. Addison Wesley Longman. https://www.researchgate.net/publication/237128463_The_computer_learner_corpus_A_verseatile_new_source_of_data_for_SLA_research.
- . 1999. "Use of Tenses by Advanced EFL Learners: Evidence from Error-tagged Computer Corpus." In *Out of Corpora – Studies in Honour of Stig Johansson*, edited by Hilde Hasselgård and Signe Oksefjell, 26:191–202. Language and Computers. Amsterdam & Atlanta: Rodopi. <http://hdl.handle.net/2078.1/76322>.
- . 2003a. "Error-tagged Learner Corpora and CALL: A Promising Synergy." *CALICO Journal* 20 (3): 465–480. https://www.researchgate.net/publication/228602144_Error-tagged_learner_corpora_and_CALL_A_promising_synergy.

- Granger, Sylviane. 2003b. "The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research." *TESOL Quarterly* 37 (3): 538–546.
<https://www.jstor.org/stable/3588404>.
- . 2008. "Learner Corpora." In *Corpus Linguistics. An International Handbook*, edited by A. Lüdeling and M. Kytö, 1:259–274. HSK 29. 1. Berlin/New York: Mouton De Gruyter.
- . 2017. "Academic phraseology: A key ingredient in successful L2 academic literacy." *Oslo Studies in English* 9 (3): 9–27.
<http://hdl.handle.net/2078.1/201048>.
- Granger, Sylviane, Estelle Dagneaux, and Fanny Meunier. 2002. *International Corpus of Learner English*. Louvain: Presses Universitaires de Louvain.
<https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html>.
- Granger, Sylviane, Gaëtanelle Gilquin, and Fanny Meunier, eds. 2015. *The Cambridge Handbook of Learner Corpus Research*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.
<https://doi.org/10.1017/CBO9781139649414>.
- Günther, Britta, and Herbert Günther. 2007. *Erstsprache, Zweitsprache, Fremdsprache: Eine Einführung*. 2nd. Pädagogik. Weinheim and Basel: Beltz.
- Hajič, Jan. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Prague: Karolinum, Charles University Press.
- Hajič, Jan, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Hajičová, Jiří Havelka, Petr Homola, et al. 2018. *PDT – Prague Dependency Treebank 3.5*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. <http://hdl.handle.net/11234/1-2621>.
- Hana, Jirka. 2007. "Czech Clitics in Higher Order Grammar." PhD diss., The Ohio State University. https://www.researchgate.net/publication/258432348_Czech_Clitics_in_Higher_Order_Grammar.
- Hana, Jirka, and Barbora Hladká. 2018a. "Syntactic annotation of a second-language learner corpus." In *Proceedings of the International Conference on Bilingual Learning and Teaching*. Hong Kong: The Open University of Hong Kong.
https://ufal.mff.cuni.cz/~hladka/2019/docs/hana-hladka-blt-2018_0.pdf.

- . 2018b. “Universal Dependencies and Non-Native Czech.” In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway*, edited by Dag Haug, Stephan Oepen, Lilja Øvrelid, Marie Candito, and Jan Hajič, 105–114. 155. Linköping University Electronic Press, Linköpings universitet.
<http://www.ep.liu.se/ecp/155/011/ecp18155011.pdf>.
- . 2019. *CzeSL – Universal Dependencies Release 0.5*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
<http://hdl.handle.net/11234/1-2927>.
- Harkins, William E. 1953. *A modern Czech grammar*. New York: King’s Crown Press.
- Hercíková, Barbora. 2009. *Přehled základní české gramatiky pro zahraniční studenty*. Praha: Karolinum.
- Hirschmann, Hagen, Anke Lüdeling, Ines Rehbein, Marc Reznicek, and Amir Zeldes. 2013. “Underuse of Syntactic Categories in Falko. A Case Study on Modification.” In *20 years of learner corpus research. Looking back, Moving ahead. Corpora and Language in Use – Proceedings 1*, edited by Sylviane Granger and Fanny Meunier. Louvain la Neuve: Presses universitaires de Louvain.
https://www.researchgate.net/publication/285136742_Underuse_of_syntactic_categories_in_Falko_A_case_study_on_modification.
- Hladká, Barbora, Martin Holub, and Vincent Kříž. 2013. “Feature Engineering in the NLI Shared Task 2013: Charles University Submission Report.” In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 232–241. Atlanta, Georgia, USA: Association for Computational Linguistics.
<http://aclweb.org/anthology/W/W13/W13-1730v2.pdf>.
- Hnátková, Milena, Vladimír Petkevič, and Hana Skoumalová. 2011. “Linguistic Annotation of Corpora in the Czech National Corpus.” In *Trudy meždunarodnoj konferencii Korpusnaja lingvistika – 2011*, 15–20. St.-Petersburg State University, Institute of Linguistic Studies (RAS), Russian State Herzen Pedagogical University.

- Hudousková, Andrea. 2013. "The Corpus CzeSL in the Service of Teaching Czech for Foreigners – Errors in the Use of the Pronoun *kteř*." In *Proceedings of the Seventh International Conference Slovko 2013; Natural Language Processing, Corpus Linguistics, E-learning*, edited by Katarína Gajdošová and Adriána Žáková, 100–107. Lüdenscheid, Germany: RAM-Verlag.
https://korpus.sk/~slovko/2013/Proceedings_Slovko_2013.pdf.
- . 2014. "Jmenné koncovky v češtině pro cizince – distribuce, frekvence a fonetika. První sonda." In *Radost z jazyků. Sborník k 75. narozeninám prof. Františka Čermáka*, edited by Vladimír Petkevič, Ana Adamovičová, and Václav Cvrček, 20:215–230. Studie z korpusové lingvistiky. Praha: Nakladatelství Lidové noviny.
- Izumi, Emi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2005. "Error Annotation for Corpus of Japanese Learner English." In *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora (LINC 2005)*, edited by Kyonghee Paik, Francis Bond, and Stephan Oepen, 71–80. Jeju Island, Korea. <https://www.aclweb.org/anthology/I05-6009/>.
- James, Carl. 1998. *Errors in language learning and use: exploring error analysis*. London and New York: Longman.
- Janda, Laura A., and Charles E. Townsend. 2002. *Czech*. The Slavic and Eurasian Language Resource Center (SEELRC), Duke University, Durham, NC.
<http://www.seelrc.org:8080/grammar/mainframe.jsp?nLanguageID=2>.
- Janssen, Maarten. 2016. "TEITOK: Text-Faithful Annotated Corpora." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, et al., 4037–4043. Paris, France: European Language Resources Association (ELRA).
http://www.lrec-conf.org/proceedings/lrec2016/pdf/651_Paper.pdf.
- . 2018. "TEITOK as a tool for Dependency Grammar." *Procesamiento del Lenguaje Natural* 61:185–188. <https://doi.org/10.26342/2018-61-28>.
- . 2020. "Integrating TEITOK and Kontext at LINDAT." In *CLARIN Annual Conference 2020, 5–7 October 2020*, edited by Costanza Navarretta and Maria Eskevich, 98–101. https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf.

