Compiling a Task-Based Corpus for the Analysis of Learner Language in Context

Detmar Meurers, Niels Ott, Ramon Ziai

SFB 833 Projekt A4, Universität Tübingen

{dm,nott,rziai}@sfs.uni-tuebingen.de

1 Motivation: Why a Task-Based Corpus

Corpora in linguistics and computational linguistics have traditionally been assembled from data sources such as newspaper texts, books and, more recently, the web. While these sources provide large quantities of language data, typically very little or nothing is known about the context under which the text has been produced. The only information an analysis can refer to is the text itself, e.g., when a sentence is analyzed using the preceding sentences for disambiguation. However, language is always produced in a concrete extra-linguistic context. This contextual setting includes world knowledge and situational knowledge, i.e., the aspects of world knowledge which are relevant to interpret the given text and the concrete task and situation that the language was produced for.

The notion of a task and the evaluation of language in context plays a particularly important role in foreign language teaching and learning (cf., e.g., Ellis 2003) and a representation of the learner's ability to use language in context and to perform tasks using appropriate task strategies has been argued to be crucial for learner modeling (Amaral and Meurers 2008). However, the so-called learner corpora created to document the language produced by language learners typically consist only of learner essays.

In this paper, we present our efforts at collecting a longitudinal learner corpus consisting of the answers to reading comprehension questions, including an explicit representation of the task context and learner information. After introducing the data sources and characteristics of the corpus we are collecting, we discuss the development of the open-source WELCOME tool, which we have created to facilitate the interdisciplinary exchange of the contextualized learner corpus between the language programs providing the data and the computational linguists working on its encoding and automatic analysis.

2 Data Sources and Characteristics

The learner corpus we have started collecting consists of answers to German reading comprehension questions written by American college students learning German. 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 and 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. Unlike Bailey and Meurers, we do not use a category "alternate answer" for answers which are semantically appropriate but are unrelated to the target answers specified by the teacher. Instead, the annotators specify new target answers for these cases, marking them as alternates.

The data for our corpus is collected in two of the largest German programs in the US, at The Ohio State University (OSU) and Kansas University (KU). We intentionally focus on the relatively homogeneous foreign language learner populations at these Midwestern universities, where the students' exposure to German is mostly through the classroom setting. In contrast, second language learners of German studying at a German university have a wide range of first language backgrounds and are exposed to the influence of everyday life interactions in German.

In addition to the primary text and task data, we also collect metadata on the learners. This includes background information such as age, gender, previous exposure to German, other foreign languages learned, and time spent in a German-speaking country. Since we plan to track individual learners over a period of at least four years, some of the metadata is bound to change. To account for this, updated versions of each student's metadata are collected at the beginning of each term, yielding a connected history of student records. These student records mark points in time that can be related to the actual learner performances in the corpus. The content assessment of the learner answers to the reading comprehension tasks can thus be directly related to the learner metadata at the corresponding points in time.

Data will be collected in all sections at all levels of instruction during the entire corpus compilation process. Both universities offer several courses at the same level simultaneously and courses at different levels are held each semester, which will result in the first large, longitudinal learner corpus including an explicit task context. The resulting corpus should be rich enough to provide empirical insights for a range of different research perspectives, from second language acquisition research into the specifics of learner language and interlanguage development to computational linguistic analysis of answers to reading comprehension questions and the use of such analysis in ICALL systems.

3 Corpus Structure and Collection

To support the distributed data entry for the kind of task-based corpus we are building, we developed the WEb-based Learner COrpus MachinE (WELCOME). Using state-of-the-art web technology, WELCOME is closer to a desktop program in its use than to a web site; yet it requires only a standard web browser for entry and assessment of the data and meta data. A screenshot of the tool showing the structured data entry for the reading comprehension task is shown in Figure 2, at the end of the paper. The system avoids the use of computational or formal linguistic terminology and instead employs terms and units from foreign language teaching and learning. Using this strategy and by organizing the interface around the language learning task from a teacher perspective, the foreign language programs in charge of entering and assessing the learner data are able to incrementally assemble a structured corpus with a complex data model in a distributed way.

