Noname manuscript No.
(will be inserted by the editor)

Evaluating and automating the annotation of a learner corpus

Alexandr Rosen · Jirka Hana · Barbora Štindlová · Anna Feldman

Received: date / Accepted: date

Abstract The paper describes a corpus of texts produced by non-native speakers of Czech. We discuss its annotation scheme, consisting of three interlinked tiers, designed to handle a wide range of error types present in the input. Each tier corrects different types of errors; links between the tiers allow capturing errors in word order and complex discontinuous expressions. Errors are not only corrected, but also classified. The annotation scheme is tested on a data set including approx. 175,000 words with fair inter-annotator agreement results. We also explore the possibility of applying automated linguistic annotation tools (taggers, spell checkers and grammar checkers) to the learner text to support or even substitute manual annotation.

Keywords learner corpus · error annotation · second language acquisition · Czech

The corpus is one of the tasks of the project Innovation of Education in the Field of Czech as a Second Language (project no. CZ.1.07/2.2.00/07.0259), a part of the operational programme Education for Competiveness, funded by the European Structural Funds (ESF) and the Czech government. The annotation tool was also partially funded by grant no. P406/10/P328 of the Grant Agency of the Czech Republic.

Alexandr Rosen
Charles University, Prague, Czech Rep.
E-mail: alexandr.rosen@ff.cuni.cz

Jirka Hana
Charles University, Prague, Czech Rep.
E-mail: jirka.hana@gmail.com

Barbora Štindlová
Technical University, Liberec, Czech Rep.
E-mail: barbora.stindlova@tul.cz

Anna Feldman
Montclair State University, Montclair, NJ, USA
E-mail: feldmana@mail.montclair.edu

1 Introduction

Learner corpora, i.e. electronic collections of texts produced by non-native speakers, are a rich source of information about specific features of learners' language. They can be annotated in the usual ways, common in other types of corpora, i.e. by metadata, morphosyntactic categories and syntactic structure, but their most interesting aspect is examples of deviant use, which can be identified, corrected and classified. Annotation of this kind is 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.

Following an overview of some learner corpora (§2) we present a learner corpus of Czech, consisting of approx. 2 million running words, compiled from texts written by students of Czech as a second or foreign language at all levels of proficiency (§3). We discuss the corpus annotation scheme, consisting of three interlinked tiers, designed to cope with a wide range of error types present in the input (§4). Tier 0 represents the transcribed input, Tier 1 corrects non-words and Tier 2 the remaining error types; links between the tiers allow capturing errors in word order and complex discontinuous expressions. Errors are not only corrected, but also classified according to a taxonomy. Annotation of this kind is supplemented by a formal classification, e.g. an error in morphology can also be specified as being manifested by a missing diacritic or a wrong consonant change.

The annotation scheme was tested in two rounds, each time on a doubly-annotated sample – first on a pilot annotation of approx. 10,000 words and later on a large data set including approx. 175,234 words, both with fair inter-annotator agreement results, calculated as several inter-annotator agreement measures (§5).

To assist the annotator and to supply additional information about deviations from the standard, we aim at a synergy of manual and automatic annotation, deriving information from the original input and from the manual annotation (§6). Some methods interact with the annotator (e.g. a spell checker within the annotation editor marks potentially incorrect forms), or use results of manual annotation, including an automatic check for consistency and compliance with the annotation guidelines. After approval by the annotator's supervisor, some error tags are specified in more detail and more tags are added automatically.

To assist the annotator even further, we also experiment with fully automatic methods (§7.1). We perform automatic emendation by a mildly context-sensitive spell checker and plan to use a grammar checker or a stochastic model to identify error types. A tagger trained on native speakers' language is used at the tier of corrected text to supply morphosyntactic categories, but we explore options of applying general-purpose tools to the original text (see, e.g. Van Rooy and Schäfer (2003); Dickinson (2010)).

2 Learner corpora

A learner corpus, also called interlanguage or L2 corpus, is a computerised textual database of language as produced by foreign/second language (L2) learners (Leech, 1998, p. xiv). It is a very useful resource in the research of second language acquisition (SLA) and foreign language teaching (FLT). It serves as a repository of authentic data about a specific variety of natural language (Granger, 2003b), namely the learner language, or interlanguage (IL).[1]

Learner corpora can be used to compare non-native and native speakers' language, or interlanguage varieties. They can be studied on the background of traditional native language 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," in comparison with the language of native speakers. Recent studies have focused primarily on the frequency of use of separate language elements (Ringbom, 1998), collocations and prefabs (Nesselhauf, 2005), lexical analysis and phrasal use (de Cock, 2003), etc.

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. Recent studies focus on lexical errors (Leńko-Szymańska, 2004), wrong use of verbal tenses (Granger, 1999) or phrasal verbs (Waibel, 2008).

Learner corpora can differ in many ways (for more details see, e.g. Granger, 2008, p. 260):

- Medium: Learner corpora can capture written or spoken texts, the latter much harder to compile, thus less common.
- 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/developmental data: Most L2 corpora are cross-sectional, gathering data from various types of learners. Only few L2 corpora are developmental (longitudinal), 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 in SLA research as quasi-longitudinal learner corpora.
- Annotation: Many learner corpora contain only raw data, some contain emendations (i.e. corrections), but only few use error tags to classify errors, e.g. Fitzpatrick and Seegmiller (2001); Granger (2003a); Abuhakema et al (2009). Some corpora use linguistic annotation, the most common is part of-speech (POS) tagging.

[1] Interlanguage 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. See Selinker (1972).

Table 2 presents an overview of some currently available learner corpora. For more details see, e.g. Pravec (2002), Nesselhauf (2005), Štindlová (2011), and Xiao (2008), for a more exhaustive list see https://www.uclouvain.be/en-cecl-lcworld.html.

Table 1 Some currently available learner corpora

Size (th. of words)	L1	TL	TL proficiency	Medium	Error annotation
ICLE *– International Corpus of Learner English*					
3,000	26	English	advanced	written	yes (part)
CLC *– Cambridge Learner Corpus*					
35,000	130	English	all levels	written	yes (part)
LINDSEI *– Louvain International Database of Spoken English*					
800	11	English	advanced	spoken	yes (part)
USE *– Uppsala Student English Corpus*					
1,200	Swedish	English	advanced	written	no
CYLIL *– Corpus of Young Learner Interlanguage*					
500	4	English	all levels	spoken	no
HKUST *– Hong Kong Univ. of Science and Technology Corpus of Learner English*					
25,000	Chinese	English	advanced	written	yes (part)
CHUNGDAHM *– Chungdahm English Learner Corpus*					
131,000	Korean	English	all levels	written	yes (part)
JEFLL *– Japanese EFL Learner Corpus*					
700	Japanese	English	beginners	written	yes (part)
MELD *– Montclair Electronic Language Learners' Database*					
1,000	16	English	advanced	written	no
NICT JLE *– NICT Japanese Learner English*					
2,000	Japanese	English	all levels	spoken	yes (part)
FALKO *– Fehlerannotiertes Lernerkorpus*					
300	5	German	advanced	written	yes
FRIDA *– French Interlanguage Database*					
200	various	French	intermediate	spoken	yes (part)
ARIDA *– Arabic Interlanguage Database*					
8,5	various	Arabic	intermediate/advanced	written	yes
FLLOC *– French Learner Language Oral Corpora*					
2,000	English	French	all levels	spoken	no
PiKUST *– Poskusni korpus usvajanja slovenščine kot tujega jezika*					
40	18	Slovene	advanced	written	yes
ASU *– ASU Corpus*					
500	various	Norwegian	advanced	written	no
CEDEL 2 *– Corpus Escrito del Español como L2*					
75	various	Spanish	all levels	written	yes (part)

