Error-tagged Learner Corpus of Czech

Jirka Hana
Charles University
Prague, Czech Republic
first.last@gmail.com

Alexandr Rosen
Charles University
Prague, Czech Republic
alexandr.rosen@ff.cuni.cz

Svatava Škodová
Technical University
Liberec, Czech Republic
svatava.skodova@tul.cz

Barbora Štindlová
Technical University
Liberec, Czech Republic
barbora.stindlova@tul.cz

Abstract

The paper describes a learner corpus of Czech, currently under development. The corpus captures Czech as used by non-native speakers. We discuss its structure, the layered annotation of errors and the annotation process.

1 Introduction

Corpora consisting of texts produced by non-native speakers are becoming an invaluable source of linguistic data, especially for foreign language educators. In addition to morphosyntactic tagging and lemmatisation, common in other corpora, learner corpora can be annotated by information relevant to the specific nonstandard language of the learners. Cases of deviant use can be identified, emended and assigned a tag specifying the type of the error, all of which helps to exploit the richness of linguistic data in the texts. However, annotation of this kind is a challenging tasks, even more so for a language such as Czech, with its rich inflection, derivation, agreement, and largely information-structure-driven constituent order. A typical learner of Czech makes errors across all linguistic levels, often targeting the same form several times.

The proposed annotation scheme is an attempt to respond to the requirements of annotating a deviant text in such a language, striking a compromise between the limitations of the annotation process and the demands of the corpus user. The three-level format allows for successive emendations, involving multiple forms in discontinuous sequences. In many cases, the error type follows from the comparison of the faulty and corrected forms and is assigned automatically, sometimes using information present in morphosyntactic tags, assigned by a tagger. In more complex cases, the scheme allows for representing relations making phenomena such as the violation of agreement rules explicit.

After an overview of issues related to learner corpora in §2 and a brief introduction to the project of a learner corpus of Czech in §3 we present the concept of our annotation scheme in §4, followed by a description of the annotation process in §5.

2 Learner corpora

A learner corpus, also called interlanguage or L2 corpus, is a computerised textual database of language as produced by second language (L2) learners (Leech, 1998). Such a database is a very powerful resource in research of second language acquisition. It can be used to optimise the L2 learning process, to assist authors of textbooks and dictionaries, and to tailor them to learners with a particular native language (L1).

More generally, a learner corpus – like other corpora – serves as a repository of authentic data about a language (Granger, 1998). In the domain of L2 acquisition and teaching of foreign languages, the language of the learners is called *interlanguage* (Selinker, 1983).[1] An interlanguage includes both correct and deviant forms. The possibility to examine learners' errors on the background of the correct language is the most important aspect of learner corpora (Granger, 1998).

Investigating the interlanguage is easier when the deviant forms are annotated at least by their correct counterparts, or, even better, by tags making the nature of the error explicit. Although

[1] *Interlanguage* is distinguished by its highly individual and dynamic nature. It is subject to constant changes as the learner progresses through successive stages of acquiring more competence, and can be seen as an individual and dynamic continuum between one's native and target languages.

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 11–19,
Uppsala, Sweden, 15-16 July 2010.

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.

Learner corpora can be classified according to several criteria:

- Target language (TL): 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.
- Medium: Learner corpora can capture written or spoken texts, the latter much harder to compile, thus less common.
- L1: The data can come from learners with the same L1 or with various L1s.
- Proficiency in TL: Some corpora gather texts of students at the same level, other include texts of speakers at various levels. Most corpora focus on advanced students.
- Annotation: Many learner corpora contain only raw data, possibly with emendations, without linguistic annotation; some include part-of-speech (POS) tagging. Several include error tagging. Despite the time-consuming manual effort involved, the number of error-tagged learner corpora is growing.

Error-tagged corpora use the following taxonomies to classify the type of error:

- Taxonomies marking the source of error: The level of granularity ranges from broad categories (morphology, lexis, syntax) to more specific ones (auxiliary, passive, etc.).
- Taxonomies based on formal types of alternation of the source text: omission, addition, misformation, mis-ordering.
- Hierarchical taxonomies based on a combination of various aspects: error domain (formal, grammatical, lexical, style errors), error category (agglutination, diacritics, derivation inflection, auxiliaries, gender, mode, etc.), word category (POS).
- Without error taxonomies, using only correction as the implicit explanation for an error.