The corpus is collected in a relational database which supports references between the different kinds of data. Reading texts are linked into specific exercises that make use of them, along with the questions that pertain to the text in the specific exercise. The original exam sheet is saved as a file upload with the exercise itself. Each exercise can have multiple questions, where each question can again have multiple target answers. Student submissions are stored for each completed exercise sheet. They consist of the student answers themselves as well as references to the other relevant parts of a submission, the exercise and the student. The student answers again contain references to their assessment, as well as to the target answer they most closely resemble. The corpus layout is visualized in Figure 1.

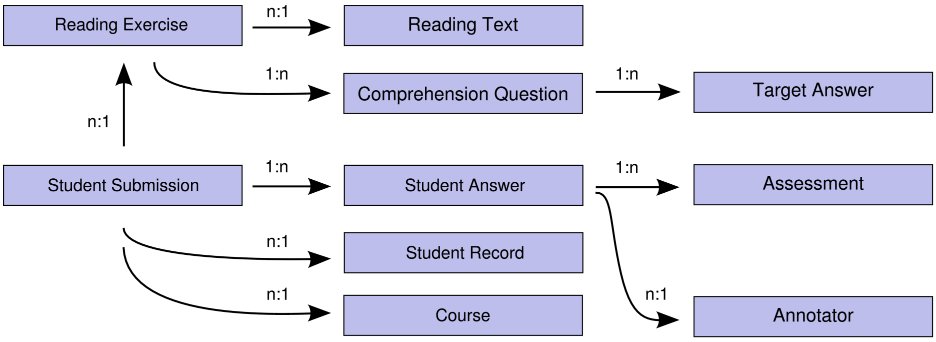


Figure 1: Corpus layout

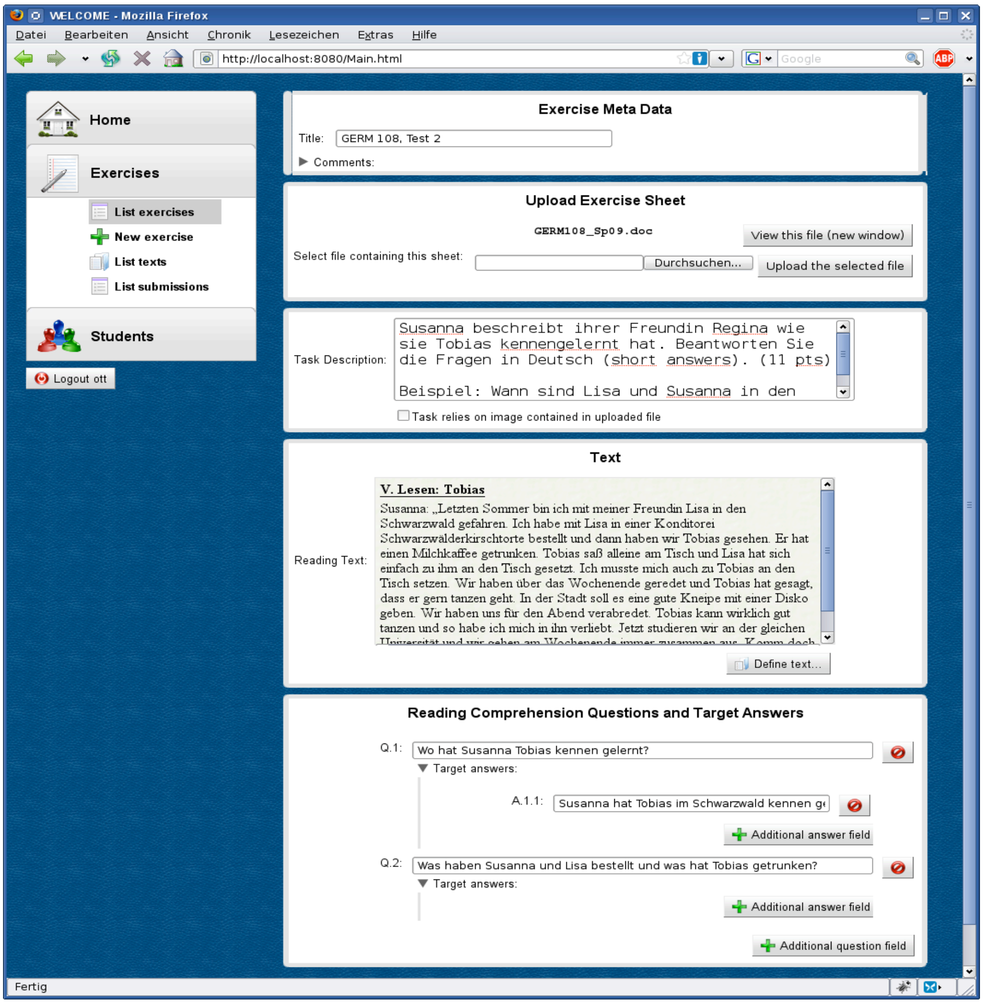
The corpus collected in this way will, to the best of our knowledge, be the first task-specific longitudinal learner corpus making both the task context and the relevant learner properties explicit. Both the corpus and the WELCOME tool will be made freely available for research use. We also intend to publish WELCOME as an open-source project that can be adapted to other approaches of collecting structured corpora that use a different data model.