3 A learner corpus of Czech

The learner corpus of Czech as a Second Language (*CzeSL*) is built as a part of a larger project, the Acquisition Corpora of Czech (*AKCES*), a research programme pursued at Charles University in Prague since 2005 (Šebesta, 2010; Hana et al, 2010; Štindlová et al, 2012c,b; Hana et al, 2012). All of the included corpora (see below) are collected and built under similar conditions, allowing for a wide range of linguistic comparisons. In addition to CzeSL, it includes the following corpora:

- *SCHOLA 2010* and *EDUCO* – recordings and transcripts of classes from Czech primary schools (about 800,000 words each, finished)
- *SKRIPT* – written texts of Czech students (about 600,000 words so far, in development)
- *IUVENT* – spoken corpus of native young Czechs' language (planned)

Table 2 summarizes the basic properties of the *CzeSL* corpus. It is focused on four main groups of learners of Czech:

- native speakers of other Slavic languages,
- native speakers of other Indo-European languages,
- native speakers of distant non-Indo-European languages, and
- young speakers of Czech with Romani background.[2]

The data collected include:

1. Written texts, produced during all range of situations throughout the language-learning process, collected as manuscripts and transcribed into an electronic format. The transcription follows rules designed to preserve many features of handwritten texts (such as self-corrections or emoticons, Štindlová, 2011, p. 106).
2. Spoken data.
3. Bachelors' and Masters' theses, written in Czech by non-native students.

Table 2 Size of various subcorpora (in thousand words, approximate)

		transcribed	annotated	doubly annotated
Foreigners	spoken	11	0	
	written	1,150	200	75
	theses	490	0	
Roma	spoken	540	0	
	written	450	170	110
Total		2,641	370	185

[2] For some members of the Czech Roma community it might be difficult to identify their first language, yet such students often exhibit a number of traits typical for the process of acquisition of Czech as a second language. 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.

The data cover all language levels according to the Common European Framework of Reference for Languages (CEFR), from real beginners (A1 level) to advanced learners (level B2 and higher), with a balanced mix of levels as much as possible.

Each text is equipped with metadata records, some of them relate to the respondent (including sociological data about the learner, such as age, gender, and language background – the first language, proficiency level in Czech, knowledge of other languages, duration and conditions of language acquisition), while other specify the character of the text and circumstances of its production (availability of reference tools, type of elicitation, temporal and size restrictions etc.).

The intended use of the Czech learner corpus is mainly pedagogical. It will be used in the education of teachers of Czech as a foreign language, it will serve as a source of examples for particular phenomena or of complete authentic texts that can be used both in the classroom and in the production of educational tools, and will help to tailor instructions and teaching materials to specific groups of learners (e.g. groups with different native languages or groups of different ages). Moreover, we expect *CzeSL* to become a resource for an extensive research of Czech as a second language and the second language acquisition in general (Štindlová, 2011).

The corpus is searchable online.[3] Queries can refer to transcripts (for privacy reasons, scans of handwritten text are not publicly accessible), error annotation, morphological tags and lemmas.

4 Error annotation of *CzeSL*

4.1 Annotation of learner corpora

In the context of second/foreign language acquisition, the learners' language is seen as an independent system, which should be analysed in its entirety, with incorrect structures as an important part. Texts produced by non-native speakers can be annotated in two different ways:

- Linguistic mark-up (e.g. part-of-speech tagging, morphological or syntactic annotation, lemmatisation etc.). In most learner corpora, at least some parts are POS-tagged by tools and tagsets originally developed for the analysis of the native language, cf., e.g. Van Rooy and Schäfer (2003). However, it is often far from obvious what kind of annotation an incorrect expression should receive.
- Error annotation, cf., e.g. Díaz-Negrillo and Fernández-Domínguez (2006). There are two different kinds of error annotation:
  - emendation: correction of erroneous text – establishing one or more *target hypotheses* about the author's intention and its expression
  - error categorisation: annotation of errors with tags from a predefined error taxonomy

  Investigating learners' language is easier when deviant forms are annotated at least by their correct counterparts, or, even better, by tags making the nature of the

[3] See http://utkl.ff.cuni.cz/learncorp.

error explicit.[4] Although learner corpora tagged this way exist, the two decades of research in this field have shown that designing a tagset for the annotation of errors is a task highly sensitive to the intended use of the corpus and the results are not easily transferable from one language to another.

4.2 Annotation scheme as a compromise

Building an error-annotated learner corpus of Czech is a challenging task. Czech, at least in comparison to languages of the existing annotated learner corpora, has a more complex morphology and a less rigid word order, which opens annotation issues that have not been addressed before.[5] Moreover, although the annotation scheme should be sufficiently informative and extensible, it should also be manageable and easily applicable, i.e. not too extensive. The resulting scheme and error typology is a compromise between the limitations of the annotation process and our research goals. Some of the issues involved, such as interference, interpretation, word order or style, do not have straightforward solutions:

Interference: Being no experts in L2 acquisition, the annotators cannot be expected to spot cases of linguistic interference of L1 or some other language known to the learner. 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' and assumed *hrad* 'castle' as the Czech equivalent of Russian *gorod* 'town, city'.

Interpretation: For some types of errors, the problem is to define the limits of interpretation. The clause *kdyby citila na tebe zlobna* is grammatically incorrect, yet roughly understandable as 'if she felt angry at you'. In such cases the task of the annotator is interpretation rather than correction. The clause can be rewritten as *kdyby se na tebe cítila rozzlobená* 'if she felt angry at you', or *kdyby se na tebe zlobila* 'if she were angry at you'; the former being less natural but closer to the original. It is difficult to provide clear guidelines.

Word order: Czech constituent order reflects information structure. It may be hard to decide (even in a context) whether an error is present. The sentence *Rádio je taky na skříni* 'A radio is also on the wardrobe' suggests that there are at least two radios

[4] However, some authors intentionally avoid categorizing errors. They see categorisation as an interpretation model, influencing access to the data. Instead, they use emendation as an implicit explanation for the errors (Fitzpatrick and Seegmiller, 2004).

[5] We are aware of four other Slavic L2 corpora. However they are either small (the first one), or under development (the other three).

- *PiKUST* (Stritar, 2009), a 35KW corpus of learner Slovene, error annotation adopted from the Norwegian *ASK* project
- *piRULEC* (Kisselev, 2012), a corpus of learner Russian, currently being built at Portland State University; a collection of academic writings of advanced foreign and heritage learners of Russian
- A 10KW corpus collected from advanced American learners of Russian (Pavlenko and Hasko, 2007)
- A corpus of theses written in several Slavic languages by non-native students of the University of Helsinki

A 7MW 'didactical/educational' part of the Russian National Corpus is sometimes referred to as a learner corpus, but in fact it includes works of fiction on a list of recommended readings in Russian schools (see http://www.ruscorpora.ru/en/corpora-structure.html).

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 skříni je taky rádio*.