In Table 1 we present a brief summary of existing learner corpora tagged by POS and/or error types, including the size of the corpus (in millions of words or Chinese characters), the mother tongue of the learners, or – in case of learners with different linguistic backgrounds – the number of mother tongues (L1), the TL and the learners' level of proficiency in TL. For an extensive overview see, for example (Pravec, 2002; Nesselhauf, 2004; Xiao, 2008).

Size	L1	TL	TL proficiency
ICLE *– Internat'l Corpus of Learner English*			
3M	21	English	advanced
CLC *– Cambridge Learner Corpus*			
30M	130	English	all levels
PELCRA *– Polish Learner English Corpus*			
0.5M	Polish	English	all levels
USE *– Uppsala Student English Corpus*			
1.2M	Swedish	English	advanced
HKUST *– Hong Kong University of Science and Technology Corpus of Learner English*			
25M	Chinese	English	advanced
CLEC *– Chinese Learner English Corpus*			
1M	Chinese	English	5 levels
JEFLL *– Japanese EFL Learner Corpus*			
0.7M	Japanese	English	advanced
FALKO *– Fehlerannotiertes Lernerkorpus*			
1.2M	various	German	advanced
FRIDA *– French Interlanguage Database*			
0.2M	various	French	intermediate
CIC *– Chinese Interlanguage Corpus*			
2M	96	Chinese	intermediate

Table 1: Some currently available learner corpora

3 A learner corpus of Czech

In many ways, building a learner corpus of Czech as a second/foreign language is a unique enterprise. To the best of our knowledge, the CzeSL corpus (Czech as a Second/Foreign Language) is the first learner corpus ever built for a highly inflectional language, and one of the very few using multi-layer annotation (together with FALKO – see Table 1). The corpus consists of 4 subcorpora according to the learners' L1:

- The Russian subcorpus represents an interlanguage of learners with a Slavic L1.
- The Vietnamese subcorpus represents a numerous minority of learners with very few points of contact between L1 and Czech.
- The Romani subcorpus represents a linguistic minority with very specific traits in the Czech cultural context.
- The "remnant" subcorpus covers texts from speakers of various L1s.

The whole extent of CzeSL will be two million words (in 2012). Each subcorpus is again divided

into two subcorpora of written and spoken texts;[2] this division guarantees the representative character of the corpus data. The corpus is based on texts covering all language levels according to the Common European Framework of Reference for Languages, from real beginners (A1 level) to advanced learners (level B2 and higher). The texts are elicited during various situations in classes; they are not restricted to parts of written examination. This spectrum of various levels and situations is unique in the context of other learner corpora.

Each text is equipped with the necessary background information, including sociological data about the learner (age, gender, L1, country, language level, other languages, etc.) and the situation (test, homework, school work without the possibility to use a dictionary, etc.).

4 Annotation scheme

4.1 The feasible and the desirable

The error tagging system for CzeSL is designed to meet the requirements of Czech as an inflectional language. Therefore, the scheme is:

- Detailed but manageable for the annotators.
- Informative – the annotation is appropriate to Czech as a highly inflectional language.
- Open to future extensions – it allows for more detailed taxonomy to be added in the future.

The annotators are no experts in Czech as a foreign language or in 2L learning and acquisition, and they are unaware of possible interferences between languages the learner knows. Thus they may fail to recognise an interferential error. 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'.[3] Similarly in *Je tam hodně sklepů* 'There are many cellars.' The formally correct sentence may strike the reader as implausible in the context, but it is impossible to identify and emend the error without the knowledge that *sklep* in Russian means 'grave', not 'cellar' (= *sklep* in Czech).

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, unlike the latter. It is difficult to provide clear guidelines.

Errors in word order represent another specific type. Czech constituent order reflects information structure and it is sometimes difficult 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 in the room, although the more likely interpretation is that among other things, there is also a radio, which happens to sit on the wardrobe. Only the latter interpretation would require a different word order: *Taky je na skříni rádio*. Similarly difficult may be decisions about errors labelled as **lexical** and **modality**.