References

Amaral, L. and D. Meurers (2008). From Recording Linguistic Competence to Supporting Inferences about Language Acquisition in Context. *Computer Assisted Language Learning*. 21 (5). http://purl.org/dm/papers/amaral-meurers-call08.html

Bailey, S. and D. Meurers (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In Proceedings of the 3rd ACL Workshop on Innovative Use of NLP for Building Educational Applications, Columbus, Ohio. pp. 107–115. http://purl.org/dm/papers/bailey-meurers-08.html

Ellis, R. (2003). Task-based Language Learning and Teaching. Oxford: OUP.



Exercise Meta Data

Title: GERM 108, Test 2

▶ Comments:

Upload Exercise Sheet

GERM108_Sp09.doc

Select file containing this sheet:

Task Description: Susanna beschreibt ihrer Freundin Regina wie sie Tobias kennengelernt hat. Beantworten Sie die Fragen in Deutsch (short answers). (11 pts)

Beispiel: Wann sind Lisa und Susanna in den

□ Task relies on image contained in uploaded file

Text

Reading Text: **V. Lesen: Tobias**

Susanna: „Letzten Sommer bin ich mit meiner Freundin Lisa in den Schwarzwald gefahren. Ich habe mit Lisa in einer Konditorei Schwarzwälderkirschtorte bestellt und dann haben wir Tobias gesehen. Er hat einen Milchkaffee getrunken. Tobias saß alleine am Tisch und Lisa hat sich einfach zu ihm an den Tisch gesetzt. Ich musste mich auch zu Tobias an den Tisch setzen. Wir haben über das Wochenende geredet und Tobias hat gesagt, dass er gern tanzen geht. In der Stadt soll es eine gute Kneipe mit einer Disko geben. Wir haben uns für den Abend verabredet. Tobias kann wirklich gut tanzen und so habe ich mich in ihn verliebt. Jetzt studieren wir an der gleichen Universität und wir gehen am Wochenende immer zusammen aus. Komm doch

Reading Comprehension Questions and Target Answers

Q.1: Wo hat Susanna Tobias kennen gelernt?

▼ Target answers:

A.1.1: Susanna hat Tobias im Schwarzwald kennen ge

Q.2: Was haben Susanna und Lisa bestellt und was hat Tobias getrunken?

▼ Target answers:

Figure 2: WELCOME user interface