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

Detmar Meurers, Niels Ott, Ramon Ziai

SFB 833

Eberhard Karls Universität Tübingen

Motivation 1: Tasks to Obtain Language in Context

- Language is produced in concrete linguistic and extra-linguistic contexts.
  - This contextual setting includes world knowledge and situational knowledge.
- We want to make the context explicit by collecting data in the setting of a concrete task.
- In what authentic settings does such data arise?

Motivation 2: Tasks in Language Learning

- Language in context plays an important role in foreign language teaching (cf., e.g., Ellis 2003).
- Observing learner's ability to perform tasks crucial for learner modeling (Amaral & Meurers, 2008)
- Yet, current learner corpora typically consist of essay data with little or no explicit information about context or concrete task requirements.

Combining the two: Collecting a Task-Based Learner Corpus

- We are compiling a longitudinal learner corpus consisting of answers to reading comprehension questions on given texts.
- The student's task is explicitly defined by the reading comprehension questions.
- The reading text in a reading comprehension task defines a concrete linguistic context.
- Questions which refer to the information explicitly or implicitly given in the text can be answered without reference to world knowledge.

➥ Context fully accessible through linguistic analysis

Obtaining the Data

- Data is collected in two of the largest German programs in the US, at Kansas University (Prof. Nina Vyatkina) and The Ohio State University (Prof. Kathryn Corl).
- Collect at four course levels over a period of four years
- The group of learners is relatively homogeneous:
  - typically English native speakers
  - exposure to German mostly limited to classroom
- Problem: How can a division of labor be achieved where
  - language instructors enter data in a distributed manner,
  - a centrally stored corpus with a complex structure and learner meta-data results?

Tackling Distributed Corpus Creation: The WELCOME Tool

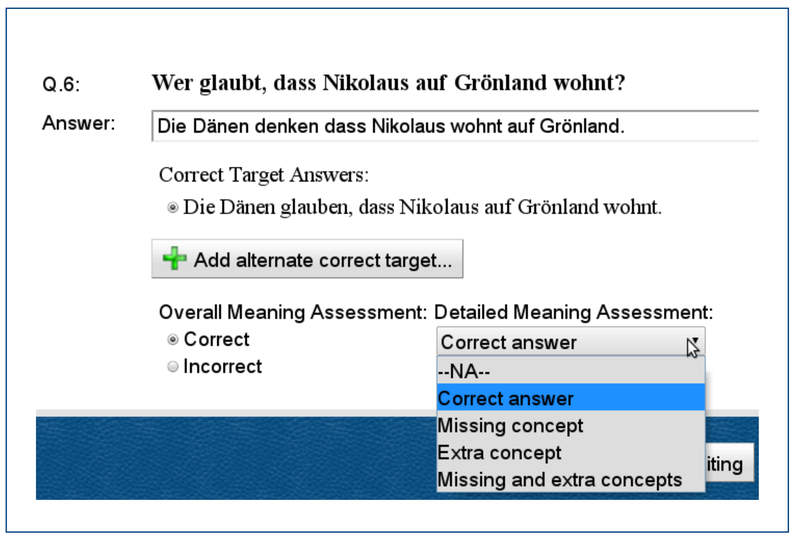
- The WEb-based Learner COrpus MachinE (WELCOME)
  - supports distributed data entry by language instructors
  - stores the complexly structured corpus and learner meta-data in a central repository
- WELCOME behaves similar to a desktop application, but only requires a web browser on the user's computer.
- Interface organized around the work-flow of language instructors, supporting incremental assembly of a structured corpus with a complex data model.
- The back-end is a relational database engine providing efficient data storage and access on the server.

Tackling Longitudinality of Meta-Data

- Problem: Some student meta-data (e.g., exposure to German) changes over time, making it necessary to record these developments in the meta-data.
- Solution: Store multiple, connected meta-data records of the same student, and associate the records with student performances via term and date information.