The phenomenon of Czech diglossia is reflected in the problem of annotating non-standard language, usually individual forms with colloquial morphological endings. The learners may not be aware of their status and/or an appropriate context for their use, and the present solution assumes that colloquial Czech is emended under the rationale that the author expects the register of his text to be perceived as unmarked.

On the other hand, there is the primary goal of the corpus: to serve the needs of the corpus users. The resulting error typology is a compromise between the limitations of the annotation process and the demands of research into learner corpora.

The corpus can be used for comparisons among learner varieties of Czech, studied as national interlanguages (Russian, Vietnamese, Romani etc.) using a matrix of statistic deviations. Similarly interesting are the heterogeneous languages of learners on different stages of acquisition. From the pedagogical point of view, corpus-based analyses have led to a new inductive methodology of data-driven learning, based on the usage of concordances in exercises or to support students' independent learning activities.

4.2 The framework

Annotated learner corpora sometimes use data formats and tools developed originally for annotating speech. Such environments allow for an arbitrary

[2] Transcripts of the spoken parts will be integrated with the rest of the corpus at a later stage of the project.

[3] All examples are authentic.

segmentation of the input and multilevel annotation of segments (Schmidt, 2009). Typically, the annotator edits a table with columns corresponding to words and rows to levels of annotation. A cell can be split or more cells merged to allow for annotating smaller or larger segments. This way, phenomena such as agreement or word order can be emended and tagged (Lüdeling et al., 2005).

However, in the tabular format vertical correspondences between the original word form and its emended equivalents or annotations at other levels may be lost. It is difficult to keep track of links between forms merged into a single cell, spanning multiple columns, and the annotations of a form at other levels (rows). This may be a problem for successive emendations involving a single form, starting from a typo up to an ungrammatical word order, but also for morphosyntactic tags assigned to forms, whenever a form is involved in a multi-word annotation and its equivalent or tag leaves the column of the original form.

While in the tabular format the correspondences between elements at various levels are captured only implicitly, in our annotation scheme these correspondences are explicitly encoded. Our format supports the option of preserving correspondences across levels, both between individual word forms and their annotations, while allowing for arbitrary joining and splitting of any number of non-contiguous segments. The annotation levels are represented as a graph consisting of a set of parallel paths (annotation levels) with links between them. Nodes along the paths always stand for word tokens, correct or incorrect, and in a sentence with nothing to correct the corresponding word tokens in every pair of neighbouring paths are linked 1:1. Additionally, the nodes can be assigned morphosyntactic tags, syntactic functions or any other word-specific information. Whenever a word form is emended, the type of error can be specified as a label of the link connecting the incorrect form at level S_i with its emended form at level S_{i+1}. In general, these labelled relations can link an arbitrary number of elements at one level with an arbitrary number of elements at a neighbouring level. The elements at one level participating in this relation need not form a contiguous sequence. Multiple words at any level are thus identified as a single segment, which is related to a segment at a neighbouring level, while any of the participating word forms can retain their 1:1 links with their counterparts at other levels. This is useful for splitting and joining word forms, for changing word order, and for any other corrections involving multiple words. Nodes can also be added or omitted at any level to correct missing or odd punctuation signs or syntactic constituents. See Figure 1 below for an example of this multi-level annotation scheme.

The option of relating multiple nodes as single segments across levels could also be used for treating morphosyntactic errors in concord and government. However, in this case there is typically one correct form involved, e.g., the subject in subject-predicate agreement, the noun in adjective-noun agreement, the verb assigning case to a complement, the antecedent in pronominal reference. Rather than treating both the correct and the incorrect form as equals in a 2:2 relation between the levels, the incorrect form is emended using a 1:1 link with an option to refer to the correct form. Such references link pairs of forms at neighbouring levels rather than the forms themselves to enable possible references from a multi-word unit (or) to another multi-word unit. See Figure 1 below again, where such references are represented by arrows originating in labels **val**.

A single error may result in multiple incorrect forms as shown in (1). The adjective *velký* 'big-NOM-SG-M(ASC)' correctly agrees with the noun *pes* 'dog-NOM-SG-MASC'. However, the case of the noun is incorrect – it should be in accusative rather than nominative. When the noun's case is corrected, the case of the adjective has to be corrected as well. Then multiple references are made: to the verb as the case assigner for the noun, and to the noun as the source of agreement for the adjective.

