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

Detmar Meurers, Niels Ott, Ramon Ziai

Universität Tübingen
SFB 833, Project A4

Workshop "Learner Corpora and Corpora for Learners"
Conference on "Multilingual Individuals and Multilingual Societies (MIMS)
SFB 538, Hamburg. October 7, 2010

Outline

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:
    - differences in situative and world knowledge
    - language not necessarily well-formed
- 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, learner corpora typically consist of essay data (cf., e.g., Granger 2008), so only the essay topic is known → contents quite unconstrained and not predictable.
  - Other activities provide more explicit, language-based context with predictable contents: reading comprehension, summarization, information-gap activities, . . .

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

Compiling a task-based learner corpus

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: nature of meaning divergence
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 ...

An English example (Bailey & Meurers 2008)

Question: What are the methods of propaganda mentioned in the article?

Target: The methods include use of labels, visual images, and beautiful or famous people promoting the idea or product. Also used is linking the product to concepts that are admired or desired and to create the impression that everyone supports the product or idea.

Learner Responses:

- A number of methods of propaganda are used in the media.
- Bositive or negative labels.
- Giving positive or negative labels. Using visual images. Having a beautiful or famous person to promote. Creating the impression that everyone supports the product or idea.

Creation & Analysis: Reading Comprehension Corpus
Detmar Meurers, Niels Ott, Ramon Ziai
Background and Motivation
Comparing meaning in context
Data from authentic tasks
Compiling a Task-Based Learner Corpus
Corpus ingredients
Obtaining the Data
WELCOME Tool
Annotating content assessment
Examples
Analysis
Current corpus snapshot
Interannotator agreement
Sources of disagreement
Examples
Meaning assessment results
Conclusion
Appendix
SFB 833
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Obtaining the Data

- Collected in two of the largest German programs in US
  - Kansas University (Prof. Nina Vyatkina)
  - Ohio State University (Prof. Kathryn Corl)
- Data is collected
  - at four course levels (beginners to advanced)
  - over a period of four years.
- Student meta-data is collected once per term.
  - These records are connected via IDs for each student, so we can track each student's development over time.
- Why are we collecting outside of Germany?
  Controlled context, with a homogeneous group of learners:
  - English native speakers (mostly)
  - exposure to German mostly limited to the classroom

Creation & Analysis: Reading Comprehension Corpus
Detmar Meurers, Niels Ott, Ramon Ziai
Background and Motivation
Comparing meaning in context
Data from authentic tasks
Compiling a Task-Based Learner Corpus
Corpus ingredients
Obtaining the Data
WELCOME Tool
Annotating content assessment
Examples
Analysis
Current corpus snapshot
Interannotator agreement
Sources of disagreement
Examples
Meaning assessment results
Conclusion
Appendix
SFB 833
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

The WELCOME tool for distributed corpus collection and annotation

- To support distributed data entry by language instructors into a centralized corpus repository, we developed the WEb-based Learner COrpus MachinE (WELCOME).
- 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, concurrent access by multiple users
  - supports incremental data manipulation and querying
  - provides export of data in an XML-based format

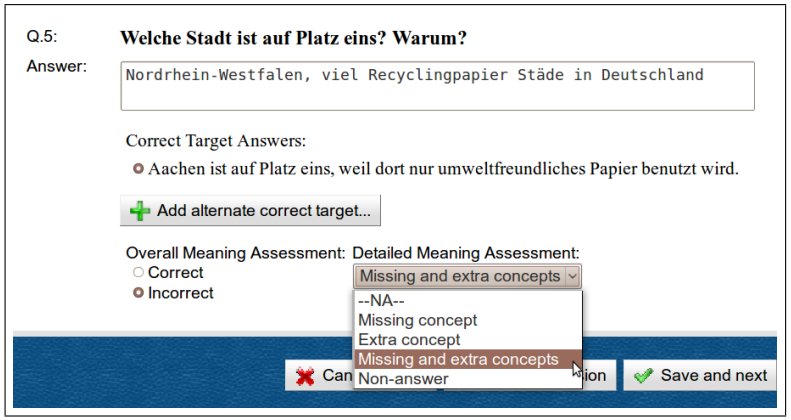
Creation & Analysis: Reading Comprehension Corpus
Detmar Meurers, Niels Ott, Ramon Ziai
Background and Motivation
Comparing meaning in context
Data from authentic tasks
Compiling a Task-Based Learner Corpus
Corpus ingredients
Obtaining the Data
WELCOME Tool
Annotating content assessment
Examples
Analysis
Current corpus snapshot
Interannotator agreement
Sources of disagreement
Examples
Meaning assessment results
Conclusion
Appendix
SFB 833
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Annotating content assessment

- Student answers are assessed by two independent annotators with respect to meaning (not form):
  - Is the answer given by the student a meaningful answer to the reading comprehension question?
- Annotation steps:
  1. For handwritten student submissions: Learner answers are independently transcribed by each annotator.
  2. Binary classification of comparison with target answer.
     - Where more than one target answer exists, annotator identifies most similar one in terms of meaning.
  3. Detailed classification of comparison with target answer:
     - identical meaning
     - missing concept, extra concept, blend
     - incomparable meaning
- Annotation scheme extends Bailey & Meurers (2008).