- Jarvis, Scott. 2012. "The Detection-Based approach: An Overview." In *Approaching Language Transfer Through Text Classification: Explorations in the Detection-Based Approach*, edited by Scott Jarvis and Scott A. Crossley, 1–33. Bristol, UK: Multilingual Matters.
- Jelínek, Tomáš. 2017. "Errors in Inflection in Czech as a Second Language and Their Automatic Classification." In *Text, Speech, and Dialogue 20th International Conference, TSD 2017, Prague, Czech Republic, August 27–31, 2017*, edited by Kamil Ekštejn and Václav Matoušek, 263–271. Lecture Notes in Artificial Intelligence series. Springer International Publishing.
- Johns, Tim. 1991. "Should you be persuaded: Two samples of data-driven learning materials." Edited by Tim Johns and Philip King. *Classroom Concordancing. English Language Research Journal* 4:1–16. https://lexically.net/wordsmith/corpus_linguistics_links/Tim%20Johns%20and%20DDL.pdf.
- Kaczmarek, Elżbieta, and Adrian Jan Zasina. 2020. *Infrastructure of the Polish Learner Corpus*. Abstract of a talk presented at the 14th Teaching and Language Corpora conference (TaLC2020), Perpignan. <https://f.hypotheses.org/wp-content/blogs.dir/6396/files/2020/07/Abstracts140720b.pdf>.
- Karlík, Petr, Marek Nekula, and Zdenka Rusínová. 1996. *Příruční mluvnice češtiny*. Praha: Nakladatelství Lidové Noviny.
- Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. "The Sketch Engine: ten years on." *Lexicography* 1 (1): 7–36. https://www.researchgate.net/publication/271848017_The_Sketch_Engine_Ten_Years_On.
- Kopřivová, Marie, and Martina Waclawíčková. 2006. *ORAL2006: korpus neformální mluvené češtiny*. Ústav Českého národního korpusu FF UK, Praha. <http://www.korpus.cz>.
- Leech, Geoffrey. 1998. "Preface." In *Learner English on Computer*, edited by Sylviane Granger, xiv–xx. London: Addison Wesley Longman.
- Leńko-Szymańska, Agnieszka. 2004. "Demonstratives as anaphora markers in advanced learners' English." In *Corpora and Language Learners*, edited by Guy Aston, Silvia Bernardini, and Dominic Stewart, 17:89–107. Studies in Corpus Linguistics. Amsterdam: John Benjamins. <https://www.researchgate.net/publication/275270846>.

- Lennon, Paul. 1991. "Error: Some Problems of Definition, Identification, and Distinction." *Applied Linguistics* 12, no. 2 (June): 180–196.
<https://doi.org/10.1093/applin/12.2.180>.
- Lüdeling, Anke. 2008. "Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora." In *Fortgeschrittene Lernervarietäten*, edited by P. Grommes and M. Walter, 119–140. Tübingen: Niemeyer.
https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiterinnen/anke/pdf/Luedeling_FLV-final.pdf.
- Lüdeling, Anke, and Hagen Hirschmann. 2015. "Error annotation systems." In *The Cambridge Handbook of Learner Corpus Research*, edited by Sylviane Granger, Gaetanelle Gilquin, and Fanny Meunier, 135–158. Cambridge Handbooks in Language and Linguistics. Cambridge University Press. <https://www.researchgate.net/publication/291835319>.
- Lüdeling, Anke, Maik Walter, Emil Kroymann, and Peter Adolphs. 2005. "Multi-level error annotation in learner corpora." In *Proceedings of Corpus Linguistics 2005*. Birmingham.
https://www.researchgate.net/publication/228352566_Multi-Level_Error_Annotation_in_Learner_Corpora.
- Lukšija, Melita. 2009. "Korpus jako zdroj dat při prezentaci předložek do/na s místním směrovým významem ve výuce češtiny pro cizince." Bachelor's Thesis, Masarykova univerzita, Kabinet češtiny pro cizince.
<https://is.muni.cz/th/emun9/>.
- . 2011. "Korpusy a česká deklinace ve výuce češtiny jako cizího jazyka." Master's thesis, Masarykova univerzita, Filozofická fakulta, Ústav českého jazyka. <https://is.muni.cz/th/trima/>.
- Machálek, Tomáš. 2017. "KonText – a modern, customizable corpus query interface." Abstract of a talk presented at the conference Corpus Linguistics 2017, Birmingham. <https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper341.pdf>.
- Marek, Michal, Pavel Pecina, and Miroslav Spousta. 2007. "Web Page Cleaning with Conditional Random Fields." In *Proceedings of the 3rd Web As a Corpus Workshop, Incorporating CLEANVAL*, 155–162. Louvain-la-Neuve, Belgium: UCL Presses Universitaires de Louvain.
<https://ufal.mff.cuni.cz/~pecina/files/cleaneval-2007.pdf>.

- Martins, André, Miguel Almeida, and Noah A. Smith. 2013. "Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 617–622. Sofia, Bulgaria: Association for Computational Linguistics. <https://www.aclweb.org/anthology/P13-2109>.
- McEnergy, Tony. 2018. "Preface." In *Learner Corpus Research: New Perspectives and Applications*, edited by Vaclav Brezina and Lynne Flowerdew, xiv–xvii. London: Bloomsbury Academic. <http://www.bloomsburycollections.com/book/learner-corpus-research-new-perspectives-and-applications/preface-tony-mcenergy/>.
- McEnergy, Tony, Vaclav Brezina, Dana Gablasova, and Jayanti Banerjee. 2019. "Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use." *Annual Review of Applied Linguistics* 39:74–92. <https://doi.org/10.1017/S0267190519000096>.
- Mendes, Amália, Sandra Antunes, Maarten Janssen, and Anabela Gonçalves. 2016. "The COUPLE2 corpus: a learner corpus for Portuguese." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, et al. Paris, France: European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2016/summaries/439.html>.
- Meunier, Fanny. 2019. "Tracking developmental patterns in learner corpora: Focus on longitudinal studies." *Selected papers on theoretical and applied linguistics* 13:34–44. <https://doi.org/10.26262/istal.v23i0.7319>.
- Meurer, Paul. 2012. "Corpuscle – a new corpus management platform for annotated corpora." In *Exploring Newspaper Language: Using the Web to Create and Investigate a large corpus of modern Norwegian*, edited by Gisle Andersen, 29–50. Studies in Corpus Linguistics 49. John Benjamins. <https://doi.org/10.1075/scl.49.02meu>.
- Meurers, Detmar. 2009. "On the Automatic Analysis of Learner Language: Introduction to the Special Issue." *CALICO Journal* 26 (3): 469–473. <http://www.sfs.uni-tuebingen.de/~dm/papers/meurers-09.pdf>.