(1) a. *Viděl velký pes.
saw big-NOM-SG-M dog-NOM-SG-M
b. Viděl velkého psa.
saw big-ACC-SG-M dog-ACC-SG-M
'He saw a big dog'

Annotation of learners' texts is often far from straightforward, and alternative interpretations are available even in a broader context. The annotation format supports alternatives, but for the time being the annotation tool does not support local disjunctions. This may be a problem if the annotator has multiple target hypotheses in mind.

4.3 Three levels of annotation

A multi-level annotation scheme calls for some justification, and once such a scheme is adopted, the question of the number of levels follows.

After a careful examination of alternatives, we have arrived at a two-stage annotation design, based on three levels. A flat, single-stage, two-level annotation scheme would be appropriate if we were interested only in the original text and in the annotation at some specific level (fully emended sentences, or some intermediate stage, such as emended word forms). The flat design could be used even if we insisted on registering some intermediate stages of the passage from the original to a fully emended text, and decided to store such information with the word-form nodes. However, such information might get lost in the case of significant changes involving deletions or additions (e.g., in Czech as a pro-drop language, the annotator may decide that a misspelled personal pronoun in the subject position should be deleted and the information about the spelling error would lost). The decision to use a multi-level design was mainly due to our interest in annotating errors in single forms as well as those spanning (potentially discontinuous) strings of words.

Once we have a scheme of multiple levels available, we can provide the levels with theoretical significance and assign a linguistic interpretation to each of them. In a world of unlimited resources of annotators' time and experience, this would be the optimal solution. The first annotation level would be concerned only with errors in graphemics, followed by levels dedicated to morphemics, morphosyntax, syntax, lexical phenomena, semantics and pragmatics. More realistically, there could be a level for errors in graphemics and morphemics, another for errors in morphosyntax (agreement, government) and one more for everything else, including word order and phraseology.

Our solution is a compromise between corpus users' expected demands and limitations due to the annotators' time and experience. The annotator has a choice of two levels of annotation, and the distinction, based to a large extent on formal criteria, is still linguistically relevant.

At the level of transcribed input (Level 0), the nodes represent the original strings of graphemes. At the level of orthographical and morphological emendation (Level 1), only individual forms are treated. The result is a string consisting of correct Czech forms, even though the sentence may not be correct as a whole. The rule of "correct forms only" has a few exceptions: a faulty form is retained if no correct form could be used in the context or if the annotator cannot decipher the author's intention. On the other hand, a correct form may be replaced by another correct form if the author clearly misspelled the latter, creating an unintended homograph with another form. All other types of errors are emended at Level 2.

4.4 Captured errors

A typical learner of Czech makes errors all along the hierarchy of theoretically motivated linguistic levels, starting from the level of graphemics up to the level of pragmatics. Our goal is to 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. Emendation is possible only when the input is comprehensible. In cases where the input or its part is not comprehensible, it is left with a partial or even no annotation.

The taxonomy of errors is rather coarse-grained, a more detailed classification is previewed for a later stage and a smaller corpus sample. It follows the three-level distinction and is based on criteria as straightforward as possible. Whenever the error type can be determined from the way the error is emended, the type is supplied automatically by a post-processing module, together with morphosyntactic tags and lemmas for the correct or emended forms (see § 5.3).

Errors in individual word forms, treated at Level 1, include misspellings (also diacritics and capitalisation), misplaced word boundaries, missing or misused punctuation, but also errors in inflectional and derivational morphology and unknown stems. These types of errors are emended manually, but the annotator is not expected label them by their type – the type of most errors at Level 1 is identified automatically. The only exception where the error type must be assigned manually is when an unknown stem or derivation affix is used.

Whenever the lexeme (its stem and/or suffix) is unknown and can be replaced by a suitable form, it is emended at Level 1. If possible, the form should fit the syntactic context. If no suitable form can be found, the form is retained and marked as unknown. When the form exists, but is not appro-

priate in context, it is emended at Level 2 – the reason may be the violation of a syntactic rule or semantic incompatibility of the lexeme.

