Ana Díaz-Negrillo, Detmar Meurers and Holger Wunsch (Tübingen University and University of Jaen)
adinegri@ujaen.es, dm@sfs.uni-tuebingen.de, wunsch@sfs.uni-tuebingen.de
Linguistic annotation of learner corpora

Generally speaking, learner data is the empirical basis of Second Language Acquisition (SLA) research, and it exemplifies typical stages and common learner problems in Foreign Language Teaching (FLT). Such data collected in learner corpora can help validate generalizations about language acquisition and support the development of new hypotheses and theories in SLA. Learner corpora can also play a role in identifying areas of relevance for FLT practice and materials design.

To find relevant classes of examples, the terminology used to single out learner language aspects of interest needs to be mapped to instances in the corpus. Effective querying of corpora for specific phenomena often requires reference to annotations (cf., e.g., Meurers, 2005; Meurers & Müller, 2009). Annotations essentially function as an index to classes of data which cannot easily be identified based on the surface form. For example, finding all sentences containing modal verbs using only the surface forms is possible, but would require a long list of all forms of the modal verbs. Even so, sentences where, for example, *can* is not actually a modal verb (e.g., *Pass me a can of beer* or *I can tuna for a living*) would be wrongly identified.

The annotation of corpora thus serves an important function, but also raises the question what type of learner language annotations are needed to support the searches for the data which are important for FLT and SLA research. A traditional focus of research on learner corpora has been the identification and classification of learner errors. As pointed out by Granger (2003), learner corpora can help overcome some of the key problems of the Error Analysis strand of SLA research in the 70s and 80s (cf., e.g., Richards, 1974; Corder, 1981). And indeed accuracy remains an important issue of interest to FLT (cf., e.g., the recent series of remedial books "Common Mistakes at . . . " by Cambridge University Press) and SLA (cf., e.g., Skehan, 1998). At the same time, prominent strands of SLA research are concerned with the stages of the acquisition process (cf., e.g., Pienemann, 1998), often independent of the accuracy of the execution of the patterns which are indicative of the different levels. In sum, SLA research essentially observes correlations of linguistic properties, whether erroneous or not. In consequence, learner corpora should ideally provide annotation of linguistic properties, including, but not limited, to errors.

In this paper, we explore what may constitute appropriate linguistic annotation schemes for learner language. We argue for an approach to Part-Of-Speech (POS) tagging of learner corpora that systematically encodes the distributional, morphological, and lexical aspects special to such interlanguage. Based on NOCE, an English learner corpus by Spanish learners, we characterize areas where the properties of learner language systematically differ from those assumed by POS annotation schemes developed for native language with a view to disclose suitable paths for learner language-specific POS tagging.

Keywords: learner corpora, POS annotation, SLA, ELT

References

Corder, S. P. 1981. *Error Analysis and Interlanguage*. Oxford: Oxford University Press.

Granger, S. 2003. Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal* 20/3: 465-480. URL http://purl.org/calico/granger03.pdf

Meurers, W. D. 2005. On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German. *Lingua* 115/11: 1619-1639. URL http://purl.org/dm/papers/meurers-03.html.

Meurers, W. D. & S. Müller. 2009. Corpora and Syntax. In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics*, vol. 2. Berlin: Mouton de Gruyter; 920-933. URL http://purl.org/dm/papers/meurers-mueller-09.html

Pienemann, M. 1998. *Language Processing and Second Language Development: Processability Theory*. Amsterdam: John Benjamins.

Richards, J. C. 1974. *Error Analysis: Perspectives on Second Language Acquisition*. London: Longman.

Skehan, P. 1998. A *Cognitive Approach to Learning Language*. Oxford: Oxford University Press.