A learner corpus of Czech: Current state and future directions

Barbora Štindlová • Svatava Škodová
Technical University of Liberec, Faculty of Education
Alexandr Rosen • Jirka Hana
Charles University Prague, Faculty of Arts • Faculty of Mathematics and Physics

Abstract

The paper describes *CzeSL*, a learner corpus of Czech, together with its design properties. We start with a brief introduction of the project within the context of *AKCES*, a program addressing Czech acquisition corpora; in connection with the programme we are also concerned with the groups of respondents, including differences due to their L1; further we comment on the choice of the sociocultural metadata recorded with each text and related both to the learner and the text production task. Next we describe the intended uses of *CzeSL*. The core of the paper deals with transcription and annotation. We explain issues involved in the transcription of handwritten texts and present the concept of a multi-level annotation scheme including a taxonomy of captured errors. We conclude by mentioning results from an evaluation of the error annotation and presenting plans for future research.

Keywords: learner corpus, Slavic languages, Czech, error annotation, error taxonomy, multi-level annotation.

1. Introduction – A learner corpus of Czech

The first learner corpus of *Czech as a Second Language* (*CzeSL*),[1] with its size of 2 million words the only large learner corpus for a Slavic language at the moment,[2] is built as a joint project of Technical University Liberec and Charles University Prague. It is a part of the programme *Acquisition Corpora of Czech* (*AKCES*), pursued at Charles University in Prague since 2005 (Šebesta 2010). In addition to *CzeSL*, *AKCES* includes the following subcorpora: (i) *SCHOLA 2010*[3] and *EDUCO* – recordings and transcripts capturing the language of Czech pupils attending primary school classes (about 800,000 words each, finished); (ii) *SKRIPT* – written texts produced by Czech

[1] 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.

[2] To the best of our knowledge, there is only one learner corpus built for a Slavic language – *PiKUST* (Stritar 2009), a corpus of Slovene as a foreign language. However, it is of a modest size of 35,000 words, and its error annotation is adopted from the Norwegian project *ASK* (Tenfjord 2009).

[3] More details and a search interface are available at http://ucnk.ff.cuni.cz/schola.php.

students (about 600,000 words so far, in development); (iii) *ROMi* – texts and speech produced by young learners with Romani background (in development); and (iv) *IUVENT* – spoken corpus of language produced by young native Czechs (planned). A consistent methodology and set of tools used throughout the program represent a significant synergic effect, allowing for comparative analyses of native and non-native language in the corpora, built on identical principles.[4]

CzeSL is focused on four main groups of learners of Czech: (1) speakers of related Slavic languages, represented mainly by Russian, other Eastern Slavic languages and Polish, (2) speakers of other Indo-European languages, with a slight majority of German, (3) speakers of distant non-Indo-European languages, mainly Chinese, Vietnamese and Arabic, and (4) pupils in primary school age with Romani background.[5] About one half (1 million words) of the corpus consists of short essays written by non-native learners (1–3), while short essays written by Romani pupils (4) account for 25%. Theses for a university degree, written by non-native students, represent the remaining 25% of the corpus. Approximately 20% (300 thousand words) of the short essays are corrected and error-annotated.

In addition to its concern with several representative groups of speakers, *CzeSL* strives to cover as much ground as possible also in other aspects. This wide-scope design property, offering extensive data with rich annotation and metadata, is meant to serve a number of different users and to satisfy varied research requirements: (a) *CzeSL* consists of both spoken and written texts, produced during a range of situations throughout the language learning process, collected as manuscripts and transcribed into an electronic format. The transcription follows rules designed to preserve important features of handwritten texts such as self-corrections (see Štindlová 2011: 106). Apart from originally handwritten texts, *CzeSL* also includes Bachelors', Masters' and doctoral theses, written in Czech by non-native students and collected in an electronic format. (b) The data cover all language levels according to the Common European Framework of Reference for Languages (CEFR, 2001), from real beginners (level A1) to more advanced learners (level B2 and higher), with a balanced mix of levels as much as possible, although levels B1 and B2 prevail over the lower grades due to their easier availability. (c) The texts are elicited in various situations; they are not restricted to parts of written or oral examination, or to argumentative or reflective essays, as in many other learner corpora. (d) *CzeSL* also includes texts, collected at regular intervals from learners attending long-term language courses. All texts of a specific author can be retrieved using metadata (see below). This will support analyses

[4] In this paper we use the term *second language* as denoting any language learned after 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. Some authors (*e.g.* Ellis 1994) use the terms second and foreign language in the same way as we do here, while others (*e.g.* Günther & Günther 2007) use them as complementary. *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. This information is included in the metadata.