- Meurers, Detmar. 2015. "Learner Corpora and Natural Language Processing." In *The Cambridge Handbook of Learner Corpus Research*, edited by Gaëtanelle Gilquin Sylviane Granger and Fanny Meunier, 537–566. Cambridge University Press.
<http://www.sfs.uni-tuebingen.de/~dm/papers/meurers-15.pdf>.
- Náplava, Jakub. 2017. "Natural Language Correction." Master's thesis, Charles University, Faculty of Mathematics and Physics.
<https://is.cuni.cz/webapps/zzp/detail/188260/>.
- Náplava, Jakub, and Milan Straka. 2019. "Grammatical Error Correction in Low-Resource Scenarios." In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, 346–356. Stroudsburg, PA, USA: Association for Computational Linguistics.
<https://www.aclweb.org/anthology/D19-5545/>.
- Naughton, James. 2005. *Czech: An Essential Grammar*. Oxon, Great Britain and New York, NY, USA: Routledge.
- Nesselhauf, Nadja. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.
- Nicholls, Diane. 2003. "The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT." In *Proceedings of the Corpus Linguistics 2003 Conference*, edited by Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery, 572–581. Lancaster, UK: Lancaster University: University Center for Computer Corpus Research on Language.
<http://ucrel.lancs.ac.uk/publications/cl2003/papers/nicholls.pdf>.
- Novák, Michal, Jiří Mírovský, Kateřina Rysová, and Magdaléna Rysová. 2019. "Exploiting Large Unlabeled Data in Automatic Evaluation of Coherence in Czech." In *Text, Speech, and Dialogue*, edited by Kamil Ekštejn, 197–210. Cham: Springer International Publishing.
- Novák, Michal, Kateřina Rysová, Jiří Mírovský, Magdaléna Rysová, and Eva Hajičová. 2017. *EVALD 2.0 for Foreigners*. Data/software. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. <http://hdl.handle.net/11234/1-2509>.

- Novák, Michal, Kateřina Rysová, Magdaléna Rysová, and Jiří Mírovský. 2017. "Incorporating Coreference to Automatic Evaluation of Coherence in Essays." In *Statistical Language and Speech Processing*, 58–69. Cham, Switzerland: Springer International Publishing.
- Osolobě, Klára. 2010. "Jak se učit česky s korpusem." In *Přednášky a besedy z XLIII. běhu Letní škola slovanských (bohemistických) studií (LŠSS)*, 112–119. Brno: Masarykova univerzita. Kabinet češtiny pro cizince.
- Pajas, Petr. 2009. *TrEd*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. <http://hdl.handle.net/11858/00-097C-0000-0001-48F7-8>.
- Pajas, Petr, and Jan Štěpánek. 2006. "XML-Based Representation of Multi-Layered Annotation in the PDT 2.0." In *Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006)*, edited by Richard Erhard Hinrichs, Nancy Ide, Martha Palmer, and James Pustejovsky, 40–47. Genova, Italy.
- Pala, Karel, Pavel Rychlý, and Pavel Smrž. 2003. "Text Corpus with Errors." In *Text, Speech and Dialogue*, edited by Václav Matoušek and Pavel Mautner, 2807:90–97. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-39398-6_13.
- Paquot, Magali. 2019. "The phraseological dimension in interlanguage complexity research." *Second Language Research* 35 (1): 121–145. <https://doi.org/10.1177/0267658317694221>.
- Paquot, Magali, and Sylviane Granger. 2012. "Formulaic language in learner corpora." *Annual Review of Applied Linguistics* 32:130–149. <https://doi.org/10.1017/S0267190512000098>.
- PDT – Prague Dependency Treebank 1.0*. 2000. Prague. <http://ufal.mff.cuni.cz/pdt/>.
- Pečený, Pavel. 2017. "Užití spojovacích prostředků v textech nerodilých mluvčích češtiny." PhD diss., Institute of Czech Language and Theory of Communication Faculty of Arts, Charles University. <https://is.cuni.cz/webapps/zzp/detail/105300/>.
- Petr, Jan, ed. 1987. *Mluvnice češtiny*. Praha: Academia.

- Petrov, Slav, Dipanjan Das, and Ryan T. McDonald. 2011. "A Universal Part-of-Speech Tagset." *CoRR* abs/1104.2086. arXiv: [1104.2086](https://arxiv.org/abs/1104.2086).
<http://arxiv.org/abs/1104.2086>.
- Pravec, Norma A. 2002. "Survey of learner corpora." *ICAME Journal* 26:81–114.
<http://icame.uib.no/ij26/pravec.pdf>.
- Preradović, Nives Mikelić, Monika Berać, and Damir Boras. 2015. "Learner Corpus of Croatian as a Second and Foreign Language." In *Multidisciplinary Approaches to Multilingualism*, edited by Kristina Cergol Kovačević and Sanda Lucia Udier, 107–126. Frankfurt am Main: Peter Lang.
https://www.researchgate.net/publication/304622939_Learner_Corpus_of_Croatian_as_a_Second_and_Foreign_Language#fullTextFileContent.
- Rakhilina, Ekaterina, Anastasia Vyrenkova, Elmira Mustakimova, Alina Ladygina, and Ivan Smirnov. 2016. "Building a learner corpus for Russian." In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, 66–75. Umeå.
<http://www.aclweb.org/anthology/W16-6509>.
- Ramasamy, Loganathan, Alexandr Rosen, and Pavel Straňák. 2015. "Improvements to Korektor: A case study with native and non-native Czech." In *ITAT 2015: Information technologies – Applications and Theory / SloNLP 2015*, edited by Jakub Yaghob, 73–80. Prague: Charles University in Prague.
<http://ceur-ws.org/Vol-1422/73.pdf>.
- Reznicek, Marc, Anke Lüdeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann, and Torsten Andreas. 2012. *Das Falko-Handbuch, Korpusaufbau und Annotationen, Version 2.01*. Technical report. Humboldt-Universität zu Berlin, Institut für deutsche Sprache und Linguistik – Korpuslinguistik. <https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko>.
- Richter, Michal. 2010. "An Advanced Spell Checker of Czech." Master's thesis, Faculty of Mathematics and Physics, Charles University.
<https://is.cuni.cz/webapps/zzp/detail/45334>.
- . 2013. *Korektor*. Software. Charles University in Prague, UFAL.
<http://hdl.handle.net/11858/00-097C-0000-000D-F67C-5>.

- Richter, Michal, Pavel Straňák, and Alexandr Rosen. 2012. "Korektor – A System for Contextual Spell-Checking and Diacritics Completion." In *Proceedings of COLING 2012: Posters*, 1019–1028. Mumbai, India: The COLING 2012 Organizing Committee, December.
<http://www.aclweb.org/anthology/C12-2099>.
- Ringbom, Håkan. 1998. "Vocabulary frequencies in advanced learner English: A cross-linguistic approach." In *Learner English on Computer*, edited by Sylviane Granger, 41–52. Harlow: Longman.
- Rio, Iria del, Sandra Antunes, Amália Mendes, and Maarten Janssen. 2016. "Towards error annotation in a learner corpus of Portuguese." In *5th NLP4CALL and 1st NLP4LA workshop in Sixth Swedish Language Technology Conference (SLTC)*. Umeå, Sweden: Umeå University.
<http://www.aclweb.org/anthology/W16-6502>.
- Rio, Iria del, and Amália Mendes. 2019. "Error annotation in the COPLE2 corpus." *Revista da Associação Portuguesa de Linguística*, no. 4, 225–239.
<https://ojs.apl.pt/index.php/RAPL/article/view/42/44>.
- Rosen, Alexandr. 2001. "A constraint-based approach to dependency syntax applied to some issues of Czech word order." PhD diss., Charles University.
<http://utkl.ff.cuni.cz/~rosen/public/THESIS/>.
- . 2017. "Introducing a corpus of non-native Czech with automatic annotation." In *Language, Corpora and Cognition*, edited by Piotr Pezik, Jacek Walinski, and Krzysztof Kosecki, 163–180. Frankfurt am Main, Bern, Bruxelles, New York, Oxford, Warszawa, Wien: Peter Lang.
http://utkl.ff.cuni.cz/~rosen/public/2016_SGT_lodz.pdf.
- Rosen, Alexandr, Jirka Hana, Barbora Štindlová, and Anna Feldman. 2014. "Evaluating and automating the annotation of a learner corpus." *Language Resources and Evaluation – Special Issue: Resources for language learning* 48, no. 1 (March): 65–92. https://www.researchgate.net/publication/234118426_Evaluating_and_automating_the_annotation_of_a_learner_corpus.
- Rozovskaya, Alla, and Dan Roth. 2010. "Annotating ESL Errors: Challenges and Rewards." In *Proceedings of NAACL'10 Workshop on Innovative Use of NLP for Building Educational Applications*. University of Illinois at Urbana–Champaign. <https://www.aclweb.org/anthology/W10-1004/>.