Table 2 gives a list of error types emended at Level 1. Some types actually include subtypes: words can be incorrectly split or joined, punctuation, diacritics or character(s) can be missing, superfluous, misplaced or of a wrong kind. The Links column gives the maximum number of positions at Level 0, followed by the maximum number of position at Level 1 that are related by links for this type of error. The Id column says if the error type is determined automatically or has to be specified manually.

Error type	**Links**	**Id**
Word boundary	m:n	A
Punctuation	0:1, 1:0	A
Capitalisation	1:1	A
Diacritics	1:1	A
Character(s)	1:1	A
Inflection	1:1	A
Unknown lexeme	1:1	M

Table 2: Types of errors at Level 1

Emendations at Level 2 concern errors in agreement, valency and pronominal reference, negative concord, the choice of a lexical item or idiom, and in word order. For the agreement, valency and pronominal reference cases, there is typically an incorrect form, which reflects some properties (morphological categories, valency requirements) of a correct form (the agreement source, syntactic head, antecedent). Table 3 gives a list of error types emended at Level 2. The Ref column gives the number of pointers linking the incorrect form with the correct "source".

Error type	**Links**	**Ref**	**Id**
Agreement	1:1	1	M
Valency	1:1	1	M
Pronominal reference	1:1	1	M
Complex verb forms	m:n	0,1	M
Negation	m:n	0,1	M
Missing constituent	0:1	0	M
Odd constituent	1:0	0	M
Modality	1:1	0	M
Word order	m:n	0	M
Lexis & phraseology	m:n	0,1	M

Table 3: Types of errors at Level 2

The annotation scheme is illustrated in Figure 1, using an authentic sentence, split in two halves for space reasons. There are three parallel strings of word forms, including punctuation signs, representing the three levels, with links for corresponding forms. Any emendation is labelled with an error type.[4] The first line is Level 0, imported from the transcribed original, with English glosses below (forms marked by asterisks are incorrect in any context, but they may be comprehensible – as is the case with all such forms in this example). Correct words are linked directly with their copies at Level 1, for emended words the link is labelled with an error type. In the first half of the sentence, **unk** for unknown form, **dia** for an error in diacritics, **cap** for an error in capitalisation. According to the rules of Czech orthography, the negative particle *ne* is joined with the verb using an intermediate node **bnd**. A missing comma is introduced at Level 1, labelled as a **p**unctuation error. All the error labels above can be specified automatically in the post-processing step.

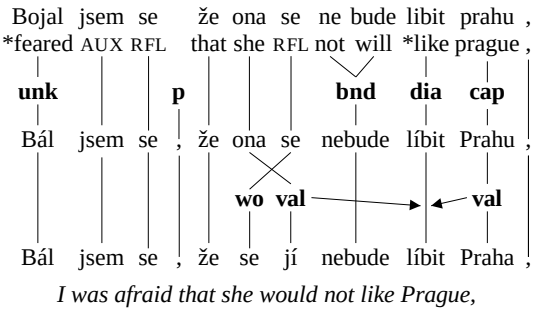
Staying with the first half of the sentence, most forms at Level 1 are linked directly with their equivalents at Level 2 without emendations. The reflexive particle *se* is misplaced as a second position clitic, and is put into the proper position using the link labelled **wo** for a word-order error.[5] The pronoun *ona* – 'she' in the nominative case – is governed by the form *líbit se*, and should bear the dative case: *jí*. The arrow to *líbit* makes the reason for this emendation explicit. The result could still be improved by positioning *Praha* after the clitics and before the finite verb *nebude*, resulting in a word order more in line with the underlying information structure of the sentence, but our policy is to refrain from more subtle phenomena and produce a grammatical rather than a perfect result.

In the second half of the sentence, there is only one Level 1 error in diacritics, but quite a few errors at Level 2. *Proto* 'therefore' is changed to *protože* 'because' – a **lex**ical emendation. The main issue are the two finite verbs *bylo* and *vadí*. The most likely intention of the author is best expressed by the conditional mood. The two non-contiguous forms are replaced by the conditional

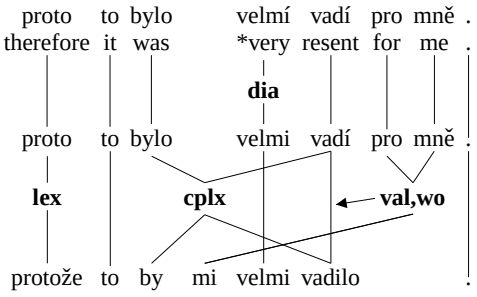
[4] The labels for error types used here are simplified for reasons of space and mnemonics.