[5] Sometimes it is difficult to decide whether Czech is the first or second language of these children. Yet the sociocultural differences between the non-Roma and some Roma communities in the Czech Republic are such that the linguistic development of Roma children may 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 will become a separate component of *AKCES*.

of temporal development of the author's interlanguage, providing the option of using some parts of *CzeSL* for longitudinal research.

Each text is equipped with detailed metadata records, for a total of 18 parameters. Some of them (12) relate to the respondent, while the remaining 6 specify the character of the text and circumstances of its production. All texts in the corpus produced by non-native speakers of Czech are assigned basic sociological data about the learners, such as age, gender, and language background (first language). Other obligatory variables describe (1) their proficiency level in Czech according to the CEFR, (2) conditions of the process of acquisition of Czech, including an indication of the institution, duration, or location – whether abroad or in the Czech Republic, (3) the textbooks used in learning Czech. More variables may be specified as an option,[6] such as the learner's knowledge of other (non-native) languages, her bilingual competence, length of stay in the country, or whether a family member has been a Czech speaker. In addition to standard metadata specifying temporal and size restrictions, we also register the availability of language reference tools and the extent and type of elicitation.

In Section 2, we sketch the intended use of the corpus, proceed to issues involved in the transcription of handwritten texts (Section 3) and argue for the specific design of our annotation scheme in Section 4. The concluding section, Section 5, deals with the error taxonomy.[7]

2. Intended use

Despite the fact that teaching Czech as a second language has acquired the status of a well-established field with a long tradition, a proper teaching methodology is not developed and available. Teachers often cope with this situation by adopting two possible strategies: (i) They have recourse to methods and techniques used for other languages, such as English or German. Since Czech is typologically different, mainly due to its rich morphology, this approach is grossly inadequate for training students in the use of rules of Czech grammar. (ii) They tend to transfer their detailed and fairly academic knowledge of Czech grammar to foreigners, in a way poorly structured for such a task, confusing the issues of presenting grammar to native and foreign students.

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. Primary school teachers receive no training in teaching Czech as a foreign language, again resorting to an individual and intuitive approach. By its inclusion in *AKCES*, *CzeSL* will become 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 (so far, there are no monographs available dealing with this topic).

---

[6] The role of such variables has been emphasized by several authors (*e.g.* Granger 2003; 2008; Tono 3003).

