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

Detmar Meurers, Niels Ott and Ramon Ziai

Universität Tübingen, SFB 833

Pre-Conference Workshop on Learner Corpora at ALOES 2010
Paris. March 25, 2010

Outline

Project background

- Project A4 in the SFB 833: *Comparing Meaning in Context: Components of a shallow semantic analysis*
- Research question:
  - How can the meaning of sentences and text fragments be analyzed and compared in realistic situations?
  - Realistic situations:
    - language not necessarily well-formed
    - differences in situative and world knowledge
- Two challenges:
  - Which linguistic representations can be robustly identified as basis of a computational approximation of meaning?
  - How can the role of the context be integrated?

⇒ Start by collecting data of authentic language in context.

Collecting data in authentic tasks

- We want to make the context explicit by collecting data in the setting of a concrete task.
  - To support evaluation of meaning, focus on tasks using information encoded in language, not world knowledge.
- In which authentic settings does such data arise?
- Language in context plays an important role in *foreign language teaching* (cf., e.g., Ellis 2003).
  - Yet, current learner corpora typically consist of essay data (cf., e.g., Granger 2008), so only the essay topic is known; contents often unconstrained and not predictable.
  - Which activities provide more explicit, language-based context? We focus on reading comprehension questions.

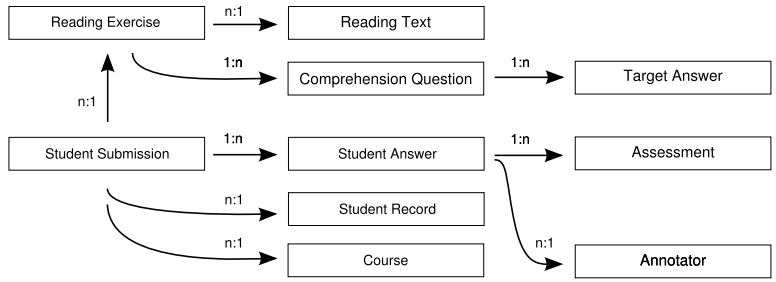
⇒ Compile a corpus with answers to reading comprehension questions written by learners of German.

Corpus ingredients

1. Texts asked about in reading comprehension
   - i.e., the explicit, language-based context
2. Comprehension questions
3. Target answers by teachers
4. Student answers
5. Teacher assessment of student answers
   5.1 binary: correct/incorrect meaning
   5.2 detailed meaning analysis
6. Student meta-data:
   6.1 age, gender
   6.2 native language
   6.3 previous exposure to German
   6.4 other languages spoken
   6.5 ...

Corpus ingredients

Database structure



Obtaining the Data

- Collected in two of the largest German programs in US
- Our project collaborates with two subcontractors:
  - Kansas University (Prof. Nina Vyatkina)
  - Ohio State University (Prof. Kathryn Corl)
- Data is collected
  - at four course levels
  - over a period of four years.
- Why are we collecting outside of Germany?
  Controlled context, with a homogeneous group of learners:
  - typically English native speakers
  - exposure to German mostly limited to the classroom

Towards effective distributed data collection

- Dissociation of source and processing of corpus data:
  - Language instructors in the US are the foreign language teaching experts in touch with the learners.
  - Computational linguists in Germany responsible for storing and processing the corpus.

  ⇒ Distributed data entry, central standardized storage
- Requirements:
  - Entering the data must be straightforward for the language instructors.
  - Approach must support the complex structure including learner meta-data.
- How can we meet these requirements?

  ⇒ Develop a web-based tool for data collection.

The WELCOME Tool

- To address the requirements, we developed the WEb-based Learner COrpus MachinE (WELCOME).
  - It supports distributed data entry by language instructors and stores all data in a central repository.
- WELCOME behaves similar to a desktop application but requires only a web browser and Internet access.
- The interface is
  - optimized around the work-flow of language instructors,
  - supports the incremental entry of data resulting in a structured corpus.
- As its back-end, it uses a relational database engine, representing and enforcing the complex corpus structure.
  - efficient, well-tested
  - supports incremental data manipulation and querying
  - allows concurrent access by multiple users
  - Data can be exported into standard XML formats.