Style: Students often use colloquial expressions, usually without being aware of their status and the appropriate context for their use. Even though these expressions might be grammatical, we emend them with their standard counterparts under the rationale that the intention of the student was to use a register that is perceived as unmarked.

Our error annotation is primarily concerned with the acceptability of the grammatical and lexical aspects of the learner's language in a narrow sense. However, we anticipate that future projects would annotate the corpus with less formal properties of speech, such as the degree of achievement of a communicative goal.

4.3 Multi-tier annotation

The optimal error annotation strategy is determined both by the goals and resources of the project and by the type of the language. A single-tier scheme could be used for a specific narrowly defined purpose, such as investigation of morphological properties of the learner language. However, in our case, to apply the single-tier scheme would be problematic. First of all, our corpus should be open to multiple research goals. Thus, a restricted set of linguistic phenomena or a single tier of analysis is not satisfactory. As a result, it is necessary to register successive emendations and to maintain links between the original and the emended forms even when the word order changes or in cases of dropped or added expressions. Another reason is the need to annotate errors spanning multiple forms, often in discontinuous positions.

In the ideal case, the annotator should be free to use an arbitrary number of tiers to suit the needs of successive emendations, choosing from a set of linguistically motivated tiers or introducing annotation tiers ad hoc. On the other hand, the annotator 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. Thus the scheme consists of three interconnected tiers – see Fig. 1, glossed in (1):

- Tier 0 – anonymised transcript of the hand-written original with some properties of the manuscript preserved (variants, illegible strings)
- Tier 1 – forms incorrect in isolation are fixed. The result is a string consisting of correct Czech forms, even though the sentence may not be correct as a whole
- Tier 2 – handles all other types of errors (valency, agreement, word order, etc.)

The correspondences between successively emended forms are explicitly expressed. Nodes at neighbouring tiers are usually linked 1:1, but words can be joined (*kdy by* as in Fig. 1) or split, deleted or added. These relations can interlink any number of potentially non-contiguous words across the neighbouring tiers. Multiple words

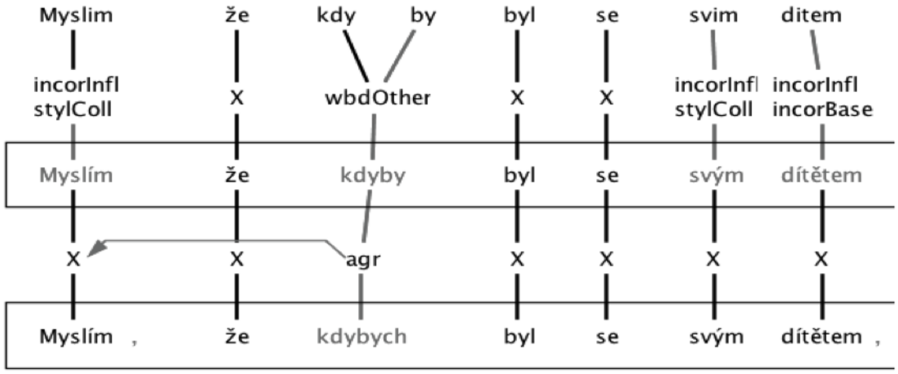

Fig. 1 Example of the three-tier error annotation scheme

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.

Whenever a word form is emended, the type of error can be specified as a label at the link connecting the incorrect form at a lower tier with its emended form at a higher tier (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).

Each node may be assigned information in addition to the form of the word, such as lemma, morphosyntactic category or syntactic function.

(1) Myslím, že kdybych byl se svým dítětem,
think$_{\mathrm{SG1}}$ that if$_{\mathrm{SG1}}$ was$_{\mathrm{MASC}}$ with my child,
'I think that if I were with my child, . . . '

Manual annotation is supported by the purpose-built annotation tool *feat*[6] and followed by automatic post-processing (see §6).

4.4 Error categorisation

A typical learner of Czech makes errors all along the hierarchy of theoretically motivated linguistic levels, from graphemics to discourse structure. For practical reasons we emend the input conservatively to arrive at a coherent and well-formed result, without any ambition to produce a stylistically optimal solution, refraining from too loose interpretation. Where a part of the input is not comprehensible, it is marked as such and left without emendation. The taxonomy of errors is based on linguistic categories, complemented by a classification of superficial alternations of the source text, such as missing, redundant, faulty or incorrectly ordered element.

[6] See http://purl.org/net/feat.

4.4.1 Errors at Tier 1

Errors in individual word forms, treated at Tier 1, include misspellings (also diacritics and capitalisation), misplaced word boundaries but also errors in inflectional and derivational morphology and unknown stems — made-up or foreign words. Except for misspellings, all these errors are annotated manually. The result of emendation is the closest correct form, which can be further modified at Tier 2 according to context, e.g. due to an error in agreement or semantic incompatibility of the lexeme. See Table 3 for a list of errors manually annotated at Tier 1. The last three error types (*stylColl*, *stylOther* and *problem*) are used also at Tier 2.

Table 3 Errors at Tier 1

Error type	Description	Example
incorInfl	incorrect inflection	*pracovají* v továrně; bydlím s *matkoj*
incorBase	incorrect word base	lidé jsou moc *mérný*; musíš to *posvětlit*
fwFab	non-emendable, made-up word	pokud nechceš slyšet *smášky*
fwNC	foreign word	váza je na *Tisch*; jsem v *truong*
flex	supplementary flag used with fwFab and fwNC marking the presence of inflection	jdu do *shopa*
wbdPre	prefix separated by a space or preposition without space	musím to *při pravit*; *veškole*
wbdComp	wrongly separated compound	*český anglický* slovník
wbdOther	other word boundary error	*mocdobře*; *atak*; *kdy koli*
stylColl	colloquial form	*dobrej* film
stylOther	bookish, dialectal, slang, hyper-correct	holka s *hnědými očimi*
problem	supplementary label for problematic cases	

Table 4 Errors at Tier 2

Error type	Description	Example
agr	violated agreement rules	to jsou *hezké* chlapci; Jana *čtu*
dep	error in valency	bojí se *pes*; otázka *čas*
ref	error in pronominal reference	dal jsem to jemu i *jejího* bratrovi
vbx	error in analytical verb form or compound predicate	musíš *přijdeš*; kluci *jsou* běhali
rflx	error in reflexive expression	dívá na televizi; Pavel *si* raduje
neg	error in negation	žádný to *ví*; *půjdu ne* do školy
lex	error in lexicon or phraseology	jsem *ruská*; dopadlo to *přírodně*
use	error in the use of a grammar category	pošta je *nejvíc blízko*
sec	secondary error (supplementary flag)	stará se o *našich holčičkách*
stylColl	colloquial expression	viděli jsme *hezký* holky
stylOther	bookish, dialectal, slang, hyper-correct expression	zvedl se mi *kufr*
stylMark	redundant discourse marker	*no*; *teda*; *jo*
disr	disrupted construction	*kratka jakost vyborné ženy*
problem	supplementary label for problematic cases	

The rule of "correct forms only" at Tier 1 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.

4.4.2 Errors at Tier 2

Emendations at Tier 2 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 4 gives a list of error types manually annotated at Tier 2. The automatically identified errors include word order errors and subtypes of the analytical forms error *vbx*.