[5] In word-order errors it may be difficult to identify a specific word form violating a rule. The annotation scheme allows for both *se* and *jí* to be blamed. However, here we prefer the simpler option and identify just one, more prominent word form. Similarly with *mi* below.

auxiliary and the content verb participle in one step using a 2:2 relation. The intermediate node is labelled by **cplx** for complex verb forms. The prepositional phrase *pro mně* 'for me' is another complex issue. Its proper form is *pro mě* (homonymous with *pro mně*, but with 'me' bearing accusative instead of dative), or *pro mne*. The accusative case is required by the preposition *pro*. However, the head verb requires that this complement bears bare dative – *mi*. Additionally, this form is a second position clitic, following the conditional auxiliary (also a clitic) in the clitic cluster. The change from PP to the bare dative pronoun and the reordering are both properly represented, including the pointer to the head verb. What is missing is an explicit annotation of the faulty case of the prepositional complement, which is lost during the Level 1 – Level 2 transition, the price for a simpler annotation scheme with fewer levels. It might be possible to amend the PP at Level 1, but it would go against the rule that only forms wrong in isolation are emended at Level 1.



I was afraid that she would not like Prague,



because I would be very unhappy about it.

Figure 1: Annotation of a sample sentence

4.5 Data Format

To encode the layered annotation described above, we have developed an annotation schema in the Prague Markup Language (PML).[6] PML is a generic XML-based data format, designed for the representation of rich linguistic annotation organised into levels. In our schema, each of the higher levels contains information about words on that level, about the corrected errors and about relations to the tokens on the lower levels. Level 0 does not contain any relations, only links to the neighbouring Level 1. In Figure 2, we show a portion (first three words and first two relations) of the Level 1 of the sample sentence encoded in our annotation schema.

[6] `http://ufal.mff.cuni.cz/jazz/pml/index_en.html`

```
<?xml version="1.0" encoding="UTF-8"?>
<adata xmlns="http://utkl.cuni.cz/czesl/">
  <head>
    <schema href="adata_schema.xml" />
    <references>
      <reffile id="w" name="wdata" href="r049.w.xml" />
    </references>
  </head>
  <doc id="a-r049-d1" lowerdoc.rf="w#w-r049-d1">
    ...
    <para id="a-r049-d1p2" lowerpara.rf="w#w-r049-d1p2">
       ...
      <s id="a-r049-d1p2s5">
        <w id="a-r049-d1p2w50">
          <token>Bál</token>
        </w>
        <w id="a-r049-d1p2w51">
          <token>jsem</token>
        </w>
        <w id="a-r049-d1p2w52">
          <token>se</token>
        </w>
         ...
      </s>
      ...
      <edge id="a-r049-d1p2e54">
        <from>w#w-r049-d1p2w46</from>
        <to>a-r049-d1p2w50</to>
        <error>
           <tag>unk</tag>
        </error>
      </edge>
      <edge id="a-r049-d1p2e55">
        <from>w#w-r049-d1p2w47</from>
        <to>a-r049-d1p2w51</to>
      </edge>
      ...
    </para>
    ...
  </doc>
</adata>
```

Figure 2: Portion of the Level 1 of the sample sentence encoded in the PML data format.

5 Annotation process

The whole annotation process proceeds as follows:

- A handwritten document is transcribed into html using off-the-shelf tools (e.g. Open Office Writer or Microsoft Word).
- The information in the html document is used to generate Level 0 and a default Level 1 encoded in the PML format.
- An annotator manually corrects the document and provides some information about errors using our annotation tool.
- Error information that can be inferred automatically is added.

