# On the Creation and Analysis of a Reading Comprehension Exercise Corpus: Evaluating Meaning in Context

Detmar Meurers, Niels Ott, Ramon Ziai
Universität Tübingen

The notion of a task and language use in context play an important role in foreign language teaching and learning. In the same vein, a representation of the learner's ability to use language in context and perform tasks using appropriate strategies has been argued to be crucial for interpreting learner language and for informing learner modeling in Intelligent Tutoring Systems (Amaral & Meurers, 2008). Most learner corpora, however, consist of learner essays with minimal meaning and form requirements. Borrowing the terminology of Bachman & Palmer (1996), the essays are *indirect responses*, primarily encoding individual background knowledge of the learners. Fitzpatrick & Seegmiller (2004) show that for such learner essays it can be tremendously difficult to agree on what the learner was trying to say, i.e., they fail to reach sufficiently high inter-annotator agreement levels for annotating target hypotheses. In light of these issues, for our research on the automated comparison of meaning we explore the creation and annotation of a corpus in which the learner language produced is more explicitly contextualized and directly related to the input provided by the exercise.

We present our efforts at collecting a longitudinal learner corpus consisting of answers to reading comprehension questions written by American college students learning German. We discuss the development of the open-source WELCOME tool, which we created to facilitate the interdisciplinary exchange of the contextualized learner corpus between the language programs at OSU and KU providing the data and the computational linguists working on its encoding and automatic analysis. Along with the learner answers, we collect the reading texts, the reading comprehension questions, and the target answers that teachers prepare as reference for the grading process.

Only reading comprehension questions asking for information that is encoded in the text are included, thereby limiting the implicit need for world knowledge to evaluate the meaning of the learner answers. The meaning of each learner answer is assessed by two independent annotators. Meaning assessment is done using a binary classification (correct vs. incorrect) as well as using a richer set of diagnosis categories encoding the nature of the divergence from the target answers specified by the teachers. Following Bailey & Meurers (2008), we distinguish "missing concept", "extra concept", "blend" (missing concept and extra material), and "non-answer" for answers which are unrelated to the topic under discussion.

Based on the first year of the four year project, we present a first evaluation of the corpus creation effort, the inter-annotator agreement obtained for meaning assessment so far, and the next steps planned.

Amaral, L. & D. Meurers (2008). From Recording Linguistic Competence to Supporting Inferences about Language Acquisition in Context. *Computer-Assisted Language Learning* 21(4), 323–338.

Bachman, L. F. & A. S. Palmer (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford, UK: Oxford University Press.

Bailey, S. & D. Meurers (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*. Columbus, Ohio, pp. 107–115.

Fitzpatrick, E. & M. S. Seegmiller (2004). The Montclair electronic language database project. In U. Connor & T. Upton (eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*, Amsterdam: Rodopi.