# Annotating foreign learners’ Czech

Alexandr Rosen[1] Svatava Škodová[2] Barbora Štindlová[2] Jirka Hana[1]

Texts produced by non-native speakers can be compiled in a corpus, usually called a *learner corpus*. Such a corpus is a precious resource mainly for authors of textbooks and researchers in 2nd language acquisition. In addition to morphosyntactic tags and lemmas, learner corpora can be annotated with information relevant to the cases of deviant use. Such cases can be identified, emended and assigned a tag specifying the type the error. 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. The annotation scheme proposed here is an attempt to strike a compromise between the limitations of the annotation process and the demands of the corpus user.

In the presentation and the full paper, we will give a brief introduction to the project of a learner corpus of Czech of the planned size of 2 million words, and present the concept of our annotation scheme, followed by a description of the annotation process.

Following the example of some previous multi-level annotation schemes (Lüdeling et al.(2005)), we propose a three-level format that supports successive emendations, including multiple forms in discontinuous sequences. In many cases, the error type follows from the comparison of 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 or valency patterns explicit.

Levels of annotation are represented as a graph, consisting of a set of interlinked parallel paths – see Figs. 1 and 2. Nodes along the paths stand for word tokens, correct or incorrect. In a sentence with nothing to correct the corresponding words in every pair of neighbouring paths are linked 1:1. Additionally, the nodes can be assigned morphosyntactic tags 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 $i$ with its emended form at level $i + 1$.

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.

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. The taxonomy of errors is rather coarse-grained (see Tables 1 and 2). It follows the three-level distinction and is based on criteria as straightforward as possible. A more detailed classification is previewed for a smaller corpus sample.

Table 1 gives a list of error types with the corresponding tags, emended at Level 1. The Links column gives the maximum number of forms at Level 0, followed by the maximum number of forms 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. 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. The cases of agreement, valency and pronominal reference are assumed to involve a source form (agreement source, syntactic head, antecedent), determining one or more target forms (agreement target, syntactic dependent, pronoun). Errors of these types occur when the target forms fail to reflect some properties (morphological categories, valency requirements) of the source form. Table 2 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”.

[1] Charles University, Prague

[2] Technical University, Liberec

| **Error type** | **Tag** | **Links** | **Id** |
|---|---|---|---|
| Word boundary | **bnd** | m:n | A |
| Punctuation | **p** | 0:1, 1:0 | A |
| Capitalisation | **cap** | 1:1 | A |
| Diacritics | **dia** | 1:1 | A |
| Character(s) | **char** | 1:1 | A |
| Inflection | **infl** | 1:1 | A |
| Unknown lexeme | **unk** | 1:1 | M |

Table 1: Types of errors at Level 1

| **Error type** | **Tag** | **Links** | **Ref** | **Id** |
|---|---|---|---|---|
| Agreement | **agr** | 1:1 | 1 | M |
| Valency | **val** | 1:1 | 1 | M |
| Pronominal reference | **ref** | 1:1 | 1 | M |
| Complex verb forms | **cvf** | m:n | 0,1 | M |
| Negation | **neg** | m:n | 0,1 | M |
| Missing constituent | **miss** | 0:1 | 0 | M |
| Odd constituent | **odd** | 1:0 | 0 | M |
| Modality | **mod** | 1:1 | 0 | M |
| Word order | **wo** | m:n | 0 | M |
| Lexis & phraseology | **lex** | m:n | 0,1 | M |