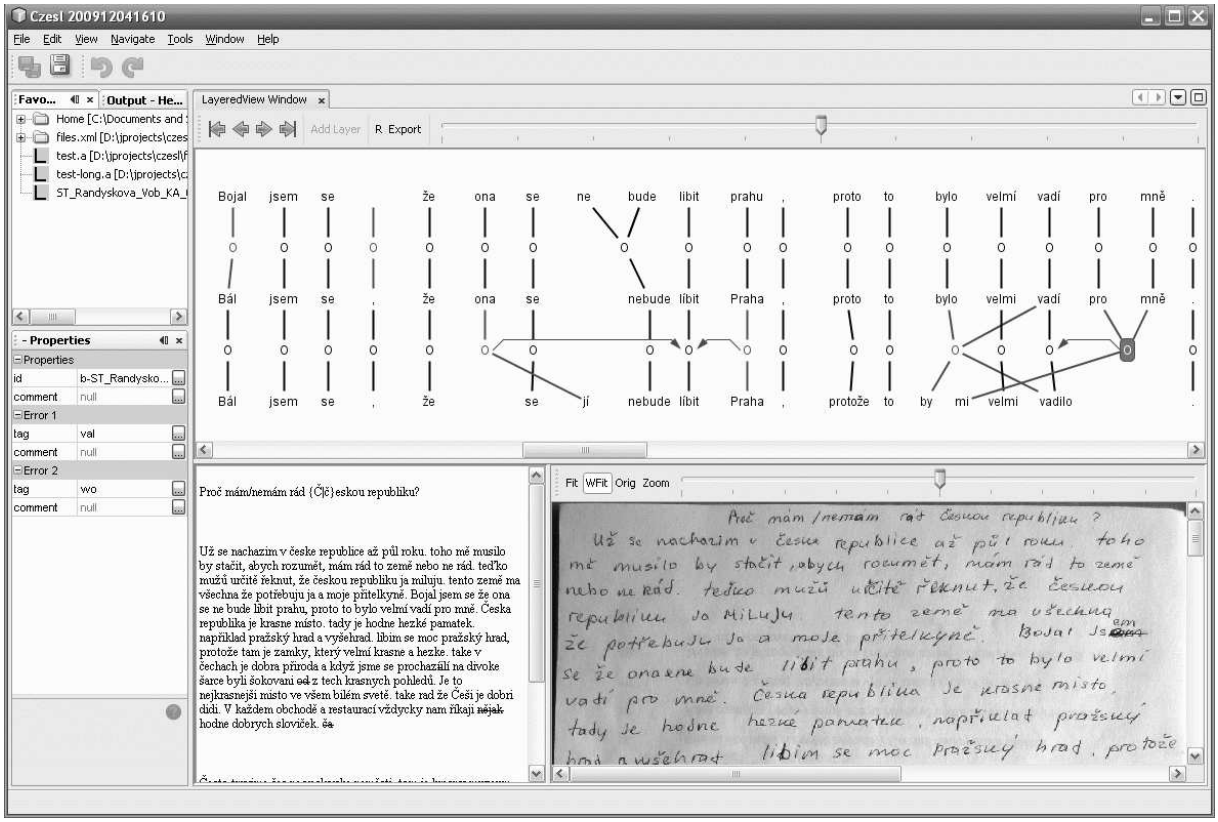


Figure 3: Sample sentence in the annotation tool.

5.1 Transcription

The original documents are hand-written, usually the only available option, given that their most common source are language courses and exams. The avoidance of an electronic format is also due to the concern about the use of automatic text-editing tools by the students, which may significantly distort the authentic interlanguage.

Therefore, the texts must be transcribed, which is very time consuming. While we strive to capture only the information present in the original hand-written text, often some interpretation is unavoidable. For example, the transcribers have to take into account specifics of hand-writing of particular groups of students and even of each individual student (the same glyph may be interpreted as *l* in the hand-writing of one student, *e* of another, and *a* of yet another). When a text allows multiple interpretation, the transcribers may provide all variants. For example, the case of initial letters or word boundaries are often unclear. Obviously, parts of some texts may be completely illegible and are marked as such.

Also captured are corrections made by the student (insertions, deletions, etc.), useful for investigating the process of language acquisition.

The transcripts are not spell-checked automatically. In a highly inflectional language, deviations in spelling very often do not only reflect wrong graphemics, but indicate an error in morphology.

5.2 Annotation

The manual portion of annotation is supported by an annotation tool we have developed. The annotator corrects the text on appropriate levels, modifies relations between elements (by default all relations are 1:1) and annotates relations with error tags as needed. The context of the annotated text is shown both as a transcribed html document and as a scan of the original document. The tool is written in Java on top of the Netbeans platform.[7] Figure 3 shows the annotation of the sample sentence as displayed by the tool.