- Rychlý, Pavel. 2007. "Manatee/Bonito – A Modular Corpus Manager." In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, 65–70. Brno. https://www.sketchengine.eu/wp-content/uploads/Manatee-Bonito_2007.pdf.
- Schmidt, Thomas. 2009. "Creating and working with spoken language corpora in EXMARaLDA." In *LULCL II 2008: proceedings of the second colloquium on Lesser used languages and computer linguistics: Bozen-Bolzano, Italy, 13th–14th November, 2008*, edited by Verena Lyding, 54:151–164. EURAC research. Bozen–Bolzano: Europäische Akademie. http://www.eurac.edu/Org/LanguageLaw/Multilingualism/Projects/LULCL_II_proceedings.htm.
- Schmidt, Thomas, Kai Wörner, Hanna Hedeland, and Timm Lehmberg. 2011. "New and future developments in EXMARaLDA." In *Multilingual Resources and Multilingual Applications. Proceedings of GSCL Conference 2011 Hamburg*. <https://www.yumpu.com/en/document/read/8609912/new-and-future-developments-in-exmaralda>.
- Šebesta, Karel. 2010. "Korpusy češtiny a osvojování jazyka." *Studie z aplikované lingvistiky* 1:11–34.
- Šebesta, Karel, Zuzanna Bedřichová, Kateřina Šormová, Barbora Štindlová, Milan Hrdlička, Tereza Hrdličková, Jirka Hana, et al. 2017. *CzeSL Grammatical Error Correction Dataset (CzeSL-GEC)*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. <http://hdl.handle.net/11234/1-2143>.
- . 2019. *AKCES-GEC Grammatical Error Correction Dataset for Czech*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. <http://hdl.handle.net/11234/1-3057>.
- Šebesta, Karel, Zuzanna Bedřichová, Barbora Štindlová, Milan Hrdlička, Tereza Hrdličková, Jirka Hana, Alexandr Rosen, et al. 2012. *AKCES 4*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. <http://hdl.handle.net/11858/00-097C-0000-000C-2293-0>.

- Šebesta, Karel, Hana Goláňová, Tomáš Jelínek, Blanka Jelínková, Michal Křen, Jana Letafková, Pavel Procházka, and Hana Skoumalová. 2013. *SKRIPT2012: akviziční korpus psané češtiny – přepisy písemných prací žáků základních a středních škol v ČR*. Ústav Českého národního korpusu FF UK, Praha. <http://www.korpus.cz>.
- Šebesta, Karel, Hana Goláňová, Jana Letafková, and Jelínková Blanka. 2016. *AKCES 1*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL). <http://hdl.handle.net/11234/1-1741>.
- Selinker, Larry. 1972. "Interlanguage." *International Review of Applied Linguistics in Language Teaching (IRAL)* 10:209–231. <https://doi.org/10.1515/iral.1972.10.1-4.209>.
- . 1983. "Interlanguage." In *Second Language Learning: Contrastive analysis, error analysis, and related aspects*, 173–196. Ann Arbor, MI: The University of Michigan Press.
- Sgall, Petr, and Jiří Hronek. 1992. *Čeština bez příkras*. Praha: H&H.
- Sgall, Petr, Jiří Hronek, Alexandr Stich, and Ján Horecký. 1992. *Variation in Language: Code switching in Czech as a challenge for sociolinguistics*. Amsterdam/Philadelphia: John Benjamins.
- Short, David. 1993. "Czech." In *The Slavonic Languages*, edited by Bernard Comrie and Greville G. Corbett, 455–532. Routledge Language Family Descriptions. Routledge.
- Šindelářová, Jaromíra, and Svatava Škodová. 2013. "Práce s korpusem ve výuce žáků-cizinců." <https://clanky.rvp.cz/clanek/c/z/17481/PRACE-S-KORPUSY-VE-VYUCE-ZAKU-CIZINCU.html/>.
- Škodová, Svatava. 2017. "Realizace vybraných komunikačních funkcí v porovnání nerodilých a rodilých mluvčích češtiny." *Studie z aplikované lingvistiky* 8:121–135. https://sites.ff.cuni.cz/studiezaplikovanelingvistiky/wp-content/uploads/sites/19/2017/11/Svatava_Skodova_121-135.pdf.
- . 2018. "Sloveso JÍT v zrcadle užití nerodilými mluvčími češtiny." In *Čeština jako cizí jazyk v průsečíku pohledů*, edited by Svatava Škodová and Milan Hrdlička. Praha: Filozofická fakulta UK v Praze. https://www.researchgate.net/publication/332698345_Sloveso_JIT_v_zrcadle_uziti_nerodilymi_mluvcimi_cestiny.

- Škodová, Svatava. 2020. "Sloveso JÍT jako reprezentant pohybové události v prostoru." *Studie z aplikované lingvistiky* 11 (2).
- . n.d. "Genitivní a lokální vazby sloves v češtině nerodilých mluvčích." In prep.
- Škodová, Svatava, Kateřina Rysová, and Magdaléna Rysová. 2019. "Comparison of Automatic and Human Evaluation of L2 Texts in Czech." *Journal of Slavic Languages* (Seoul, Korea), 93–102.
<https://doi.org/10.30530/JSL.2019.04.24.1.93>.
- Škodová, Svatava, Barbora Štindlová, Alexandr Rosen, Tomáš Jelínek, and Barbora Hladká. 2019. *Příručka k morfologické anotaci češtiny nerodilých mluvčích*. Univerzita Karlova, Praha.
<https://doi.org/10.13140/RG.2.2.34952.78080>.
- Šotolová, Eva. 2008. *Vzdělávání Romů*. Praha: Karolinum.
- Spoustová, Drahomíra, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. "The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech." In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007*, 67–74. Praha, Czechia: Association for Computational Linguistics. <https://www.aclweb.org/anthology/W07-1709/>.
- Stemle, Egon W., Adriane Boyd, Maarten Janssen, Therese Lindström Tiedemann, Nives Mikelić Preradović, Alexandr Rosen, Dan Rosén, and Elena Volodina. 2019. "Working together towards an ideal infrastructure for language learner corpora." In *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*, edited by Andrea Abel, Aivars Glaznieks, Verena Lyding, and Lionel Nicolas, 427–468. Corpora and Language in Use – Proceedings 5. Louvain-la-Neuve: Presses universitaires de Louvain.
- Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. "brat: a Web-based Tool for NLP-Assisted Text Annotation." In *Proceedings of the Demonstrations Session at EACL 2012*. Avignon, France: Association for Computational Linguistics. <https://www.aclweb.org/anthology/E12-2021/>.
- Štícha, František. 2013. *Akademická gramatika spisovné češtiny*. Praha: Academia.