5 Evaluation of the error mark-up

There is no widely accepted metric evaluating the consistency of annotation of learner corpora. In the current annotation practice of non-native speakers' corpora, it is common to have ill-formed texts tagged by a single annotator, despite problems in reliability and evaluation. A general shift towards multiple annotation of learner corpora is imminent.

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 §5.1 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) also 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.

5.1 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. 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. Tier 0, the original 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 Tier 0 tokens and then calculate differences relative to that tier. When there are multiple tokens on Tier 0 corresponding to a token on the relevant tier, we project the tag on the first Tier 0 token only.[7]

Table 5 summarizes the distribution of selected error tags for a 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 by an asterisk) are used only in the evaluation as a more general error category.[8] 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).

5.2 A pilot annotation

Early in the project, we calculated IAA on a pilot sample. It consisted of 67 texts totalling 9848 tokens, most of them written by native speakers of Russian; the texts are classified according to the CEFRL scale as A2 or B1 (Štindlová, 2011). The sample was corrected and assigned error tags according to the error taxonomy presented above in §4.4 by 14 annotators. They were split into two groups: Annotators A and Annotators B. Each group annotated the whole sample independently. On average each annotator processed 1,475 words in 11 texts. The annotator agreement is reported in Table 5.

5.3 Full corpus

Using the feedback gained from the pilot experiment we improved the annotation manual and 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.[9] The evaluation was extended to all usable texts doubly-annotated so far, i.e. to 1,396 texts totalling 175,234 words.

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. 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.

[7] Note that this is different from our previously reported results (Štindlová et al, 2012a), where we projected the tag to all such tokens. Also note that, in Štindlová et al (2012a) we switched the numbers for *incorInfl* and *incorStem* by mistake.

[8] The error taxonomy is hierarchical – 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*).

[9] For the share of different learner groups according to L1 see Table 2.

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 as a colloquial form (annotating it as *stylColl*). In such cases, the annotation manual might instruct the annotator to prefer a certain tag. However, even though this would lead to a higher IAA, it would conceal the fact that these expressions are perceived differently by different native speakers.

Table 5 Inter-annotator agreement on selected tags

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/preposition)	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
vbx	Complex verb form error	0.13	15	0.17	233

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** category 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. 6

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 (consequent) errors (*sec*) and – surprisingly – also errors in analytical verb forms / complex predicates (*vbx*) 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 usage 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' emendation differ in their meaning distinctly, the annotators agree in their emendations in most of the cases (2); if the lexemes show semantic proximity, the annotators highly disagree in the emendation and therefore also in the consequent annotation (3).

(2) **T0:** *v pekařství kupuju* ***housenky***
'I buy **caterpillars** in the baker's shop'
T2: A1: … ***housky***$_{\text{LEX}}$ 'buns'
A2: … ***housky***$_{\text{LEX}}$ 'buns'

(3) **T0:** *kdýž se divá na* ***druhý*** *kultury*
'when one looks at **other** cultures'
T2: A1: *když se dívá na* ***druhé***$_{\text{AGR+STYLCOLL}}$ *kultury* 'other'
A2: *když se dívá na* ***jiné***$_{\text{LEX+AGR+STYLCOLL}}$ *kultury* 'different'

Tables 6 and 7 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 B used the *incorInfl* tag instead, far less common were cases when Annotator B assumed the error to be one of the *fw** tags, finally in 574 cases Annotator B used multiple tags, but none of them was *incorBase* (so we cannot say which one of those corresponds to A's *incorBase* tag).

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

- *incorBase* (error in stem) for *incorInfl* (error in inflection)
  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. In quantified NPs, some quantifiers (numerals above five) are syntactical heads followed by the genitive in the nominative and accusative and agreeing attributes in other cases, some quantifiers (numerals below five) are always agreeing attributes, and some quantifiers (e.g. *mnoho* 'many') are always syntactical heads.

Table 6 Confusion matrix on T1 for all data (selected tags)

	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 7 Confusion matrix on T2 for all data (selected tags)

	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	

5.4 Error tags depend on emendation

Analysis of the tagged data (see Table 8) 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 emended form (the target hypothesis). For example, while *agr* has an overall agreement of 0.69, it is 0.82 for identical emendations, but only 0.24 if the target (T2) hypotheses are different. The situation of other tags is similar. See (4) for an example.

(4) **T0:** *a kdyz stratil manzel*
T2: A1: *a když ztratí*$_{\text{AGR}}$ *manžela*$_{\text{DEP}}$
‘and when she loses her husband’
A2: *a když se*$_{\text{RFLX}}$ *ztratil manžel*
‘and when the husband got lost’

However, sometimes annotators arrived to identical emendations, but still interpreted the original text differently, thus labeling it with different error tags. In some cases, this is manifested by different emendations on the lower tier, i.e. T1. For example, consider the expression in (5): 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 layer. A2 interpreted *tezky* as simply as incorrect form, and corrects it directly to *těžké*. Both approaches make sense, and it is difficult to choose between them without knowing more about the language of the speaker.

(5) **T0:** *tezki období*
‘a difficult period’
A1: **T1:** *těžký*$_{\text{incorStem+incoInfl}}$ *období*
T2: *těžké*$_{\text{agr+stylColl}}$ *období*
A2: **T1:** *těžké*$_{\text{incorStem+incoInfl}}$ *období*
T2: *těžké období*

In all these cases, tagging is correct vis-à-vis the selected emendation. Currently, we investigate the impact of emendation on error annotation at the individual tiers, but we can already 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).

Table 8 Tag IAA depends on emendation agreement

	tag	total		same emendations		different emendations	
		κ	avg. tags	κ	avg. tags	κ	avg. tags
T1	incor*	0.88	14380	0.95	12376	0.48	2004
	incorBase	0.82	10780	0.89	9323	0.44	1456
	incorInfl	0.71	4679	0.79	3887	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	1396	0.51	1088	0.20	307
T2	agr	0.69	2622	0.82	2050	0.24	572
	dep	0.58	3064	0.71	2303	0.19	760
	rflx	0.42	141	0.58	98	0.05	43
	lex	0.32	1815	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

5.5 Outline of the possible causes of the annotators disagreement

We can identify the following causes of the annotators' disagreements:

1. Invalid or imprecise annotation scheme: Generally, the annotators' disagreement can be caused by the annotation scheme itself. If it includes invalid tags or misses some necessary tags, or if the definition of a tag misleads the annotator. In the case of trial tagging of a sample of *CzeSL* data, it was problematic in several points, such as poorly distinguished subtypes of word boundary error (*wbd*), fuzzy definition of the error in pronominal reference (*ref*), also in contrast to the *agr* and *dep* types, or 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.

3. Different target hypotheses: 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 emendation on both tiers, makes reasons for possible disagreements explicit.

6 Automatic extension of manual annotation

So far, the annotation is largely a manual enterprise, quite demanding in terms of annotators' time and expertise. We aim to shift much of the burden to automatic tools, either by aiding human annotators and/or following up on their work, or by processing the texts from scratch, without any human involvement.[10] Let us explore the former option first.

Manually emended and error-annotated text can be assigned additional information by automatic tools in the following two ways:

1. As far as the emended text approximates standard language, at least in grammatical correctness, a tagger/lemmatiser can be applied with an error rate similar to that for standard texts (Spoustová et al, 2007). The resulting annotation (morphosyntactic tags and lemmas) can then be projected to the original forms. See §6.1 below.
2. Some manually assigned error tags can be specified in more detail using formal rules. In fact, some of these tags were designed with this aim in mind. Rules for other tags can be completely formalised and the tags can be assigned fully automatically. See §6.2.