Content Assessment

- Student answers are assessed by two independent annotators with respect to meaning (not form).
  - Binary classification: appropriate vs. inappropriate
  - Fine-grained classification of comparison with target answers (Bailey & Meurers, 2008) is extended:
    - both appropriate and inappropriate answers can be sub-classified into *missing concept*, *extra concept*, and *blend*
    - sub-class *non-answer* kept for inappropriate
    - Instead of an *alternate answer* category for appropriate but unexpected answers, these can be added to target answer set by the annotators.



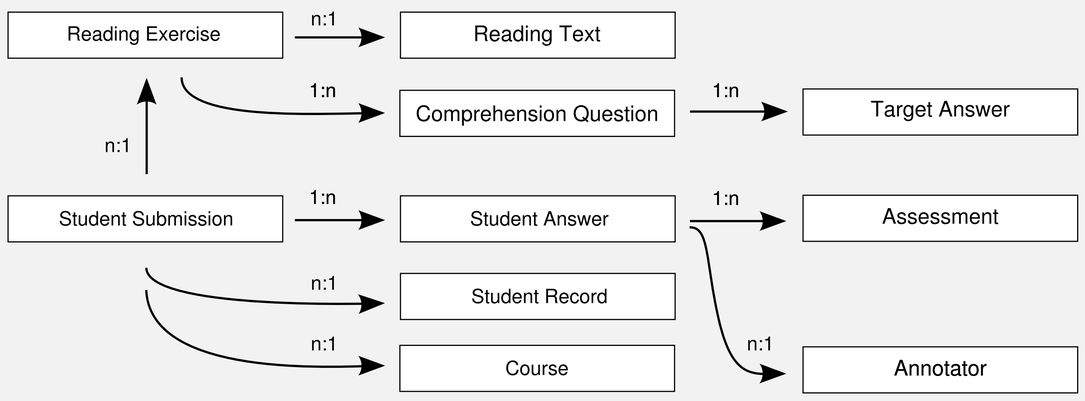
- The learner answers are copied from the handwritten submissions by both annotators since transcribing handwritten text is an interpretative task already.