- Štindlová, Barbora. 2011. "Evaluace chybové anotace v žákovském korpusu češtiny." PhD diss., Charles University, Faculty of Arts.
<https://is.cuni.cz/webapps/zzp/detail/25046/>.
- . 2013. *Žákovský korpus češtiny a evaluace jeho chybové anotace*. Praha: Univerzita Karlova v Praze, Filozofická fakulta.
- Štindlová, Barbora, and Alexandr Rosen. 2012. "Návod k anotaci chybového korpusu." <https://doi.org/10.13140/RG.2.2.24106.64968>.
- Štindlová, Barbora, Alexandr Rosen, Jirka Hana, and Svatava Škodová. 2012. "CzeSL – an error tagged corpus of Czech as a second language." In *Corpus Data across Languages and Disciplines*, edited by Piotr Pezik, 28:21–32. Łódź Studies in Language. Frankfurt am Main: Peter Lang.
<http://utkl.ff.cuni.cz/~rosen/public/2011-czesl-palc.pdf>.
- Straka, Milan, and Michal Richter. 2015. *Korektor 2*. Software. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. <http://hdl.handle.net/11234/1-1469>.
- Straka, Milan, and Jana Straková. 2017. "Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe." In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99. Vancouver, Canada: Association for Computational Linguistics.
<http://www.aclweb.org/anthology/K/K17/K17-3009.pdf>.
- Straková, Jana, Milan Straka, and Jan Hajič. 2014. "Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition." In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 13–18. Baltimore, Maryland: Association for Computational Linguistics.
<http://www.aclweb.org/anthology/P/P14/P14-5003.pdf>.
- Stritar, Mojca. 2009. "Slovene as a Foreign Language: The Pilot Learner Corpus Perspective." *Slovenski jezik – Slovene Linguistic Studies* 7:135–152.
<https://doi.org/10.17161/SLS.1808.5274>.
- Tarone, Elaine. 2006. "Interlanguage." In *Encyclopedia of Language and Linguistics*, edited by Keith Brown, 747–751. Boston: Elsevier.

- Tenfjord, Kari, Jon Erik Hagen, and Hilde Johansen. 2009. "Norsk andrespråkscorpus (ASK) – design og metodiske forutsetninger." *NOA norsk som andrespråk* 25 (1): 52–81. <http://hdl.handle.net/1956/4442>.
- Tenfjord, Kari, Paul Meurer, and Knut Hofland. 2006. "The ASK Corpus – a Language Learner Corpus of Norwegian as a Second Language." In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA).
http://www.lrec-conf.org/proceedings/lrec2006/pdf/573_pdf.pdf.
- Tetreault, Joel, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. "Native Tongues, Lost and Found: Resources and Empirical Evaluations in Native Language Identification." In *Proceedings of COLING 2012*, 2585–2602. Mumbai, India: The COLING 2012 Organizing Committee.
<https://www.aclweb.org/anthology/C12-1158>.
- Tetreault, Joel, and Martin Chodorow. 2008. "Native Judgements of Non-Native Usage: Experiments in Preposition Error Detection." In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, 24–32. Manchester.
<https://www.aclweb.org/anthology/W08-1205.pdf>.
- Thomas, James. 2006. "Using corpora in Language Teaching and Learning." *Teaching English with Technology: A Journal for Teachers of English* 6 (1).
http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-3741e069-9f42-4941-aeaf-309f8ca4c2fd/c/4._Using_Corpora_in_Language_Teaching_and_Learning__by_James_Thomas__2006-1_.pdf.
- Tono, Yukio. 2003. "Learner corpora: design, development and applications." In *Proceedings of the 2003 Corpus Linguistics Conference*, edited by Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery, 800–809. Lancaster, UK: Lancaster University.
<http://www.scribd.com/doc/8254550/Learner%E2%80%93Corpora>.
- Tydlitátová, Ludmila. 2016. "Native Language Identification of L2 Speakers of Czech." Master's thesis, Faculty of Mathematics and Physics, Charles University.
<http://ufal.mff.cuni.cz/~hana/bib/tydlitativa-2016-BScThesis.pdf>.

- Vališová, Pavlína. 2009. "Korpus jako zdroj dat systémového popisu české konjugace při výuce češtiny jako cizího jazyka." Master's thesis, Masarykova univerzita. <https://is.muni.cz/th/u2b1o/>.
- . 2016. "Korpus ve výuce češtiny jako cizího jazyka – typy cvičení." In *Čeština jako cizí jazyk: VIII. Sborník příspěvků z VIII. mezinárodního symposia o češtině jako cizím jazyku*, 129–141.
- Veselovská, Ludmila. 1995. "Phrasal Movement and X-Morphology: Word Order Parallels in Czech and English Nominal and Verbal Projections." PhD diss., Palacký University. https://www.academia.edu/3269918/Phrasal_movement_and_X_morphology_Word_order_parallels_in_Czech_and_English_nominal_and_verbal_projections.
- Vetchinnikova, Svetlana. 2019. *Phraseology and the Advanced L2 Learner*. 38–65. Cambridge University Press. <https://doi.org/https://doi.org/10.1017/9781108758703>.
- Vokáčová, Martina. 2016. "Vliv gramatických profilů českých substantiv na jejich osvojování nerodilými mluvčími." Master's thesis, Filozofická fakulta Univerzity Karlovy. <https://is.cuni.cz/webapps/zzp/detail/178323/>.
- Volodina, Elena, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, et al. 2019. "The SweLL Language Learner Corpus: From Design to Annotation." *Northern European Journal of Language Technology* 6:67–104. <https://doi.org/10.3384/nejlt.2000-1533.19667>.
- Volodina, Elena, Arild Matsson, Dan Rosén, and Mats Wirén. 2019. "SVALA: an Annotation Tool for Learner Corpora generating parallel texts." In *Learner Corpus Research conference (LCR-2019), Warsaw, 12–14 September 2019, Book of Abstracts*. <https://www.ep.liu.se/ecp/159/023/ecp18159023.pdf>.
- Volodina, Elena, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016. "SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 206–212. Portorož, Slovenia: European Language Resources Association (ELRA). <https://www.aclweb.org/anthology/L16-1031>.