The tools for extending manual annotation can also be used to check its quality of manual annotation, especially to identify tags that are probably missing or incorrect (see §6.3).

6.1 Automatic addition of linguistic information

Emended sentences at T2 can be tagged with morphosyntactic categories and lemmas using standard tools (see, e.g. Jelínek, 2008; Jelínek and Petkevič, 2011). Each word is assigned a lemma and a tag from a standard morphological tagset Hajič (2004). Applying standard methods to T1, consisting of forms which may be correct only in isolation and which may also be wrongly ordered, can produce unreliable results.[11] Instead of a fully disambiguated tag and lemma, T1 is tagged using potentially ambiguous morphological analysis of isolated forms in combination with the tag and lemma assigned at T2 as follows:

- If the forms at both tiers are identical, the tag and lemma assigned at T2 is used.
- If the forms are different, but their lemmas are identical, then that lemma and the appropriate tags are used. For example, if the T1 form is *má* 'has' or 'my' and the T2 form is *mou* 'my', we assign *má* the lemma *můj* 'my'.
- If the T1 form's lemma is different from the lemma at T2, the T1 form receives all possible morphological tags. For example, *má* would be labeled both as a verb with the lemma *mít* 'to have' and as the possessive pronoun with the lemma *můj* 'my'.

[10] See also Jelínek et al (2012).

[11] Depending on the quality of the original and the requirements on the result, some learner texts can be tagged or even parsed automatically, see, e.g. de Haan (2000); de Mönnink (2000); Díaz-Negrillo et al (2010).

Error type	Error description	Example
Cap0	capitalisation: incor. lower case	*evropě/Evropě*; *štědrý/Štědrý*
Cap1	capitalisation: incor. upper case	*Staré/staré*; *Rodině/rodině*
Vcd0	voicing assimilation: incor. voiced	*stratíme/ztratíme*; *nabítku/nabídku*
Vcd1	voicing assimilation: incor. vcless	*zbalit/sbalit*; *nigdo/nikdo*
VcdFin0	word-final voicing: incor. voiceless	*kdyš/když*; *vztach/vztah*
VcdFin1	word-final voicing: incor. voiced	*přez/přes*; *pag/pak*
Vcd	voicing: other errors	*protoše/protože*; *hodili/chodili*
Palat0	missing palatalisation (*k,g,h,ch*)	*amerikě/Americe*; *matkě/matce*
Je0	*je/ě*: incorrect *ě*	*ubjehlo/uběhlo*; *Nejvjetší/Největší*
Je1	*je/ě*: incorrect *je*	*vjeděl/věděl*; *vjeci/věci*
Mne0	*mě/mně*: incorrect *mě*	*zapoměla/zapomněla*
Mne1	*mě/mně*: incor. *mně, mňe, mňě*	*mněla/měla*; *rozumněli/rozuměli*
ProtJ0	protethic *j*: missing *j*	*sem/jsem*; *menoval/jmenoval*
ProtJ1	protethic *j*: extra *j*	*jse/se*; *jmé/mé*
ProtV1	protethic *v*: extra *v*	*vosm/osm*; *vopravdu/opravdu*
EpentE0	*e* epenthesis: missing *e*	*domček/domeček*
EpentE1	*e* epenthesis: extra *e*	*rozeběhl/rozběhl*; *účety/účty*

Table 9 Examples of automatically assigned errors on T1

6.2 Automatic extension and modification of error annotation

Some error types can be detected automatically. This is especially true about formal errors at T1, identifiable by a simple comparison of the corresponding forms at T0 and T1. Errors at T2 are more difficult to classify automatically, thus only a limited number of phenomena are tagged this way.

The manually assigned T1 tags include the following three types of errors: wrong form (*incor*), incorrect word boundaries (*wbd*), and neologism or foreign word (*fw*). The automatically assigned 'formal' errors complement these manual tags as an additional dimension of annotation. For example, **chrozba/hrozba* 'threat' is manually annotated as *incorBase* (the *h/ch* error is in the stem), and **każdécho/každého* 'every$_{masc.sg.gen/acc}$' as *incorInfl* (the *h/ch* error is in the *ého* ending). However, in both cases, the correct *h* is incorrectly devoiced, thus the *h/ch* error is annotated as *formVcd1*.[12]

The formal T1 error tags express the way in which a T1 form differs from the original incorrect T0 form. Most of these tags (such as "missing character", "switch error" or even "error in diacritics") only identify surface manifestations. However, a few error types are characterised by linguistic concepts, such as voicing assimilation or palatalisation. It is the possibility of their automatic detection that puts them in the same class with the truly formal error types.

Table 9 provides examples of some currently handled automatically assigned errors on T1. Some errors affect only spelling with no change in pronunciation (capitalisation, diacritics in *dě/tě/ně*, voicing assimilation, etc.). Other errors always affect pronunciation (vowel quantity, *e* epenthesis). Some errors might affect pronunciation in some contexts, but not others (writing *i/y*, the *c/k* substitution).

[12] In Czech phonology, *h* and *ch* [x] act as voicing counterparts.

Most of the T2 error tags are assigned manually, because the variability of incorrect structures is too high to allow for reliable automatic error tagging. Thus, only limited amount of information is added automatically:

- The reflexivity error tag (*rflx*) is added if another type of error concerns a reflexive pronoun.
- Manually assigned error tags for compound verb forms (*vbx*) are sub-divided as errors in: analytical verb forms (*cvf*), phase or modal verbs (*mod*), and copular predicates (*vnp*). The distinction uses lemmas and morphological tags.
- Tags marking deleted and inserted words are added (*odd*, *miss*).
- Word order corrections are tagged (*wo*). The annotator reorders the words as necessary, but does not tag the altered order. The label is assigned automatically to one or more misplaced forms using lemmas and tags on T2.

6.3 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 analyser, it is likely to be an incorrect word which should be emended. Also, if a word was emended 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 emendation and error annotation, but provides a good approximate measure of the quality of annotation and draws the annotator's attention to potential errors.

7 Fully automatic annotation

Despite the benefits of annotators' insight and judgment, manual annotation, or even manual annotation supplemented by automatic annotation, is tedious and costly. On the other hand, automatic tools are more error-prone and cannot produce the sort of sophisticated annotation envisaged in the present project. Aware of these pros and cons, we explore how far we can get without manual annotation. Due to the lack of methods targeting learner texts, we confronted some 'native Czech' tools (two taggers and a spell checker) with ill-formed input.

7.1 Automatic emendation

One of the options to (partially) automate the task of emendation is to use a proofreading tool – a spell checker or a grammar checker. So far, we have experimented with *Korektor* (Richter, 2010), a spell checker that has some functionalities of a grammar checker, using a combination of lexicon, morphology and a syntax model.[13]

[13] Flor and Futagi (2011) 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.

The tool was tested on a subset of the pilot set of annotated texts (see §5.2), produced by learners at intermediate or higher levels of proficiency, yet among the total 9,372 tokens (7,995 tokens excluding punctuation) 918 (10%) were not recognised by the morphological analyser included in a Czech POS tagger (see *Morče* in Spoustová et al, 2007). Even more forms were judged as faulty by the annotators: 1,189 (13%) were corrected in the same way by both annotators at T1 and 1,519 (16%) at T2.