5.3 Postprocessing

Manual annotation is followed by automatic post-processing, providing the corpus with additional information:

[7] `http://platform.netbeans.org/`

- Level 1: lemma, POS and morphological categories (this information can be ambiguous)
- Level 2: lemma, POS and morphological categories (disambiguated)
- Level 1: type of error (by comparing the original and corrected strings), with the exception of lexical errors that involve lemma changes (e.g. **kadeřnička – kadeřnice* 'hair-dresser')
- Level 2: type of morphosyntactic errors caused by agreement or valency error (by comparing morphosyntactic tags at Level 1 and 2)
- Formal error description: missing/extra expression, erroneous expression, wrong order
- In the future, we plan to automatically tag errors in verb prefixes, inflectional endings, spelling, palatalisation, metathesis, etc.

6 Conclusion

Error annotation is a very resource-intensive task, but the return on investment is potentially enormous. Depending on the annotation scheme, the corpus user has access to detailed error statistics, which is difficult to obtain otherwise. An error-tagged corpus is an invaluable tool to obtain a reliable picture of the learners' interlanguage and to adapt teaching methods and learning materials by identifying the most frequent error categories in accordance with the learner's proficiency level or L1 background.

We are expecting plentiful feedback from the error annotation process, which is just starting. As the goal of a sizable corpus requires a realistic setup, we plan to experiment with more and less detailed sets of error types, measuring the time and inter-annotator agreement. A substantially more elaborate classification of errors is previewed for a limited subset of the corpus.

At the same time, the feedback of the annotators will translate into the ongoing tuning of the annotation guidelines, represented by a comprehensive error-tagging manual. We hope in progress in dealing with thorny issues such as the uncertainty about the author's intended meaning, the inference errors, the proper amount of interference with the original, or the occurrence of colloquial language. In all of this, we need to make sure that annotators handle similar phenomena in the same way.

However, the real test of the corpus will come with its usage. We are optimistic – some of the future users are a crucial part of our team and their needs and ideas are the driving force of the project.

7 Acknowledgements

We wish to thank other members of the project team, namely Milena Hnátková, Tomáš Jelínek, Vladimír Petkevič, and Hana Skoumalová for their numerous stimulating ideas, acute insight and important feedback. We are especially grateful to Karel Šebesta, for all of the above and for initiating and guiding this enterprise.

The work described in this paper is funded by the European Social Fund and the government of the Czech Republic within the operational programme 'Education for Competitiveness' as a part of the project 'Innovation in Education in the Field of Czech as a Second Language' (project no. CZ.1.07/2.2.00/07.0259).

References

Sylviane Granger, editor. 1998. *Learner English on Computer*. Addison Wesley Longman, London and New York.

Geoffrey Leech. 1998. Preface. In Granger Sylviane, editor, *Learner English on Computer*, pages xiv–xx. Addison Wesley Longman, London and New York.

Anke Lüdeling, Maik Walter, Emil Kroymann, and Peter Adolphs. 2005. Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics 2005*, Birmingham.

Nadja Nesselhauf. 2004. Learner corpora and their potential for language teaching. In John McHardy Sinclair, editor, *How to use corpora in language teaching*, Studies in corpus linguistics, pages 125–152. Benjamins, Amsterdam/Philadelphia.

Norma A. Pravec. 2002. Survery of learner corpora. *ICAME Journal*, 26:81–114.

Thomas Schmidt. 2009. Creating and working with spoken language corpora in EXMARaLDA. In *LULCL II: Lesser Used Languages & Computer Linguistics II*, pages 151–164.

Larry Selinker. 1983. Interlanguage. In Betty W. Robinett and Jacquelyn Schachter, editors, *Second Language Learning: Contrastive analysis, error analysis, and related aspects*, pages 173–196. The University of Michigan Press, Ann Arbor, MI.

Richard Xiao. 2008. Well-known and influential corpora. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, volume 1 of *Handbooks of Linguistics and Communication Science [HSK] 29.1*, pages 383–457. Mouton de Gruyter, Berlin and New York.