Different ways of tagging errors in learner corpora

The most distinctive feature of learner corpora is the fact that texts written by language learners contain errors, or deviant forms, or non-native variants of the target language if you wish. To provide a systematic analysis of those errors, learner corpora typically include error annotation, indicating errors in the text. This is traditionally done by an error code assigned to the incorrect part of the text, optionally accompanied by a target hypothesis, i.e. a reformulation of the error in the native standard of the target language. While standards are emerging for linguistic annotation of corpora including standard native language, choosing appropriate categories for annotating errors is not easy. The codes usually reflect their interpretation in terms of a standard grammar (spelling, morphological paradigms, morphosyntactic categories, agreement, government or valency, etc.), and thus their design and application to various phenomena of a non-native language is far from trivial. Moreover, the interplay of categories presumably responsible for the phenomena is not easily represented by tags assigned to the linear text of the original.

We compare the traditional linear approach to error annotation with two alternative approaches: (i) the approach introduced in the COPLE2 corpus (del Río et al. 2016), in which errors are not indicated by a code, but rather the error is provided with an orthographic, a morphosyntactic, and a lexical correction, which together provide detailed information about the type of error; and (ii) the approach introduced in the CzeSL corpus, in which the erroneous sentence is aligned with two corrected versions of of the sentence, with alignments between the words in the three variants, as well as error codes (Hana et al. 2014).

For the comparison, we will describe what the various options provide, using the two tools in their respective project as examples: *feat*[1] for the parallel scheme approach, and TEITOK (Janssen 2016) for the multi-layered correction approach. We look at how the different approaches can represent complex cases of overlapping errors (where a group of words is involved in various distinct errors), discontinuous errors (where the words involved in an error are not next to each other), word order errors, and secondary errors (where the correction of one error leads to another error). We will not only discuss how the three different paradigms can represent the complex error cases, but also how once the errors are correctly represented in the corpus, they can be used for concrete search queries to answer research questions.

Error annotation of learner corpora is often combined with linguistic annotation, as it is applied in corpora of standard (native) language. If a corrected version of the text is available, standard tools (such as tokenizers, taggers, and parsers) can be used to apply such tools with high accuracy, and the result can even be projected to or linked with the original uncorrected text. Annotating the uncorrected original text is typically a more challenging task, both in terms of accuracy of existing tools, and in terms of missing concepts. We will show how the paradigms discussed above integrate with annotations of this type.

[1] https://bitbucket.org/jhana/feat-hg/wiki/Home

References

del Río, I., Antunes, S., Mendes, A., and Janssen, M. (2016). Towards error annotation in a learner corpus of portuguese. In *5th NLP4CALL and 1st NLP4LA workshop in Sixth Swedish Language Technology Conference (SLTC)*, Umeå University, Sweden.

Hana, J., Rosen, A., Štindlová, B., and Štěpánek, J. (2014). Building a learner corpus. *Language Resources and Evaluation*, 48(4):741–752.

Janssen, M. (2016). TEITOK: Text Faithful Corpora. In: *Proceedings of LREC 2016.* ELRA. Portorož, Slovenia, pp. 4037–4043.