Related Work

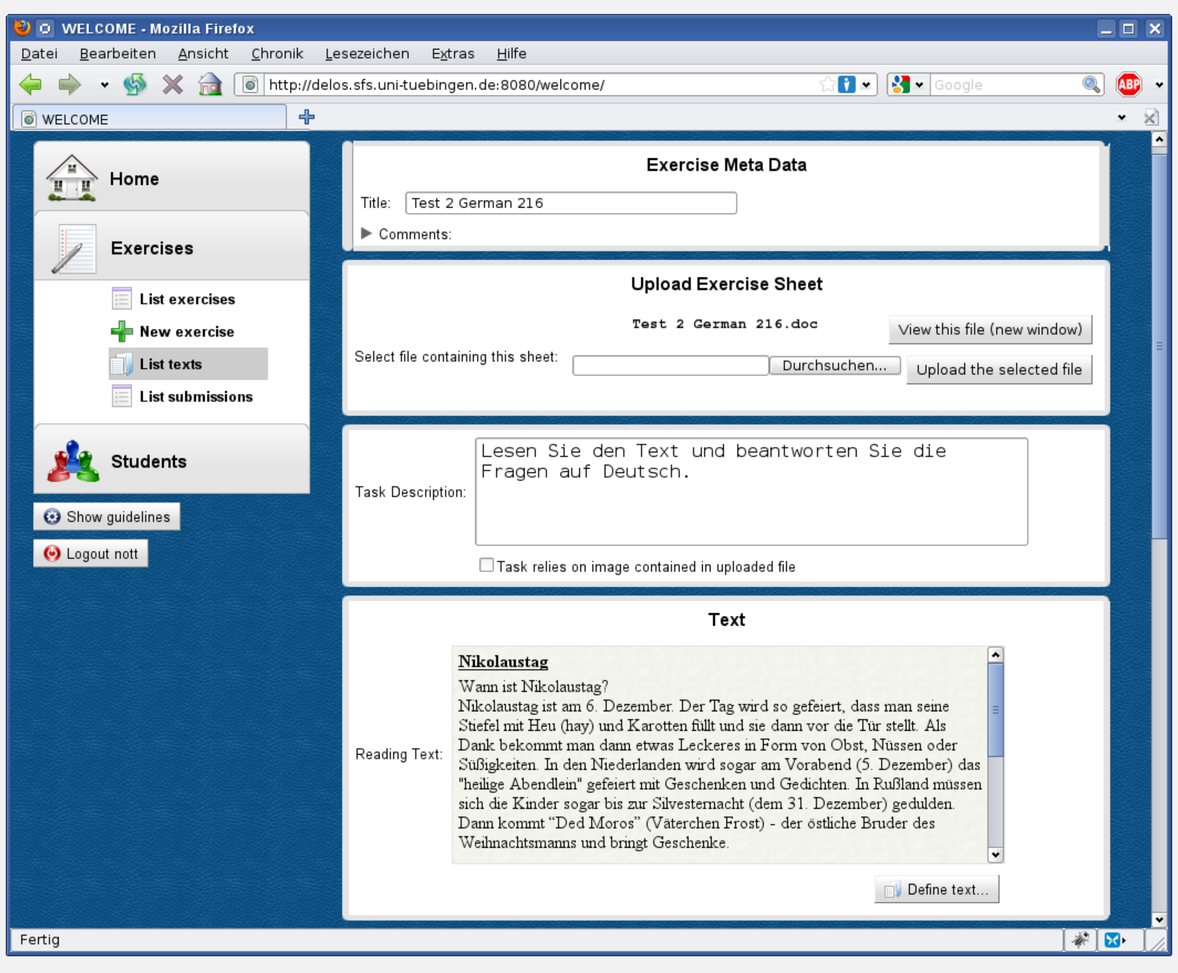
- In contrast to most learner corpus work, the task-based setting allows us to analyze and compare meaning (not only form).
- The task-oriented TRAINS corpus (Heeman & Allen, 1995) collects dialogues based on a concrete, not-linguistically encoded micro-world.
- We are interested in the semantic appropriateness of learner answers with respect to the question given the "micro-world" made explicit by the reading text.
  - All relevant information for evaluating the learner answers is linguistically accessible in the questions and the text they are about.

Corpus Ingredients and Layout

- Reading texts
- Comprehension questions
- Target answers by teachers
- Student answers and assessment
- Student meta-data:
  - age, gender
  - native language
  - previous exposure to German
  - other languages spoken
  - . . .



Screenshot of the WELCOME Tool



Uses of the Corpus

A task-based learner corpus of reading comprehension exercises and student answers can provide insights for different research perspectives, such as

- the automatic comparison of meaning representations in SFB 833 Project A4 "Comparing Meaning in Context: Components of a Shallow Semantic Analysis".
- Learner language and interlanguage development in theories of second language acquisition.
- Linguistic analysis of language in context (e.g., information structure).
- Application of automatic meaning evaluation in Intelligent Language Tutoring Systems.

References

Amaral, L. & D. Meurers (2008). From Recording Linguistic Competence to Supporting Inferences about Language Acquisition in Context: Extending the Conceptualization of Student Models for Intelligent Computer-Assisted Language Learning. *Computer-Assisted Language Learning* 21(4), 323–338.

Bailey, S. & D. Meurers (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In J. Tetreault, J. Burstein & R. D. Felice (eds.), *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications, held at ACL 2008*. Columbus, Ohio: Association for Computational Linguistics, pp. 107–115.

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford, UK: Oxford University Press.

Granger, S. (2008). Learner Corpora in Foreign Language Education. In N. V. Deusen-Scholl & N. H. Hornberger (eds.), *Encyclopedia of Language and Education. Volume 4: Second and Foreign Language Education*, Springer Science and Business Media, pp. 337–351. 2nd ed.

Heeman, P. A. & J. Allen (1995). *The Trains 93 Dialogues*. Tech. rep., The University of Rochester, Computer Science Department. TRAINS Technical Note 94-2.