Results of the spell checker were compared with those of the morphological analyser and with forms at T1 and T2, provided both annotators were in agreement. The spell checker was run in three (batch) modes: (i) “autocorrect” (as proofreader), (ii) “remove-diacritics” followed by “diacritics” (as diacritics assigner), and (iii) same as in (ii), followed by “autocorrect”, the latter two to test the hypothesis that diacritics is a frequent source of errors.

Although the morphological analyser includes a guesser, it makes no attempt to correct an unknown word form, only guesses its morphosyntactic tag and lemma. The spell checker is deemed to be successful for a given form if the morphological analyser treats it as unknown and the spell checker suggests a correction, or if the analyser treats the form as known and the spell checker leaves it intact.

Table 10 shows figures for the morphological analyser. The rows give results for the three modes: *autocorrect* (i), *diacritics* for “remove-diacritics” followed by “diacritics” (ii), and *autocorrect + diacritics* for the full sequence (iii). The column “corrected” gives the counts for forms corrected by the spell checker run in the relevant mode. The column “unknown” gives the number of cases where the morphological analyser happens to flag a form corrected by the spell checker as unknown. The results of the analyser are assumed as truth for the purpose of calculating precision (“unknown”/“corrected”) and recall (“unknown”/918, the latter figure representing all forms unknown to analyser).

Precision is not really a fair measure here, because the analyser never flags forms which are correct in isolation but faulty in a context, while the spell checker often manages to use local context to replace a form X with an orthographically close but morphosyntactically quite different for Y: *podlé → podle, jejích → jejich, žit → žít, libí → líbí, ze → že, divá → dívá, drahy → drahý, mel → měl, jích → jich, čine → číně*. Interestingly, diacritics seem to represent a substantial share of problems in learners’ writings, and the preprocessing of the input by the diacritics remover and assigner (iii) means a significant improvement.

Table 10 Comparison with morphological analyser, which identified the total of 918 unknown forms

mode	corrected	unknown	precision	recall	F-measure
autocorrect	1151	888	0.77	0.97	0.86
diacritics	1176	795	0.68	0.87	0.76
autocorrect + diacritics	1315	906	0.69	0.99	0.81

Corrections made by the annotators can be compared verbatim with those proposed by the spell checker. The spell checker scores whenever the form proposed by the relevant mode matches the form at T1 or T2, respectively. The two annotators must agree about the corrected form, only then it is seen as fit for comparison.

At T1 the total number of corrections (1189) is higher than the number of forms unknown to the morphological analyser (918) because the annotators correct also misspellings which look like homographs with an existing form. Such faulty forms are never detected by the morphological analyser. As a result, recall of the spell checker is lower when its performance is compared with T1 than when with the results of the morphological analyser. Precision stays roughly the same as in the previous comparison because in one aspect T1 is similar to the analyser: it still largely abstracts from context. E.g. annotators are instructed to leave errors due to missed grammatical concord for T2. The data are shown in Table 11 – the column "corrected" is identical to that in Table 10, but the "wrong" column shows the number of cases where the two annotators agree about an emended form, identical to the suggestion of the spell checker.

Table 11 Comparison with corrections at T1, where annotators agreed on the total of 1189 wrong forms

mode	corrected	wrong	precision	recall	F-measure
autocorrect	1151	846	0.74	0.71	0.72
diacritics	1176	780	0.66	0.66	0.66
autocorrect + diacritics	1315	904	0.69	0.76	0.72

It is interesting to investigate cases where the spell checker does not agree with the annotators, but both the spell checker and the annotators indicate an error (170 such cases at T1 for the *autocorrect + diacritics* mode). In some of these cases, the simple *autocorrect* mode without the diacritics component fares better (in 30 cases out of 170). It seems that removing and reassigning diacritics takes the spell-checker too far (Table 12). In some cases the T1 and T2 versions differ and none of the methods matches the contextually correct version of T2 (*pláž, lépe*).

Table 12 Where simple autocorrect mode is better

T0	*autocorrect+diacritics*	T1=*autocorrect*	T2	T2 gloss
plaži	*pláží*	*pláži*	*pláž*	'beach'
tydnů	*týdnů*	*týdnu*	*týdnu*	'week$_{dat/loc}$'
lepšé	*lepše*	*lepší*	*lépe*	'better'
jide	*lidé*	*jde*	*jde*	'goes'
vždicky	*vodičky*	*vždycky*	*vždycky*	'always'

In 150 cases the spell checker suggests a correction when T1 prefers the original, but in 37 cases the spell checker agrees with an annotator at T2 (in 16 cases with both), which means that the real precision is higher. The rest of the cases are mostly inflectional issues, often due to misassigned diacritics, but also quite a few errors in the annotation (shared by both annotators).

T2 is problematic for evaluation in its own right. Some error types handled here are due to wrong word order, style, phraseology and a few other that go beyond simple spell checking, even in a broader sense of some degree of contextual sensitivity. The figures in Table 13, otherwise similar to Table 11, should be interpreted accordingly.

Table 13 Comparison with corrections at T2, where annotators agreed on the total of 1519 wrong forms

mode	corrected	wrong	precision	recall	F-measure
autocorrect	1151	687	0.60	0.45	0.51
diacritics	1176	640	0.54	0.42	0.47
autocorrect + diacritics	1315	745	0.57	0.49	0.53

The two-stage annotation scheme suggests the option to distinguish corrections of forms that are wrong in any context, from those that could be correct in isolation, or in a different context, i.e. to test the grammar-checking capabilities of the spell checker. However, *Korektor* does not quite match the annotation scheme. It is only possible to find a few individual cases of successful corrections of missed agreement or case government (in the order of tens). Again, as in all the previous cases, the mode combining diacritics remover, assigner and proofreader is the best scenario.

The results seem to justify the option to integrate the spell checker into the annotation workflow, even though its suggestions may not quite match the two distinct tiers without tuning to the specific task and annotation scheme. We have already applied *Korektor* in the *autocorrect* mode to all transcribed texts in *CzeSL*, including the texts without any manual annotation.[14]

7.2 Automatic error tagging

For an experiment in automatic tagging we used two taggers, based on different concepts: *Morče* (Votrubec, 2006) uses a morphological analyser, preferring lexical and morphological diagnostics over syntactic context, while *TnT* (Brants, 2000) has the opposite strategy, relying on a lexicon extracted from training data. Both taggers were trained on the same tagset and include a method to handle unknown words. Because of the different strategies the taggers use to tag correct input, they respond differently to various types of deviations. A mutual comparison of their results is thus as interesting as their evaluation against gold standard, which – in the case of ill-formed input – is a difficult concept anyway.

Identifying all errors would involve comparing manual annotations at T2 form-by-form with the original text at T0. In the current absence of such data, we used data

[14] After registration at http://www.korpus.cz/english/dohody.php the result is available for on-line searches as *czesl-plain*, one of the synchronous specialized subcorpora of the Czech National Corpus. See http://www.korpus.cz/english/czesl-plain.php for a description and http://www.korpus.cz/corpora/ for the search interface.

obtained from the easier task of comparing T0 to T1, where all erroneous forms are emended to a closest correct version, disregarding context.