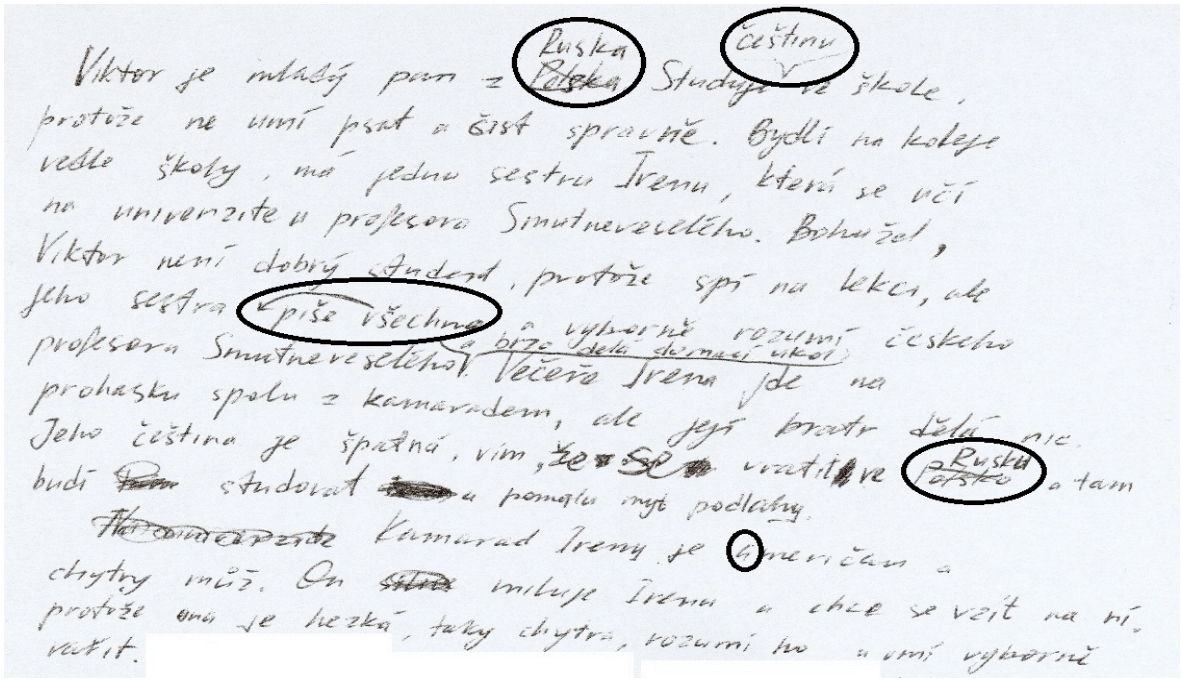
[7] For more details about some technical aspects of the compilation of the *CzeSL* corpus see Hana *et al.* (2012) and Jelínek *et al.* (2012).

The programme *Czech as a foreign language* has been available only recently as a three-year BA course at Technical University in Liberec (TU), and as a two-year MA course at Charles University in Prague (CU). Texts collected for *CzeSL* are already in use in the training of teachers both at TU and CU 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.

In the following, the design and structure of the corpus is presented in more detail.

3. Transcription

Since most original texts are handwritten,[8] they are transcribed according to detailed rules using off-the-shelf tools (*e.g.*, Open Office Writer or Microsoft Word). A set of codes is used to capture the author's corrections and other properties of the manuscript (*e.g.*, for future research of handwriting of students with a different native writing system, for investigating the process of language acquisition, or to enable multiple interpretation – the same glyph may be interpreted as *i* in the handwriting of one student, *e* of another, and *a* of yet another). An additional reason for collecting handwritten texts is to avoid the use of a spell checker, because the result would not reflect the student's skills. In a highly inflectional language such as Czech, deviations in spelling very often do not only reflect wrong graphemics, but also indicate errors in morphology. In Figure 1 a sample text is presented with some of the author's self-corrections in circles.[9]



[8] Electronic texts (BA, MA and Ph.D. theses) represent a minority.

[9] For more details about the rules of transcription see Štindlová (2011).

Figure 1: A sample handwritten text

The handwritten text is transcribed as in Figure 2. Boldface highlights specific features of handwriting – self-corrections (deletion ~~*Polska*~~, insertion *{češtinu}<in>*, text movement *{piše všechno -> všechno piše}*), and transcription codes in angle brackets. For example, the author replaced the form *Polska* 'Poland' by *Ruska* 'Russia', which is transcribed as strikeout text.

> Viktor je mladý pan z **~~Polska~~** Ruska. Studuje **{češtinu}<in>** ve škole, protože ne umí psat a čist spravně. 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 českeho profesora Smutneveselého **{a brzo delá domací ukol}<in>**. Večeře Irena jde na prohasku spolu z kamaradem, ale její bratr dělá nic. Jeho čeština je špatná, vím, že se vratit ve **~~Polsko~~** Rusk**~~o~~**u a tam budí studovat u pomalu myt podlahy.
> Kamarad Ireny je **{A|a}**meričan a chytry můž. On miluje Irenu a chce se vzít na ní. protože ona je hezká, taky chytra, rozumí ho a umí vyborně vařit.

