A proposal for a new error annotation system of Czech learner corpora

The analysis of corpora consisting of texts of non-native speakers has become an important tool for understanding of the process of learning a second language and the development of adequate teaching methodologies. In this paper, we propose a new concept of error annotation of texts produced by learners of Czech as a second language which is both simpler and more accurate than previous error annotation systems such as (Wisniewski et al. 2014, Rosen 2016, Jelínek et al. 2012). We also describe the procedure of re-annotation of existing learner corpora by the proposed annotation system.

The largest corpus of Czech learner corpora to date is the corpus CzeSL (Štindlová et al. 2013). In this corpus, the error annotation uses two levels of emendation. At the lower level, erroneous word forms are corrected; the result of the higher level of annotation is a correct sentence. To each word correction on both levels, an error label (of about twenty types) is then manually assigned.

We propose an annotation system that will only use the final, correct emendation (not two levels like CzeSL), significantly simplifying annotator work and facilitating the reproducibility of the error annotation using NLP tools. Our error annotation is based on the levels of linguistic description: we identify orthographic errors (ORT), phonological and morphological errors (MPHON), errors of inflection (MORPH), syntactical errors (SYN) and lexical errors and errors of use (LEX); with optional more detailed sub-labels (e.g. SYN:dep – syntactic error of dependency, ORT:cap – orthographic error of capitalization). In cases where there are two or more possible causes of the error, several error tags may be assigned, with one chosen as the most likely (most relevant). For example, the phrase *přijdou mnoho lidi* "many people will come" with the wrong form of *lidi* (Nom.pl) instead of *lidí* (Gen.pl) may be an orthographic error (omission of diacritics), morphological error (erroneous case form) or syntactic error (incorrect case choice); the primary error tag is MORF, with ORT and SYN as alternative tags.

The error annotation can be more accurate due to the fact that the precise location of errors inside the words are marked. For example, the word *kamar**atky*** "friends" in the phrase *Mám mnoho kamaratky* "I have many friends" instead of *kamarádek* has three separate errors:

a/á MPHON + ORT:dia (missing diacritics marking vowel length);

t/d MPHON + ORT:assim (wrong consonnant due to assimilation of voiced/voiceless consonnants);

ky/ek MORPH + SYN:dep (wrong choice of suffix, nominative instead of genitive plural);

each will be marked and error-annotated separately.

In order to get data for machine learning and automatic annotation, we use already annotated CzeSL data, namely the original text (transcribed) and the corrected text (final emendation). In the future, we will use also automatically corrected texts using a combination of rule-based corrections and a stochastic spell-checker and text correction tool (Richter et al., 2012).

The actual annotation of student texts combines automatic text pre-processing, manual annotation in the Brat environment (Stenetorp et al., 2012; see Fig. 1[1] and Fig. 2) and automatic post-processing of annotated text.

Preprocessing identifies individual differences between original and corrected text, it marks these differences and whenever possible, automatically assigns error types (on lower levels of language description, i.e. most of the ORT and MPHON errors). This automatic error tagging greatly reduces the task of the manual annotator, and as it is rule-based, it is relatively accurate: 3% of errors (on a 100 error sample), almost any type of automatic annotation error can be corrected. In addition to the error-annotation, a simple morphematic analysis is performed.

The manual annotator verifies the automatically labeled errors, assigns each identified error an error-label and checks for others, unidentified errors. The corrected text cannot be changed in Brat, but can be marked as not properly corrected (to be corrected outside of Brat). Automatic postprocessing assigns, morphological tags and lemmas to both original and corrected word forms, for some types of annotator-labeled error tags, sub-labels or flags are added. As a separate information, it records which characters on the part of the original and corrected word form are part of the identified error (eg. in *Prahě/Praze* : *hě/ze*).

We intend to build a corpus of texts produced by learners of Czech and annotated by the proposed error-annotation systém. It would increase our understanding of interlanguage and lead to better teaching methods of Czech as a second language.

1 Fig. 1. shows a part of a text in the Brat annotation environment, pre-processed and ready for manual annotation. Fig. 2 shows data used by the Brat environment: the text itself and standoff annotation. Asterisk marks morphematic analysis (prefix*stem*suffix).

References

Jelínek, T., Štindlová, B., Rosen, A. and Hana, J. (2012). Combining manual and automatic annotation of a learner corpus. In P. Sojka et al. (eds), Text, Speech and Dialogue – Proceedings of TSD 2012, no. 7499 in LNCS, p. 127–134.