Table 2: Types of errors at Level 2

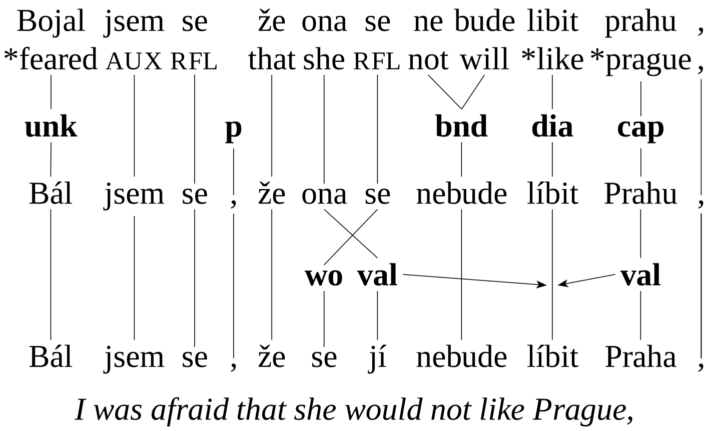
Bojal jsem se že ona se ne bude libit prahu ,
*feared AUX RFL that she RFL not will *like *prague ,
unk
p
bnd
dia
cap
Bál jsem se , že ona se nebude líbit Prahu ,
wo
val
val
Bál jsem se , že se jí nebude líbit Praha ,
I was afraid that she would not like Prague,

Figure 1: Annotation of a sample sentence, part I

**Comments on Fig. 1** The first line is Level 0, imported from the transcribed original (asterisked forms are incorrect in any context). Correct words are linked directly with their copies at Level 1. For emended words the link is labelled with an error type. Here, all error labels for emendations at Level 1 can be assigned automatically in post-processing. Most forms at Level 1 are linked directly with their equivalents at Level 2 without emendations. The reflexive particle *se* is misplaced as a 2P clitic, and reordered using the link labelled **wo**. The pronoun *ona* – 'she' in the nominative case – is governed by the form *líbit se*, and should bear the dative: *jí*. The arrow to *líbit* makes the reason for this emendation explicit.[2]

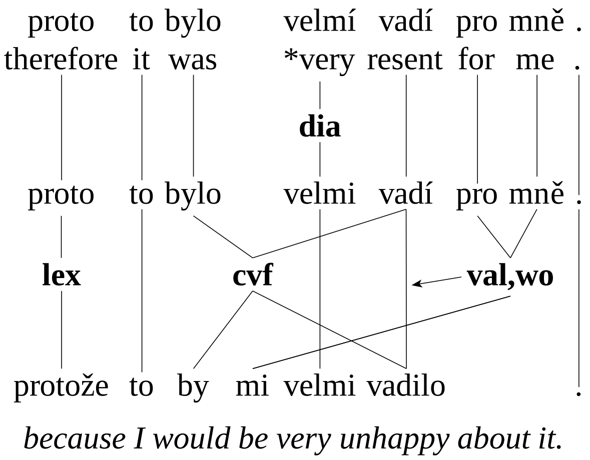
proto to bylo velmí vadí pro mně .
therefore it was *very resent for me .
dia
proto to bylo velmi vadí pro mně .
lex
cvf
val,wo
protože to by mi velmi vadilo .
because I would be very unhappy about it.

Figure 2: Annotation of a sample sentence, part II

**Comments on Fig. 2** In the rest of the sentence, *proto* 'therefore' is changed to *protože* 'because' – a **lex**ical emendation. The main issue are the two finite verbs *bylo* and *vadí*. The author's most likely intention is best expressed by the conditional. Therefore, the two non-contiguous forms are replaced by the conditional auxiliary and the content verb participle in one step using a 2:2 relation. The intermediate node is labelled with **cvf** for complex verb forms. The prepositional phrase *pro mně* 'for me' is another complex issue. Its proper form is *pro mě* (homophonous 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 2P clitic, following the conditional auxiliary. The change from PP to the bare dative pronoun and the reordering are both properly represented, including the pointer to the head verb.[3]

## References

Lüdeling, A., Walter, M., Kroymann, E., & Adolphs, P. (2005). Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics 2005*, Birmingham.

[2] The result could still be improved by ordering *Praha* after the clitics and before *nebude*, resulting in a word order more in line with the information structure of the sentence, but our policy is to prefer less intervention, and to produce a grammatical rather than a perfect result.

[3] What is missing is an explicit annotation of the faulty case of the prepositional complement (*mně* or *mne* instead of *mě*), which is lost during the Level 1 – Level 2 transition, the price for a simpler annotation scheme with fewer levels.