Analyzing learner corpora: Which categories for interlanguage?

Detmar Meurers (Tübingen)

Learner corpora as collections of language produced by second language learners have been systematically collected since the 90s, and with readily available collections such as the ICLE for English and FALKO for German there is a growing empirical basis on which theories of second language acquisition and the interlanguage systems can be informed. Yet, as soon as the research questions go beyond the acquisition of vocabulary and constructions with unambiguous surface indicators, corpora must be enhanced with linguistic annotation to support efficient retrieval of the data that is relevant for such research questions.

In contrast to the different types of linguistic annotation schemes which have been developed for native language corpora, the discussion on which linguistic analysis and annotation is meaningful and appropriate for learner language is only starting. When formulating linguistic generalizations, one generally relies on a long tradition of linguistic analysis that has established an inventory of categories and properties to abstract away from the surface strings. In this talk, we will see that traditional linguistic categories are not necessarily an appropriate index into the space of interlanguage realizations and their systematicity, which research into second language acquisition aims to capture. Complementing the language explicitly given in the corpus, we also consider the need for information about the task which resulted in the corpus and the learners who produced it for interpreting and annotating learner data.