Figure 2: The sample transcribed

4. Annotation scheme

The language of a learner of Czech may deviate from the standard in a number of aspects at the same time: spelling, morphology, morphosyntax, semantics, pragmatics or style. To show some of the options, the transcribed example in Figure 2 is shown again in Figure 3 with the spelling, morphological and morphosyntactic problems identified and the handwriting-specific annotation resolved. Forms wrong in any context (due to an error in spelling or morphology) are set in boldface, forms wrong due to a morphosyntactic or lexical anomaly are underlined. Some forms may be faulty for both reasons; these are in bold and underlined.

> Viktor je mladý pan z Ruska. Studuje češtinu ve škole, protože ne umí **psat** a **čist spravně**. 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 všechno **piše** a **vyborně** rozumí **českeho** profesora **Smutneveselého** a brzo **delá** domací **ukol**. Večeře Irena jde na **prohasku** spolu z **kamaradem**, ale její bratr dělá nic. Jeho čeština je špatná, vím, že se **vratit** ve Rusku a tam budí studovat u pomalu **myt** podlahy.
> **Kamarad** Ireny je Američan a **chytry můž**. On miluje Irenu a chce se vzít na ní. protože ona je hezká, taky **chytra**, rozumí ho a umí **vyborně** vařit.

Figure 3: Spelling, morphological and morphosyntactic problems identified

The highlighted and underlined parts are incorrect. The transcribed version (Figure 3) and a corrected version are shown in Table 1. Corrected characters in italics and

glossed translations are included for the benefit of the reader; they are not a part of the corpus.

Viktor je mladý pan z Ruska. [Viktor is a young Mr. from Russia.]	Viktor je mladý pán z Ruska. [Viktor is a young man from Russia.]
Studuje češtinu ve škole, protože ne umí **psat** a **čist spravně**. [He studies Czech at school, because he can not **write** and **read correctly**.]	Studuje češtinu ve škole, protože neumí **psát** a **číst správně**. [He studies Czech at school, because he cannot **write** and **read correctly**.]
Bydlí na koleje vedle školy, má jednu sestru Irenu, která se učí na **univerzite** u profesora **Smutneveselého**. [He lives at residence halls$_{GEN}$ next to the school, has one sister Irena, who is a student of professor **Smutněveselý** at the **university**.]	Bydlí na kolej*i* vedle školy, má jednu sestru Irenu, která se učí na **univerzitě** u profesora **Smutněveselého**. [He lives at residence halls$_{LOC}$ next to the school, has one sister Irena, who is a student of professor **Smutněveselý** at the **university**.]
Bohužel, Viktor není dobrý student, protože spí na lekci, ale jeho sestra všechno **piše** a **vyborně** rozumí **českeho** profesora **Smutneveselého** a brzo **delá** domací **ukol**. [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.]	Bohužel, Viktor není dobrý student, protože spí na lekci, ale jeho sestra všechno **píše** a **výborně** rozumí **české*mu*** profesor*ovi* **Smutněveselé*mu*** a brzo **dělá** domácí ***úkoly***. [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.]
Večeře Irena jde na **prohasku** spolu z **kamaradem**, ale její bratr dělá nic. [Dinner Irena goes for a **walk** with her **friend**, but her brother does nothing.]	Večer Irena jde na **procházku** spolu *s* **kamarádem**, ale její bratr *ne*dělá nic. [In the evening Irena goes for a **walk** with her **friend**, but her brother doesn't do anything.]
Jeho čeština je špatná, vím, že se **vratit** ve Rusku a tam budí studovat u pomalu **myt** podlahy. [?] [His Czech is poor, I know that he will to **to return** to Russia and there he wakes study at slowly **wash** floors.]	Jeho čeština je špatná, vím, že se **vrátí** *do* Rusk*a* a tam **bude** studovat *a* pomalu **mýt** podlahy. [?] [His Czech is poor, I know that he will **return** to Russia and there he **will** study and slowly **wash** floors.]
Kamarad Ireny je Američan a **chytry můž**. [Irena's **boyfriend** is an American and a **smart guy**.]	**Kamarád** Ireny je Američan a **chytrý muž**. [Irena's **boyfriend** is an American and a **smart guy**.]
On miluje Irenu a chce se vzít na ní. protože ona je hezká, taky **chytra**, rozumí ho a umí **vyborně** vařit. He loves Irena and wants to **marry on her**, because she is pretty, also **smart**, she understands him$_{GEN}$ and is an **excellent** cook.	On miluje Irenu a chce ***si ji* vzít, protože** ona je hezká, taky **chytrá**, rozumí *mu* a umí **výborně** vařit.¨ He loves Irena and wants to **marry her**, because she is pretty, also **smart**, she understands him$_{DAT}$ and is an **excellent** cook.