Table 14 presents data extracted from a sample of 93 texts including 12,681 word tokens, with 1,323 tokens (8.9%) identified as ill-formed by the morphological analyser. The two taggers agreed on the same tag in 405 cases, i.e. in 28.8% of the total of ill-formed tokens, and disagreed in 918 cases (71.2%). The figures are additionally split by 12 morphological categories constituting the tag. Column 1 (T0m x T0t) shows in which categories the two taggers disagree at T0 for the 918 tokens, where their tags do not match at least in one category. Agreement is significantly lower between categories largely determined by syntactic context (POS, Gender, Number, Case) as opposed to those determined lexically. Columns 2 (T0m x T1) and 3 (T0t x T1) show agreement rates of tags assigned by *Morče* and *TnT*, respectively, to all tokens at T0[15] in comparison with tags assigned by *Morče* to the corresponding tokens at T1.[16] *Morče* shows better results overall and in most categories. Columns 4 and 5 show agreement rates for an ill-formed subset of the sample used in Columns 2 and 3. Interestingly, *TnT* shows significantly better results, except in the categories of Person and Tense.

Table 14 Tags on T0 and T1 – percentages of agreement

	T0m x T0t	T0m x T1	T0t x T1	T0m x T1	T0t x T1
No. of tokens	918	2589	2589	314	314
Entire tag	0	84.1	79.0	19.1	26.1
POS	39.2	89.6	88.7	43.9	52.5
SubPOS	37.1	89.2	87.9	42.0	49.7
Gender	23.9	88.8	88.2	36.0	46.5
Number	36.9	91.1	91.2	49.0	63.1
Case	31.2	89.0	86.5	43.0	51.3
Possessive Gender	98.6	99.8	99.9	98.4	99.7
Possessive Number	99.5	99.8	99.7	99.0	99.7
Person	68.1	96.3	94.2	81.8	76.1
Tense	70.6	96.7	95.3	83.1	77.4
Grade	78.3	96.4	96.9	75.2	81.5
Negation	74.4	95.3	93.8	73.9	74.2
Voice	70.6	96.7	95.5	83.1	78.7

The difference between the two taggers is also reflected in the share of different POS categories assigned to ill-formed words. Table 15 shows that *Morče* has a more even distribution, but strongly disprefers all verbal categories.

[15] The size of the sample is smaller than in the previous comparison at T0 only due to a more demanding procedure to obtain the data at T1.

[16] The reason why *Morče* was used to tag T1 is because it is currently the best tagger of Czech and we were only interested in the cross-tagger comparison on the ill-formed input at T0.

Table 15 Numbers of tags assigned to ill-formed words

POS	*Morče*	*TnT*	POS	*Morče*	*TnT*	POS	*Morče*	*TnT*
adjective	158	94	noun	499	441	finite verb	32	129
adverb	118	21	preposition	10	–	particle	8	–
gradable adverb	31	11	infinitive	7	41	l-participle	10	119
						passive pcple	1	29

To sum up, the comparison of the two taggers confirms the assumption that the differences in their strategies will have a significant effect on the interpretation of faulty forms. A more general observation concerns the comparison of the success rate of the two taggers on the ill-formed input: *TnT* loses ground in a context with many errors but outperforms *Morče* on faulty forms, while *Morče* strongly disprefers verbs and works better in general.

8 Conclusion

We described a corpus of Czech texts produced by non-native learners of Czech, focussing on error annotation. Results of its evaluation show fair inter-annotator agreement. We also explored and implemented some options of partially or even fully automating the annotation of learner texts.

It is no simple task to design an annotation scheme for a learner corpus and to maintain consistency in the annotated texts, both in a way that would reflect most demands of the corpus users. One of the main reasons is that annotating learner texts tends to be a highly specific enterprise, and even seemingly similar projects do not offer enough guidance – solutions are often too specific to a language or to the project concept and user requirements. On the other hand, annotation itself is quite rewarding due to the plentiful feedback from the annotators about all aspects of the task and, of course, about the learners' interlanguage.

More specifically, our experience shows that the rules for tagging morphosyntactic errors are relatively easy to formalise and it is thus possible to obtain a high inter-annotator agreement for them. However, we were unable to obtain a similarly robust annotation of semantic errors, which are much more dependent on subjective judgement. It is even unclear whether it is desirable to aim to standardize their annotation. Obviously, we should aim to prevent and correct differences that are clear mistakes. Some of it can be done by a better selection of annotators, some by clearer instructions and some by providing better tools to annotators For example, we have seen less errors in the *incorBase* and *incorInfl* errors after integrating a spell-checker into the annotation tool – some of these annotation errors were simply due to annotators overlooking the incorrect word.

The pilot study, where two POS taggers and a spell checker were applied to ill-formed input, confirmed the viability of a partially or even fully automatic annotation as an alternative to manual-only annotation, especially when the demand for large data is higher than concerns about the error rate. It remains to be seen to what extent the comparison of results of multiple taggers, based on different tagging strategies, can lead to usable interpretations of faulty forms.

Acknowledgements The authors are grateful to Tomáš Jelínek and Svatava Škodová for their essential contributions to this work, and also to other members of the project team, namely Milena Hnátková, Vladimír Petkevič, and Hana Skoumalová. This research was supported by the Education for Competitiveness programme, funded by the European Structural Fund and the Czech government as a project no. CZ.1.07/2.2.00/07.0259. It was also co-funded by the GACR grant no. P406/10/P328, and by the NAKI programme of the Czech Ministry of Culture, project no. DF11P01OVV013.

References

Abuhakema G, Feldman A, Fitzpatrick E (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

Bedřichová Z, Šebesta K, Škodová S, Šormová K (2011) Podoba a využití korpusu jinojazyčných a romských mluvčích češtiny: CZESL a ROMi [Form and utilization of a corpus of non-native and Romany speakers of Czech: CZESL and ROMi]. In: Čermák F (ed) Korpusová lingvistika Praha 2011: 2 - Výzkum a výstavba korpusů, Ústav Českého národního korpusu, Nakladatelství Lidové noviny, Praha, Studie z korpusové lingvistiky, vol 15, pp 93–104

Brants T (2000) TnT – a statistical part-of-speech tagger. In: Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000), Seattle, WA

de Cock S (2003) Recurrent sequences of words in native speaker and advanced learner spoken and written english. PhD thesis, Université catholique de Louvain, Louvain-la-Neuve

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1):37–46

Dickinson M (2010) Generating learner-like morphological errors in Russian. In: Proceedings of the 23nd International Conference on Computational Linguistics (COLING-10), Beijing, URL \url{http://jones.ling.indiana.edu/~mdickinson/papers/dickinson-coling10.html}

Díaz-Negrillo A, Fernández-Domínguez J (2006) Error tagging systems for learner corpora. Resla 19:83–102

Díaz-Negrillo A, Meurers D, Valera S, Wunsch H (2010) Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. Language Forum 36(1–2):139–154, URL http://purl.org/dm/papers/diaz-negrillo-et-al-09.html, special Issue on Corpus Linguistics for Teaching and Learning. In Honour of John Sinclair

Fitzpatrick E, Seegmiller S (2001) The montclair electronic language learner database. In: Proceedings of the International Conference on Computing and Information Technologies (ICCIT)

Fitzpatrick E, Seegmiller S (2004) The Montclair electronic language database project. In: Connor U, Upton TA (eds) Applied Corpus Linguistics: A Multidimensional Perspective, Rodopi, p 223–238