- Votrubec, Jan. 2006. "Morphological Tagging Based on Averaged Perceptron." In *WDS'06 Proceedings of Contributed Papers*, 191–195. Praha, Czechia: Matfyzpress, Charles University. ISBN: 80-86732-84-3. https://www.mff.cuni.cz/veda/konference/wds/proc/pdf06/WDS06_134_i3_Votrubec.pdf.
- Waibel, Birgit. 2008. *Phrasal verbs. German and Italian learners of English compared*. Saarbrücken: VDM.
- White, Lydia. 2003. "On the nature of interlanguage representation: Universal Grammar in the second language." In *Handbook of second language acquisition*, edited by Catherine J. Doughty and Michael H. Long, 9–42. Blackwell.
- Wirén, Mats, Arild Matsson, Dan Rosén, and Elena Volodina. 2019. "SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora." In *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8–10 October 2018*. Linköpings Universitet Electronic Press. <http://www.ep.liu.se/ecp/159/023/ecp18159023.pdf>.
- Wisniewski, Katrin, Claudia Woldt, Karin Schöne, Andrea Abel, Verena Blaschitz, Barbara Štindlová, and Kateřina Vodičková. 2014. *The MERLIN annotation scheme for the annotation of German, Italian, and Czech learner language*. Technical report. <http://merlin-platform.eu/docs/MERLIN-annotation-scheme.pdf>.
- Xiao, Richard. 2008. "Well-known and influential corpora." In *Corpus Linguistics. An International Handbook*, edited by Anke Lüdeling and Merja Kytö, 1:383–457. Handbooks of Linguistics and Communication Science [HSK] 29.1. Berlin and New York: Mouton de Gruyter.
- Zasina, Adrian Jan. 2019. "Korpusový přístup ve výuce češtiny jako cizího jazyka." PhD diss., Univerzita Karlova, Filozofická fakulta, Ústav českého národního korpusu. <https://dspace.cuni.cz/handle/20.500.11956/115540>.
- Zasina, Adrian Jan, and Svatava Škodová. 2020. "Konvence a variabilita v užívání prefixů u sloves pohybu nerodilými mluvčími češtiny." In *Konvence a kreativita v českém jazyce a literatuře, 2–4 September 2019, Cieszyn*, edited by Mieczysław Balowski. In print. Poznań: Uniwersytet Adama Mickiewicza.
- Zeldes, Amir, Florian Zipser, and Arne Neumann. 2013. *PAULA XML Documentation*. <http://hal.inria.fr/hal-00783716>.

- Zinsmeister, Heike, Ulrich Heid, and Kathrin Beck. 2014. "Adapting a part-of-speech tagset to non-standard text: The case of STTS." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 4097–4104. Reykjavík, Iceland. http://www.lrec-conf.org/proceedings/lrec2014/pdf/721_Paper.pdf.
- Znotina, Inga. 2017. "Computer-aided error analysis for researching Baltic interlanguage." In *Rural environment. Education. Personality. (REEP) Proceedings of the 10th International Scientific Conference*, edited by Vija Dislere and Zane Beitere-Šelegovska, 238–244. Jelgava: Latvia University of Agriculture. https://llufb.llu.lv/conference/REEP/2017/Latvia-Univ-Agricult-REEP-2017_proceedings-238-244.pdf.

Index of Authors

- Abuhakema, Ghazi, 27
Aharodnik, Katsiaryna, 214, 215
Almeida, Miguel, 121
Altenberg, Bengt, 22
Amaral, Luiz, 214
Augustinova, Tania, 240, 244
- Barkhuizen, Gary, 16, 17
Beck, Kathrin, 15
Bedřichová, Zuzanna, 245
Berać, Monika, 34, 190
Bermel, Neil, 233
Bley-Vroman, Robert, 60
Boras, Damir, 34, 190
Borin, Lars, 30
Bořkovicová, Máša, 245
Boušková, Petra, 14
Boyd, Adriane, 15, 36, 111
Bykh, Serhiy, 214
- Čermák, František, 202, 204
Chiarcos, Christian, 25
Chodorow, Martin, 88
Christ, Oliver, 179, 190
Cohen, Jacob, 88
Corder, S. P., 60
Cvrček, Václav, 204
- Dagneaux, Estelle, 25
- Dahlmeier, Daniel, 169
Das, Dipanjan, 125
Díaz-Negrillo, Ana, 59, 79, 125, 221
Dickinson, Markus, 15, 122
Dušková, Libuše, 15
- Ellis, Nick C., 22
Ellis, Rod, 16, 17, 26, 61
Evert, Stefan, 30
- Feldman, Anna, 27
Fernández-Domínguez, Jesús, 59, 79
Filip, Hana, 237
Fitzpatrick, Eileen, 27, 105
Flor, Michael, 139
Forsberg, Markus, 30
Fronek, Josef, 233
Futagi, Yoko, 139
- Gilquin, Gaëtanelle, 22, 27, 201
Granger, Sylviane, 22, 25–27, 36, 44,
54, 79, 201
Granstedt, Lena, 38, 51, 61
Günther, Britta, 26
Günther, Herbert, 26
- Hagen, Jon Erik, 32
Hajič, Jan, 13, 120, 121, 194, 227
Hana, Jirka, 169, 244

- Hardie, Andrew, 30
 Harkins, William E., 233
 Heid, Ulrich, 15
 Hercíková, Barbora, 15
 Hirschmann, Hagen, 15, 59, 60, 62,
 63, 79, 80
 Hladká, Barbora, 169, 214
 Hnátková, Milena, 121
 Hofland, Knut, 32
 Holub, Martin, 214
 Hronek, Jiří, 233
 Hudousková, Andrea, 146, 208, 209,
 219

 Isahara, Hitoshi, 79
 Israel, Ross, 15
 Izumi, Emi, 79

 James, Carl, 15, 80
 Janda, Laura A., 233
 Janssen, Maarten, 190
 Jarvis, Scott, 214
 Jelínek, Tomáš, 194
 Johansen, Hilde, 32
 Johns, Tim, 203

 Kaczmarska, Elżbieta, 217
 Karlík, Petr, 233, 242
 Kilgarriff, Adam, 179, 190
 Kopřivová, Marie, 234
 Křen, Michal, 204
 Kříž, Vincent, 214

 Lee, Sun-Hee, 15
 Leech, Geoffrey, 25
 Lennon, Paul, 61, 62
 Leńko-Szymańska, Agnieszka, 22, 44
 Lüdeling, Anke, 59, 60, 62, 63, 72,
 79, 80, 95, 107

 Lukšija, Melita, 207, 208

 Machálek, Tomáš, 179
 Marek, Michal, 212
 Martins, André, 121
 Matsson, Arild, 38
 McDonald, Ryan T., 125
 McEnery, Tony, 25
 Mendes, Amália, 34, 105, 190
 Meunier, Fanny, 22, 25, 27, 201
 Meurer, Paul, 30, 32, 33
 Meurers, Detmar, 41, 69, 88, 95, 214
 Mírovský, Jiří, 211

 Náplava, Jakub, 139, 140, 169, 213,
 214, 227
 Naughton, James, 233
 Nekula, Marek, 233, 242
 Nesselhauf, Nadja, 22, 32
 Neumann, Arne, 31
 Ng, Hwee Tou, 169
 Nicholls, Diane, 33, 79
 Novák, Michal, 211, 219

 Oliva, Karel, 244
 Osolsobě, Klára, 207

 Pajas, Petr, 136, 175
 Pala, Karel, 212
 Paquot, Magali, 22
 Pečený, Pavel, 209
 Pecina, Pavel, 212
 Petkevič, Vladimír, 121
 Petr, Jan, 233
 Petrov, Slav, 125
 Pravec, Norma A., 32
 Preradović, Nives Mikelić, 34, 190