Table 1: The transcribed and the corrected versions with translations

To cope with the multi-level options of erring in Czech and to satisfy the goals of the project, our annotation scheme answers the following requirements:

1. Preservation of the original text alongside with the emendations
2. Successive emendations
3. Ability to capture errors in single forms as well as in multi-word discontinuous expressions
4. Syntactic relations as supplementary information for some error types: agreement, valency, pronominal reference
5. Automatic assignment of errors when possible, based on comparing faulty and corrected forms, using morphosyntactic tags, assigned by a tagger

To meet these requirements, we use a multilevel annotation scheme, supporting successive emendations. As a compromise between several theoretically motivated levels and practical concerns about the process of annotation, the scheme offers two annotation levels. This enables the annotators to register anomalies in isolated forms separately from the annotation of context-based phenomena but saves them from difficult theoretical dilemmas.

Level 0 is the level of transcribed input, where the words represent the original strings of graphemes, with some properties of the handwritten original preserved in the mark-up. Level 1 gives orthographical and morphological emendation of isolated forms as a text consisting of existing Czech forms; the sentence as a whole can still be incorrect. 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 Level 2, where all other types of deviations are treated, resulting in a grammatically correct sentence. This includes errors such as those in syntax (agreement, government), lexicon, word order, usage, style, reference, or negation.

Levels of annotation 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, deleted or added by the annotator. Whenever a word form is emended, the type of error can label the link connecting the incorrect form with its emended version (such as *incorInfl* or *incorBase* for morphological errors in inflectional endings and stems). The sample text is shown in Figure 4 as displayed by the tool used by the annotators.[10]

The whole annotation process proceeds as follows:

1. The transcript is converted into the annotation format, where Level 0 roughly corresponds to the tokenized transcript and Level 1 is set as equal to Level 0 by

[10] The tool *feat* (*F*lexible *E*rror *A*nnotation *T*ool) is an environment for layered error annotation of learner corpora, see Hana *et al.* (2010). It is freely available from http://purl.org/net/feat.

default. Both are encoded as PML (an XML-based format for structural linguistic annotation, see Pajas & Štěpánek 2006).

2. The annotator manually corrects the document and provides some information about errors using our annotation tool *feat*.
3. Automatic post-processing provides additional information about lemma, part-of-speech and morphological categories for emended forms.
4. Error information that can be inferred automatically is added by comparing original and corrected strings: type of spelling alternation, missing/redundant expression, and inappropriate word order.

