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
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