Longitudinality of Meta-Data

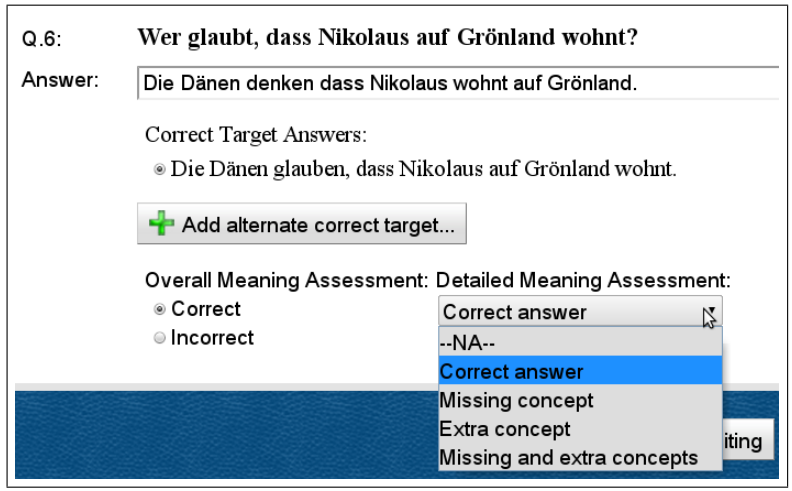
- Student meta-data change over time
  - e.g., exposure to German
- Ideally, one would collect the meta-data together with each reading comprehension task.
  - Massive overhead, incompatible with integration of data collection into regular classes
- Compromise: Student meta-data collected once per term.
  - These records are connected via IDs for each student, so we can track each student's development over time.
- The reading comprehension answers are stored with a specific date, supporting more fine-grained tracking of development.

Content Assessment

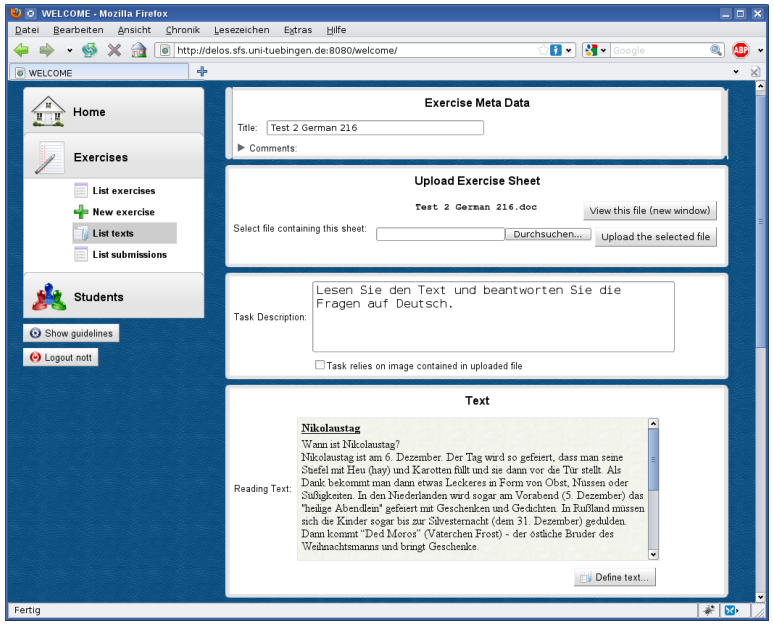
- Student answers are assessed by two independent annotators with respect to meaning (not form).
  1. The learner answers are independently transcribed from the handwritten submissions by each annotator.
     - Why two annotators? Transcribing handwritten text is an interpretation.
  2. Binary classification: appropriate vs. inappropriate
     - 'Is the answer given by the student a valid answer to the reading comprehension question?'
  3. Fine-grained classification of comparison with target answers based on Bailey & Meurers (2008)
     - For appropriate and inappropriate answers: *missing concept*, *extra concept*, *blend*
     - Additional answer category for inappropriate answers: *non-answer*
     - Alternate answers, which are appropriate but differ in contents from target answer, can be added by annotators.

Content Assessment

Example



WELCOME Demo (http://purl.org/icall/welcome)



Conclusion

- We motivated the creation of task-based corpora of authentic language data in context.
- We are collecting a longitudinal learner corpus of German reading comprehension exercises.
  - includes rich structure: context, student data and meta-data, teacher targets and assessment
- WELCOME tool supports distributed data entry and central, standardized corpus storage.
  - We will make the tool freely available for research.
- Corpus resulting from our collaboration with US German programs serves as empirical basis for our research on comparing meaning in context. It more generally supports:
  - SLA research on learner language development
  - linguistic research into language in context, e.g., interaction of syntax and information structure

References

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. URL http://aclweb.org/anthology-new/W08-0913.

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.