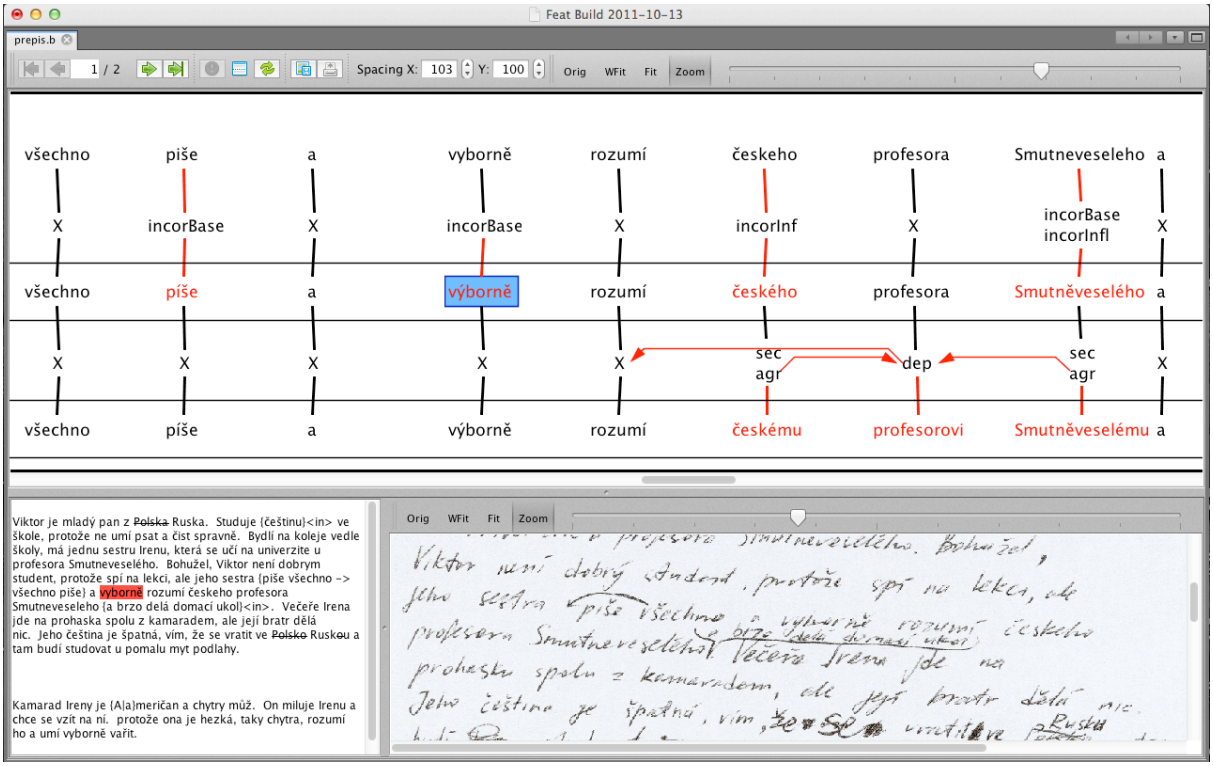


Figure 4: The sample in the annotation tool feat

5. Error taxonomy

The taxonomy of errors is based on previous research of frequent error types and reflects implicitly stated research hypotheses about the acquisition of an inflectional language.[11] To a large extent the taxonomy uses standard linguistic categories, complemented by a classification of superficial alternations of the source text, such as missing, redundant, faulty or incorrectly ordered element. The tagset consists of 22 error tags, 8 for Level 1, 11 for Level 2, and 3 that can be used at both levels. They are

[11] For some taxonomies used in previous projects see, *e.g.*, Díaz-Negrillo *et al.* (2006), Nicholls (2003), Izumi *et al.* (2005), or Granger (2003a).

supplemented by tags generated automatically by comparing original with emended forms and manually assigned tags.

Errors in individual word forms, treated at Level 1 (see Table 2), 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.

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, "fabricated" word	pokud nechceš slyšet *smášky*
fwNC	foreign word	váza je na *Tisch*; jsem v *truong*
flex	with *fwFab* and *fwNC*: inflected	jdu do *shopa*
wbdPre	word boundary: prefix or preposition	musím to *při pravit*; *veškole*
wbdComp	word boundary: compound	*český anglický* slovník
wbdOther	other word boundary error	*mocdobře*; *atak*; *kdy koli*
stylColl	colloquial form	*dobrej* film
stylOther	bookish, dialectal, hypercorrect	holka s *hnědými očimi*
problem	problematic cases	

Table 2: Errors at Level 1

Emendations at Level 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 3 gives a list of error types manually annotated at Level 2. The automatically identified errors include word order errors and subtypes of the error in analytical verb forms (*vbx*).

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 or compound verb form	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	stará se o *našich holčičkách*
stylColl	colloquial expression	viděli jsme *hezký* holky
stylOther	bookish, dialectal, hypercorrect	rozbil se mi *hadr*
stylMark	redundant discourse marker	*no*; *teda*; *jo*
disr	disrupted construction	*kratka jakost vyborné ženy*
problem	problematic cases	

Table 3: Errors at Level 2

Rather than aiming at perfect Czech, we emend the input conservatively, modifying incorrect and inappropriate forms and expressions to arrive at a coherent and well-formed result, without any ambition to produce a stylistically optimal solution. Word order, for example, is corrected only when the input is ungrammatical.