 Ragheb, Marwa, 122
 Rakhilina, Ekaterina, 37

- Ramasamy, Loganathan, 139, 140, 213
- Reznicek, Marc, 35
- Richter, Michal, 139, 192, 212, 213, 227
- Ringbom, Håkan, 22
- Rio, Iria del, 34, 105, 190
- Rosen, Alexandr, 78, 90, 139, 140, 158, 192, 212, 213, 227, 244
- Roth, Dan, 88
- Roxendal, Johan, 30
- Rozovskaya, Alla, 88
- Rusínová, Zdenka, 233, 242
- Rychlý, Pavel, 179, 190, 212
- Rysová, Kateřina, 211, 212, 219
- Rysová, Magdaléna, 211, 212, 219
- Schmidt, Thomas, 31, 72, 136
- Šebesta, Karel, 45, 46, 169, 172
- Seegmiller, Steve, 27, 105
- Selinker, Larry, 15, 26, 60
- Sgall, Petr, 233
- Short, David, 233
- Šindelářová, Jaromíra, 204
- Škodová, Svatava, 113, 147, 204, 206, 208–210, 212
- Skoumalová, Hana, 121
- Smith, Noah A., 121
- Smrž, Pavel, 212
- Šotolová, Eva, 245
- Spousta, Miroslav, 212
- Spoustová, Drahomíra, 213
- Stemle, Egon W., 25, 37, 59, 189
- Stenertorp, Pontus, 175
- Štěpánek, Jan, 136
- Štícha, František, 204
- Štindlová, Barbora, 32, 53, 78, 89, 140
- Straka, Milan, 121, 130, 139, 140, 169, 194, 212, 214, 227
- Straková, Jana, 121, 130, 194, 227
- Straňák, Pavel, 139, 140, 192, 212, 213, 227
- Stritar, Mojca, 32, 41
- Tapper, Marie, 22
- Tarone, Elaine, 15
- Tenfjord, Kari, 32
- Tetreault, Joel, 88, 214
- Thomas, James, 204
- Tono, Yukio, 54, 201
- Townsend, Charles E., 233
- Tydlitátová, Ludmila, 214, 215
- Uchimoto, Kiyotaka, 79
- Vališová, Pavlína, 204, 207, 208
- Veselovská, Ludmila, 238
- Vetchinnikova, Svetlana, 22
- Vokáčová, Martina, 209
- Volodina, Elena, 38, 51, 61
- Votrubec, Jan, 227
- Waclawičová, Martina, 234
- Waibel, Birgit, 22, 44
- White, Lydia, 16
- Wirén, Mats, 31
- Wisniewski, Katrin, 36, 111
- Xiao, Richard, 32
- Zasina, Adrian Jan, 204, 208, 209, 217
- Zeldes, Amir, 31
- Zinsmeister, Heike, 15
- Zipser, Florian, 31
- Znotina, Inga, 191

Index of Corpora

- A-GEC, *see* AKCES-GEC
AKCES, **41**, **45**, **47**, **169**, **210**, **245**
AKCES 1, **46**, **172**
AKCES 2, **46**
AKCES 3, **157**
AKCES 4, **46**, **155**, **157**, **170**, **172**
AKCES Grammatical Error
Correction Dataset, *see*
AKCES-GEC
AKCES-GEC, **155**, **169**, **214**
ASK, **32**, **39**, **41**
- C-GEC, *see* CzeSL-GEC
Cambridge Learner Corpus, **33**, **39**,
176, **183**, **227**
Cambridge Reference Corpus, **33**
Chyby, **212**
CLC, *see* Cambridge Learner Corpus
CNC, **43**, **46**, **201**, **203**, **207**, **208**, **218**
COPLE2, **34**, **35**, **39**, **190**
Corpus de Português Língua
Estrangeira / Língua
Segunda, *see* COPLE2
Corpus of Arabic learners of Czech,
193
Croatian Learner Text Corpus, *see*
CroLTeC
CroLTeC, **34**, **39**, **190**, **217**
- Czech National Corpus, *see* CNC
CzeFL-LONG, **46**
CzeSL Grammatical Error
Correction Dataset, *see*
CzeSL-GEC
CzeSL in TEITOK, **31**, **32**, **34**, **54**,
57, **87**, **121**, **151**, **155**, **156**,
170, **217**, **223**
CzeSL-GEC, **155**, **169**, **213**, **214**
CzeSL-LONG, **46**
CzeSL-man, **54**, **108**, **130**, **135**, **136**,
141, **145**, **155**, **156**, **163**,
167–170, **191**, **212**, **213**
CzeSL-man v.0, **156**
CzeSL-man v.0, a2, **156**
CzeSL-man v0, **31**, **163**, **165**, **169**,
172, **213**, **223**, **224**
CzeSL-man v1, **30**, **121**, **163**, **164**,
166, **168**, **169**, **223**, **224**
CzeSL-man v1 downloadable, **56**,
164, **165**
CzeSL-man v1 searchable, **57**, **121**,
165, **180**
CzeSL-man v2, **31**, **57**, **121**, **167**, **176**,
223, **224**
CzeSL-MD, **31**, **140**, **155**, **168**, **170**,
224
CzeSL-plain, **42**, **155**, **157**, **158**, **163**,

- 172, 180, 223, 228
- CzeSL-SGT, 30, 55, 57, 68, 121, 139, 151, 155, 157, **158**, 159–166, 168, 170, 180, 181, 213, 215, 219, 223, 231
- CzeSL-TH, 107, 134, 151, 155, **168**, 170, 229
- CzeSL-UD, 124, 130, 152, 155, **169**, 170, 175, 224
- EARLYFAMILY 2018, 47
- Ein fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache, *see* Falko
- Falko, 31, **35**, 37, 39, 62, 67, 71
- Frog Story Corpus, 47
- Frog, where are you?, 47
- Google Web1T, 139
- ICLE, **36**, 39
- International Corpus of Learner English, *see* ICLE
- MERLIN, 31, **36**, 39, 50, 71, 111, 203, 209, 211, 218, 221
- Multilingual Platform for European Reference Levels, *see* MERLIN
- Norsk andrespråkskorpus, *see* ASK
- Open Cambridge Learner Corpus, **34**
- ORAL 2006, 234
- PDT, 121, 175
- PiKUST, 32, 41
- Prague Dependency Treebank, *see* PDT
- RLC, **37**, 39, 217
- ROMi 1.0, 47
- Russian Learner Corpus, *see* RLC
- SCHOLA 2010, 45, 46
- SKRIPT 2012, 46, 151, 155, 170, 172
- SKRIPT 2015, 21, 31, 46, 151, 155, 170, 172, 190
- SweLL, **38**, 39
- SYN2015, 166
- TLE, 169
- WebColl, 212, 213