Flor M, Futagi Y (2011) Automatic correction of non-word misspellings and generation of learner language corpora. In: Learner Corpus Research 2011 – 20 years of learner corpus research: Looking back, moving ahead, Centre for English Corpus Linguistics, Université catholique de Louvain, Louvain-la-Neuve

Granger S (1999) Use of tenses by advanced EFL learners: Evidence from error-tagged computer corpus. In: Hasselgård H, Oksefjell S (eds) Out of Corpora - Studies in Honour of Stig Johansson, Atlanta, Amsterdam, URL http://hdl.handle.net/2078.1/76322

Granger S (2003a) Error-tagged learner corpora and call: A promising synergy. CALICO Journal 20(3):465–480

Granger S (2003b) Error–tagged learner corpora and CALL: A promising synergy. CALICO journal 20:465–480

Granger S (2008) Learner corpora. In: Lüdeling A, Kytö M (eds) Corpus Linguistics. An International Handbook, HSK 29. 1., vol 1, Mouton De Gruyter, Berlin/New York, pp 259–274

de Haan P (2000) Tagging non-native English with the TOSCA-ICLE tagger. In: Mair C, Hundt M (eds) Corpus Linguistics and Linguistic Theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999, Rodopi, Amsterdam, pp 69–80

Hajič J (2004) Disambiguation of Rich Inflection (Computational Morphology of Czech). Karolinum, Charles University Press, Prague

Hana J, Rosen A, Škodová S, Štindlová B (2010) Error-tagged learner corpus of Czech. In: Proceedings of the Fourth Linguistic Annotation Workshop, Association for Computational Linguistics, Uppsala, Sweden, URL http://utkl.ff.cuni.cz/~rosen/public/hanaetal_law2010.pdf

Hana J, Rosen A, Štindlová B, Jäger P (2012) Building a learner corpus. In: Calzolari N, Choukri K, Declerck T, Doğan MU, Maegaard B, Mariani J, Odijk J, Piperidis S (eds) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey

Jelínek T (2008) Nové značkování v Českém národním korpusu [A new tagging system in the Czech National Corpus]. Naše řeč 91:13–20

Jelínek T, Petkevič V (2011) Systém jazykového značkování korpusů současné psané češtiny [A system of linguistic markup of corpora of contemporary written Czech]. In: Petkevič V, Rosen A (eds) Korpusová lingvistika Praha 2011: 3 – Gramatika a značkování korpusů, Ústav Českého národního korpusu, Nakladatelství Lidové noviny, Praha, Studie z korpusové lingvistiky, vol 16, pp 154–170

Jelínek T, Štindlová B, Rosen A, Hana J (2012) Combining manual and automatic annotation of a learner corpus. In: Sojka P, Horák A, Kopeček I, Pala K (eds) Text, Speech and Dialogue – Proceedings of the 15th International Conference TSD 2012, no. 7499 in Lecture Notes in Computer Science, Springer, pp 127–134

Kisselev O (2012) Heritage language learning: A corpus-based inquiry. Sixth Heritage Language Research Institute

Leech G (1998) Preface. In: Granger S (ed) Learner English on Computer, Addison Wesley Longman, London, p xiv–xx

Leńko-Szymańska A (2004) Demonstratives as anaphora markers in advanced learners' English. In: G Aston SBDS (ed) Corpora and Language Learners, John Benjamins, Amsterdam, p 89–107

Lüdeling A (2008) Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In: Grommes P, Walter M (eds) Fortgeschrittene Lernervarietäten, Niemeyer, Tübingen, p 119–140

Meurers D (2009) On the automatic analysis of learner language: Introduction to the special issue. CALICO Journal 26(3):469–473, URL http://purl.org/dm/papers/meurers-09.html

de Mönnink I (2000) Parsing a learner corpus? In: Mair C, Hundt M (eds) Corpus Linguistics and Linguistic Theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999, Rodopi, Amsterdam, pp 81–90

Nesselhauf N (2005) Collocations in a Learner Corpus. John Benjamins, Amsterdam

Pavlenko A, Hasko V (2007) Russian emotion vocabulary in American learners' narratives. The Modern Language Journal 91:213–234

Pravec NA (2002) Survey of learner corpora. ICAME Journal 26:81–114

Richter M (2010) Pokročilý korektor češtiny [An advanced spell checker of Czech]. Master's thesis, Faculty of Mathematics and Physics, Charles University, Prague

Ringbom H (1998) Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In: Granger S (ed) Learner English on Computer, Longman, Harlow, p 41–52

Rozovskaya A, Roth D (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–Champ, URL http://cogcomp.cs.illinois.edu/page/publication_view/212

Selinker L (1972) Interlanguage. IRAL 10:209–231

Spoustová D, Hajič J, Votrubec J, Krbec P, Květoň P (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, Association for Computational Linguistics, Praha, Czechia, pp 67–74

Stritar M (2009) Slovene as a foreign language: The pilot learner corpus perspective. Slovenski jezik – Slovene Linguistic Studies 7:135–152

Šebesta K (2010) Korpusy češtiny a osvojování jazyka [Corpora of Czech and language acquistion]. Studie z aplikované lingvistiky/Studies in Applied Linguistics 1:11–34

Štindlová B (2011) Evaluace chybové anotace v žákovském korpusu češtiny [Evaluation of error mark-up in a learner corpus of Czech]. PhD thesis, Charles University, Faculty of Arts, Prague

Štindlová B, Škodová S, Hana J, Rosen A (2012a) CzeSL – an error tagged corpus of Czech as a second language. In: Pęzik P (ed) PALC 2011 – Practical Applications in Language and Computers, Lódż 13–15 April 2011, Peter Lang, Łódź Studies in Language, URL http://utkl.ff.cuni.cz/~rosen/public/2011-czesl-palc.pdf, to appear

Štindlová B, Škodová S, Hana J, Rosen A (2012b) Proceedings of The Learner Corpus Research 2011, Louvain-la-Neuve, 15–17. September 2011. In: A learner corpus of Czech: current state and future directions, Presses Universitaires de Louvain,

Louvain-la-Neuve, Corpora and Language in Use, in print

Štindlová B, Škodová S, Rosen A, Hana J (2012c) Annotating foreign learners' Czech. In: Ziková M, Dočekal M (eds) Slavic Languages in Formal Grammar. Proceedings of FDSL 8.5, Brno 2010, Peter Lang, Frankfurt am Main, pp 205–219

Tetreault J, Chodorow M (2008) Native judgements of non-native usage: Experiments in preposition error detection. In: COLING Workshop on Human Judgements in Computational Linguistics, Manchester

Van Rooy B, Schäfer L (2003) An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus. In: Archer D, Rayson P, Wilson A, McEnery T (eds) Proceedings of the Corpus Linguistics 2003 Conference, UCREL, Lancaster University, Lancaster, p 835–844

Votrubec J (2006) Morphological tagging based on averaged perceptron. In: WDS'06 Proceedings of Contributed Papers, Matfyzpress, Charles University, Praha, Czechia, pp 191–195

Waibel B (2008) Phrasal verbs. German and Italian learners of English compared. VDM, Saarbrücken

Xiao R (2008) Well-known and influential corpora. In: Lüdeling A, Kytö M (eds) Corpus Linguistics. An International Handbook, Handbooks of Linguistics and Communication Science [HSK] 29.1, vol 1, Mouton de Gruyter, Berlin and New York, pp 383–457