Overall, we are convinced that annotation guided by formal criteria is useful at least as a base for comparison with native speakers' language, automatic (error) annotation, and for annotating communicative adequacy, style, etc. in the future. As a further step towards a common ground for the comparison and guidance for the annotators, grammatical and lexical aspects of the learner language are emended and tagged to conform to the rules of Standard Czech.

A doubly annotated sample (10,000 word forms) 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. For more details see Štindlová *et al.* (2012).

6. Conclusion and outlook

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 optimal presentation of the Czech language for non-native speakers. 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.

In order to support varied types of use and to maintain consistency of annotation, we opted for an annotation scheme and error taxonomy based on grammatical deviations from the standard, without one specific focus. This strategy fits well with the typological properties of Czech and allows for extensions now or in the future – both

into new domains of annotated phenomena, and into more efficient annotation processes, such as automatic assignment of more detailed error categories (implemented), automatic morphological analysis (implemented) and syntactic analysis or semi-automatic emendation and error tagging using a spell and grammar checker, integrated with the annotation tool (in preparation).

References

CEFR (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Applied Linguistics Non Series. Cambridge University Press.

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

Ellis, R. (1994). *The Study of Second Language Acquisition*. OUP, Oxford, 12.

Granger, S. (2003). The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly.* 37(3):538–545.

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

Granger, S. (2008). Learner corpora. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics*. *An International Handbook*. Mouton De Gruyter, Berlin/New York, 259–274.

Günther, B. and Günther P. (2007). *Erstsprache, Zweitsprache, Fremdsprache: Eine Einführung.* Beltz Pädagogik, Belz.

Hana, J., Rosen, A., Škodová, S. and Štindlová, B. (2010). Error-tagged learner corpus of Czech. In *Proceedings of the Fourth Linguistic Annotation Workshop.* Uppsala, Sweden: Association for Computational Linguistics, 11–19.

Hana, J., Rosen, A., Štindlová, B., and Jäger, P. (2012). Building a learner corpus. In Calzolari, N. *et al*., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Günther, B., Günther P. (2004). Erstsprache und Zweitsprache: Einführung aus pädagogischer Sicht. Weinheim, Basel, Beltz.

Izumi, E., Uchimoto, K., and Isahara, H. (2005). Error annotation for corpus of Japanese learner English. In *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora* (LINC 2005), Korea, 71–80.

Jelínek, T., Štindlová, B., Rosen, A., and Hana, J. (2012). Combining manual and automatic annotation of a learner corpus. In *Proceedings of the 15th International Conference TSD 2012*. To appear.

Nicholls, D. (2003). The Cambridge Learner Corpus: error coding and analysis for lexicography and ELT. *Proceedings of the Corpus Linguistics 2003 Conference, 28–31 March.* Lancaster, 572–581.

Pajas, P. and Štěpánek, J. (2006). XML-Based Representation of Multi-Layered Annotation in the PDT 2.0. In R. E. Hinrichs, N. Ide, M. Palmer, J. Pustejovsky (eds.) *Proceedings of LREC 2006 Workshop on Merging and Layering Linguistic Information, Genoa, Italy: ELRA*, 40–47.

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]*. Doctoral dissertation, Charles University, Faculty of Arts, Prague.

Štindlová, B., Rosen, A., Hana, J. and Škodová, S. (2012). CzeSL – An error tagged corpus of Czech as a second language. In P. Pęzik (ed.) *PALC 2011 – Practical Applications in Language and Computers (Łódź 13–15 April 2011), Łódź Studies in Language.* Peter Lang.

Tenfjord, K., Hagen, J. E., and Johansen, H. (2009). Norsk andrespråkskorpus (ASK) – design og metodiske forutsetninger. *NOA norsk som andrespråk*, 25(1): 52–81.

Tono, Y. (2003). Learner corpora: design, development and applications. In *Proceedings of the Corpus Linguistics 2003 Conference.* Lancaster, United Kingdom, 800–809.