Štindlová, B., Škodová, S., Hana, J., & Rosen, A. (2013). A learner corpus of Czech: current state and future directions. In S. Granger et al. (eds), Twenty Years of Learner Corpus Research: Looking back, Moving ahead, Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve. Presses Universitaires de Louvain.

Richter M., Straňák P., Rosen A. (2012). Korektor – A System for Contextual Spell-checking and Diacritics Completion. In Kay M., Boitet C.: Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012). Mumbai, India, p. 1-12.

Rosen, A. (2016). Building and using corpora of non-native Czech. In B. Brejová (ed), Proceedings of the 16th ITAT: Slovenskočeský NLP workshop (SloNLP 2016), vol. 1649 of CEUR Workshop Proceedings, Bratislava, Slovakia, p. 80–87.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou S. and Tsujii, J. (2012). Brat: a Web-based Tool for NLP-Assisted Text Annotation. In Proceedings of the Demonstrations Session at EACL 2012, 102–107.

Wisniewski, K., Woldt, C., Schöne, K., Abel, A., Blaschitz, V., Štindlová, B., Vodičková, K. (2014). The MERLIN annotation scheme for the annotation of German, Italian, and Czech learner language. Available online www.merlin-platform.eu.

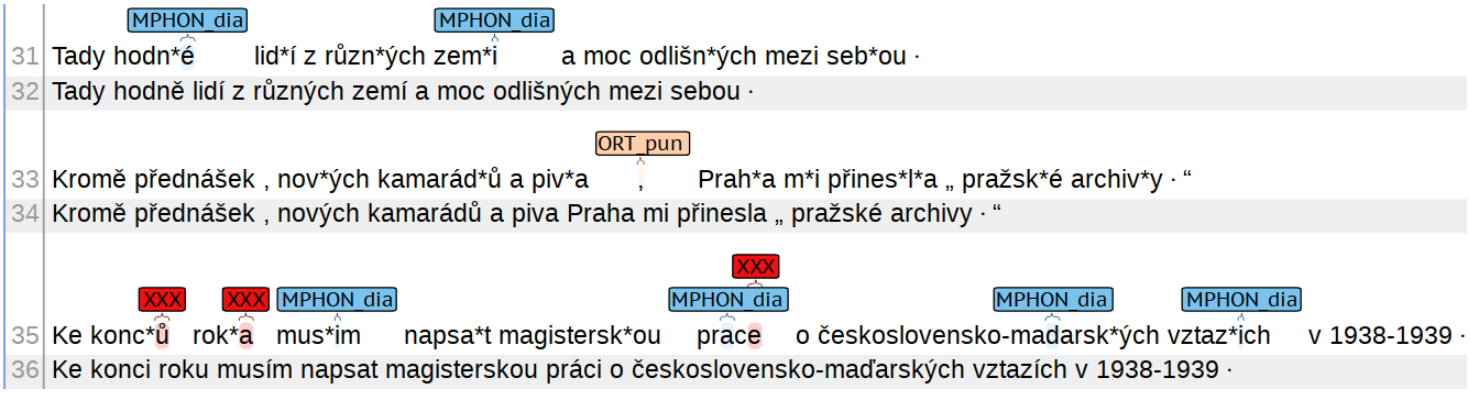

Fig. 1. Error annotation in brat: input for manual annotation.

```
Tady hodn*é lid*í z různ*ých zem*i a moc odlišn*ých mezi seb*ou .
Tady hodně lidí z různých zemí a moc odlišných mezi sebou .
Kromě přednášek , nov*ých kamarád*ů a piv*a , Prah*a m*i přines*l*a „ pražsk*é archiv*y . “
Kromě přednášek , nových kamarádů a piva Praha mi přinesla „ pražské archivy . “
Ke konc*ů rok*a mus*im napsa*t magistersk*ou prace o československo-madarsk*ých vztaz*ich v 1938-1939 .
Ke konci roku musím napsat magisterskou práci o československo-maďarských vztazích v 1938-1939 .

T43	MPHON_dia 1806 1807	é
T44	MPHON_dia 1829 1830	i
T45	ORT_pun 1966 1967	,
T46	XXX 2103 2104	ů
T47	XXX 2109 2110	a
T48	MPHON_dia 2115 2116	i
T49	MPHON_dia 2142 2143	a
T50	XXX 2144 2145	e
T51	MPHON_dia 2165 2166	d
T52	MPHON_dia 2181 2182	i
```

Fig. 2. Text data and standoff annotation used in the brat environment.