Index of Tools

- ANNIS, 36, 37, 39, 224
- Bonito, 179
- brat, 31, 140, 147–149, 168, 173, **175**,
176–178
- CNC KonText, *see* KonText
- ConSpel, 139
- Corpus Workbench, *see* CWB, 179
- Corpuscle, 30, 33
- CQP, 190
- CWB, 30, 38, 176, 190
- EvalD, 211
- EXMARaLDA, 31, 36, 136, 173
- feat, 31, 38, 56, 72, 76, 134–136, 151,
164, **173**, 174, 178, 194,
196, 197, 228
- IMS Open Corpus Workbench, *see*
CWB
- KonText, 30, 32, 43, 46, 56, 157, 158,
161, 165, 166, 172, 173, 176,
178, **179**, 180–184, 186, 189,
190, 208, 223, 225, 227
- Korektor, 139, 140, 152, 158, 159,
181, 192, 212, 213, 215, 227
- Korp, 30, 32, 38
- LINDAT, 46, 47
- Manatee, 179, 190
- Microsoft Excel, 36
- MorphoDiTa, 121, 152, 194
- NoSketch Engine, 179
- Oxygen, 32
- PAULA, 31
- PML-TQ, 31
- SeLaQ, 31, 163, 173, **177**, 178, 227
- Sketch Engine, 30, 33, 34, 39, 173,
176, 178, **179**, 180, 183,
186, 187, 189, 190, 227
- Speed, 173, **174**
- SVALA, 31, 32, 38, 39
- SyD, 208
- TEITOK, 31, 32, 34, 35, 46, 108,
133, 136, 139, 149–152, 172,
173, 177, 178, **189**, 213,
223–225, 227, 229
- TrEd, 130, 152, 153, 173, **175**
- TrEd-ud, 130, 153, 175
- TurboParser, 121
- UDPipe, 130, 152, 153, 194, 197
- Word Sketch, 179
- XMLmind, 132

Index

- alternatives
 - categorization, 80, 110, 148
 - interpretation, 78
 - target hypothesis, *see* target hypothesis: alternative
 - transcription, 77
- annotation
 - automatic, 97, 120, 121, 136, 141, 143, 144, 158, 170, 175, 219
 - editor, 173, 175, 189
 - error, 27, 59, 143
 - evaluation, 79, 87, 130
 - lemma, 126
 - linguistic, 28, 119
 - manual, 43, 134, 147, 152, 163, 168–170, 173–175, 189
 - morphosyntax, 76, 86, 87, 125, 158
 - morphs, 140, 141
 - scheme, 30
 - syntax, 121, 122, 128, 175
 - textual, 27
- anonymization, 31, 53
- authenticity, 202
- CCz, *see* Colloquial Czech
- CEFR, *see* Common European Framework of Reference
- clitic, 20, 21, 68, 238, 244
- collection, 49
- Colloquial Czech, 68, 103, 111, 233
- Common Czech, *see* Colloquial Czech
- Common European Framework of Reference, 15, 37, 42
- computer-aided error analysis, 22
- contrastive analysis, 59
- corpus
 - cross-sectional, 27
 - quasi-longitudinal, 27
 - longitudinal, 27, 43, 46
 - native Czech, 25, 46, 170
- Corpus Query Language, 179, 180, 186, 198
- crosslinguistic influence, *see* interference
- curriculum, 16
- data-driven learning, 203, 207
- DDL, *see* data-driven learning
- developmental corpus, *see* corpus: longitudinal
- developmental patterns, 16
- diglossia, 68
- emendation, *see* target hypothesis

- epenthesis, 102, 109, 110
- error, 61
- agreement, 19, 21, 76, 77, 83, 86
 - aspect, 112
 - auxiliary, 83, 86, 87
 - case, 19, 86, 87, 109, 112, 159, 210
 - categorization, 29, 81, 83, 221
 - category, 29, 76, 93
 - clitic, 86, 87
 - diacritics, 68, 101
 - domain, 29, 108, 147
 - embedded, *see* error: successive
 - follow-up, 77, 86
 - formal, 80, 97, 136
 - grammar-based, 80, 81, 219
 - implicit categorization, 105, 150
 - inflection, 19, 20, 81, 86, 109, 111, 113, 124, 127, 144, 159, 209
 - interferential, 66
 - lexical, 83, 87, 91, 92, 114, 209
 - location, 108
 - metathesis, 103
 - pronoun, 21, 209
 - pronunciation, 97, 100, 101, 111
 - real-word, 159
 - reflexive, 20, 21, 86
 - span, 61, 81, 109, 112, 149, 175
 - spelling, 81, 86, 97, 99, 100, 111, 113, 159
 - stem, 86
 - successive, 62, 70
 - taxonomy, 29, 78, 113
 - valency, 19, 76, 83, 86, 87
 - word boundary, 21, 31, 70, 81, 105, 185, 197
 - word order, 20, 67, 83, 86, 87
- error analysis, 59
- error annotation, *see* error: annotation
- error tag, *see* error: category
- error tagset, *see* error: categorization
- false friends, 66
- foreign language, *see* second language
- format, 223
- conversion, 176
 - inline, 30, 70, 136, 191
 - multi-tier, 31, 35, 39, 69, 136, 173, 177
 - stand-off, 35, 39, 136, 173, 194
 - structure, 183
 - tabular, 30, 36, 71
 - vertical, 140, 179
- handwriting, 51, 132, 191
- homonymy, 21, 87, 124, 219
- HTML, 53, 132
- IAA, *see* inter-annotator agreement
- IL, *see* interlanguage
- information structure, 67, 244
- intelligibility, 125
- inter-annotator agreement, 87, 88, 130
- interference, 66
- interlanguage, 15, 16, 22, 26, 43, 44, 60, 107, 202
- interlingual errors, 17
- interpretation, 66, 126, 128
- intralingual errors, 17
- L2, *see* second language
- license, 31, 43
- manuscript, *see* handwriting
- metadata, 43, 54, 202, 220
- morphosyntax, 120
- multi-word unit, 76

- natural language processing, 17, 18, 211
- NLP, *see* natural language processing
- normalization, *see* target hypothesis
- orders of acquisition, 16, 18
- overgeneralization, 17
- palatalization, 102
- PML, *see* Prague Markup Language
- Prague Markup Language, 133, 136, 173
- principle of positive assumption, 61
- pro-drop, 70
- protection of personal rights, 50
- prothesis, 103
- pseudonymization, 31
- query interface, 32, 177, 179, 189, 223
- reconstruction, *see* target hypothesis
- Romani ethnolect, 25, 42, 43, 46, 245
- SCz, *see* Standard Czech
- search interface, *see* query interface
- second language, 14, 15, 17, 26
- second language acquisition, 14–16, 37, 59, 201
- second language teaching, 201
- secondary error, *see* error: follow-up
- SLA, *see* second language acquisition
- stages of acquisition, *see* orders of acquisition
- Standard Czech, 18, 64, 233
- style, 87
- syncretism, 18
- target hypothesis, 22, 28, 62, 66, 93, 120, 168
- alternative, 28, 77
- automatic, 139
- successive, 29, 35, 87
- TEI, *see* Text Encoding Initiative
- testing, 16
- Text Encoding Initiative, 33, 39, 57, 136, 189
- TH, *see* target hypothesis
- token, 30, 124, 180, 191, 223
- transcription, 28, 51, 132, 222
- transfer, *see* interference
- UD, *see* Universal Dependencies
- Universal Dependencies, 122, 169, 175, 194
- word order, 67, 244
- XML, 39, 53, 132, 136, 191