Syntactic Misuse, Overuse and Underuse: A Study of a Parsed Learner Corpus and its Target Hypothesis

Anke Lüdeling, Amir Zeldes, Marc Reznicek,
Ines Rehbein, Hagen Hirschmann

Humboldt-Universität zu Berlin

This talk is concerned with using syntactic annotation of learner language and the corresponding target hypothesis to find structural acquisition difficulties in German as a foreign language. Using learner data for the study of acquisition patterns is based on the idea that learners do not produce random output but rather possess a consistent internal grammar (interlanguage; cf. [1] and many others). Analysing learner data is thus an indirect way of assessing the interlanguage of language learners. There are two main ways of looking at learner data, error analysis and contrastive interlanguage analysis [2, 3]. A careful analysis of errors makes it possible to understand learners' hypotheses about a given grammatical phenomenon. Contrastive interlanguage analysis is not concentrated on errors but compares categories (of any kind) of learner language with the same categories in native speaker language. Learners' underuse of a category (i.e. a significantly lower frequency in learner language than in native speaker language) can be seen as evidence for the perceived difficulty of that category (either because learners fail to acquire it, or because they deliberately avoid it).

While some learner corpora are annotated (manually or automatically) with part-of-speech or lemma information [4], or even error types, there are as yet only very few attempts to annotate them syntactically (some exceptions are [5] or [6]. Parsing learner data is very difficult because of the learner errors but would be very helpful for the analysis of errors and overuse/underuse of syntactic structures and categories. In our paper we therefore discuss how the comparison of parsed learner data and the corresponding target hypotheses helps in understanding syntactic properties of learner language.

We use the Falko corpus which contains essays of advanced learners of German as a foreign language and control essays by German native speakers [7]; the corpus is freely available[1]. Since it is very difficult to decide what an error is and often there can be different hypotheses about the 'correct' structure the learner utterance

[1] http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung-en/falko/standardseite-en

is evaluated against [8] both subcorpora are annotated manually with several layers of target hypotheses, as well as automatically with part-of-speech, lemma, and edit error tags [9].

The original learner data and the target hypotheses were parsed with a state-of-the-art statistical parser trained on the TiGer treebank [10]. Since the target hypotheses are aligned with the original data we can identify those sections in the data where parsing of the original fails but parsing of the target hypothesis is possible. We can then see which syntactic structures are assigned to the target hypothesis and use this as a diagnostic for syntactic learner errors. We can also analyse the syntactic categories in the learner data quantitatively against the native speaker data.

References

[1] Larry Selinker. Interlanguage. *IRAL*, 10/3:31–54, 1972.

[2] Sylviane Granger. From CA to CIA and back. An integrated approach to computerized bilingual and learner corpora. In Karin Aijmer, editor, *Papers from a Symposium on Text-based Cross-linguistic Studies Lund 4 - 5 March 1994*, page 37–51. Lund University Press,, 1996.

[3] Sylviane Granger. Learner corpora. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, pages 259–275. Mouton de Gruyter, Berlin, 2008.

[4] Ana Díaz-Negrillo, Detmar Meurers, Salvador Valera, and Holger Wunsch. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36(1–2), 2010.

[5] Markus Dickinson and Marwa Ragheb. Dependency Annotation for Learner Corpora. In *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8)*, 2009.

[6] Niels Ott and Ramon Ziai. Evaluating Dependency Parsing Performance on German Learner Language. In *Proceedings of the Ninth Workshop on Treebanks and Linguistic Theories (TLT-9)*, Tartu, 2010.

[7] Anke Lüdeling, Seanna Doolittle, Hagen Hirschmann, Karin Schmidt, and Maik Walter. Das Lernerkorpus Falko. *Deutsch als Fremdsprache*, 2:67–73, 2008.

[8] Anke Lüdeling. Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In Maik Walter and Patrick Grommes, editors, *Fortgeschrittene Lernervarietäten*, pages 119–140. Niemeyer, Tübingen, 2008.

[9] Marc Reznicek, Maik Walter, Karin Schmid, Anke Lüdeling, Hagen Hirschmann, and Cedric Krummes. *Das Falko-Handbuch. Korpusaufbau und Annotationen Version 1.0*, 2010.

[10] Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. TIGER: Linguistic Interpretation of a German Corpus. *Research on Language & Computation*, 2:597–620, 2004.