Annotating content assessment

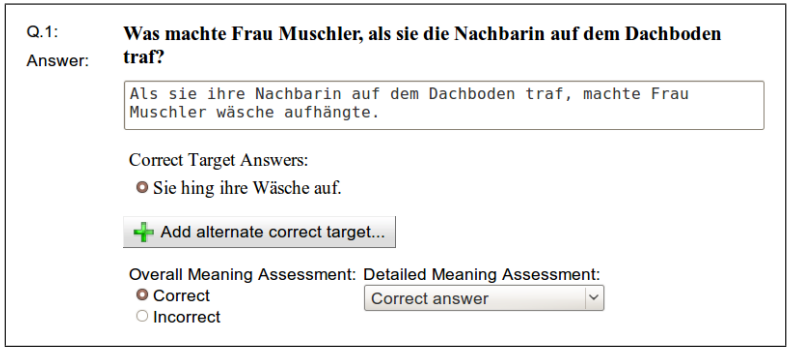
Example in WELCOME



Creation & Analysis: Reading Comprehension Corpus
Detmar Meurers, Niels Ott, Ramon Ziai
Background and Motivation
Comparing meaning in context
Data from authentic tasks
Compiling a Task-Based Learner Corpus
Corpus ingredients
Obtaining the Data
WELCOME Tool
Annotating content assessment
Examples
Analysis
Current corpus snapshot
Interannotator agreement
Sources of disagreement
Examples
Meaning assessment results
Conclusion
Appendix
SFB 833
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Annotating content assessment

Example in WELCOME (2)



Creation & Analysis: Reading Comprehension Corpus
Detmar Meurers, Niels Ott, Ramon Ziai
Background and Motivation
Comparing meaning in context
Data from authentic tasks
Compiling a Task-Based Learner Corpus
Corpus ingredients
Obtaining the Data
WELCOME Tool
Annotating content assessment
Examples
Analysis
Current corpus snapshot
Interannotator agreement
Sources of disagreement
Examples
Meaning assessment results
Conclusion
Appendix
SFB 833
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Current corpus snapshot

- Discuss snapshot after one of four years (27.9.2010), based on the data collected at Kansas University:
  - 34 texts
  - 242 questions
    - so far 138 with student answers scored by two annotators
  - 4.908 student answers, written by 181 students
    - so far primarily beginner data, some intermediate
- Other components of the corpus (not discussed here):
  - Ohio State University:
    - parallel to Kansas University
  - Tübingen University:
    - 20 texts, 100 native speakers (control group)
    - 144 questions, 4.414 answers

Creation & Analysis: Reading Comprehension Corpus
Detmar Meurers, Niels Ott, Ramon Ziai
Background and Motivation
Comparing meaning in context
Data from authentic tasks
Compiling a Task-Based Learner Corpus
Corpus ingredients
Obtaining the Data
WELCOME Tool
Annotating content assessment
Examples
Analysis
Current corpus snapshot
Interannotator agreement
Sources of disagreement
Examples
Meaning assessment results
Conclusion
Appendix
SFB 833
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Interannotator agreement

- 3.635 student answers with two annotations (binary):
  - agreement between reviewers: 3114 (85.7%)
  - $\kappa = 0.650$
- 3.628 student answers with two detailed annotations
  - agreement between reviewers: 3108 (85.7%)
  - $\kappa = 0.757$
- agreement of 85% to 88% in Bailey & Meurers (2008)

Use of Categories by individual annotators

Binary

	correct meaning	incorrect meaning
Annotator Y	66.20%	33.80%
Annotator K	77.75%	22.25%

Detailed

	identical	extra concept	missing concept	blend	incomparable
Y	55.28%	3.85%	15.10%	23.71%	2.04%
K	60.29%	1.71%	14.25%	23.21%	0.36%

Creation & Analysis: Reading Comprehension Corpus
Detmar Meurers, Niels Ott, Ramon Ziai
Background and Motivation
Comparing meaning in context
Data from authentic tasks
Compiling a Task-Based Learner Corpus
Corpus ingredients
Obtaining the Data
WELCOME Tool
Annotating content assessment
Examples
Analysis
Current corpus snapshot
Interannotator agreement
Sources of disagreement
Examples
Meaning assessment results
Conclusion
Appendix
SFB 833
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Sources of disagreement

Binary Assessment and Detailed Differences

- How many cases are agreements in detailed classes, but disagreement in binary evaluation?

⇒ 359 (9,9% of all answers)

ident.	extra concept	missing concept	blend	incomp.
0	8 (2.23%)	282 (78.55%)	69 (19.22%)	0

Creation & Analysis: Reading Comprehension Corpus
Detmar Meurers, Niels Ott, Ramon Ziai
Background and Motivation
Comparing meaning in context
Data from authentic tasks
Compiling a Task-Based Learner Corpus
Corpus ingredients
Obtaining the Data
WELCOME Tool
Annotating content assessment
Examples
Analysis
Current corpus snapshot
Interannotator agreement
Sources of disagreement
Examples
Meaning assessment results
Conclusion
Appendix
SFB 833
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Sources of disagreement

Examples (1)

Question: *Welche Filme kann man im November sehen?*

Target: *Im November kann man Schwarz auf Weiß und (500) Days of Summer sehen.*

Student: *Man kann Schwarz auf Weiß sehen*

- Annotator Y: incorrect, missing concept
- Annotator K: correct, missing concept

⇒ Annotation guidelines should specify how complete enumerations must be for correct answer.
  - depends on question type and other triggers

Creation & Analysis: Reading Comprehension Corpus
Detmar Meurers, Niels Ott, Ramon Ziai
Background and Motivation
Comparing meaning in context
Data from authentic tasks
Compiling a Task-Based Learner Corpus
Corpus ingredients
Obtaining the Data
WELCOME Tool
Annotating content assessment
Examples
Analysis
Current corpus snapshot
Interannotator agreement
Sources of disagreement
Examples
Meaning assessment results
Conclusion
Appendix
SFB 833
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Sources of disagreement

Examples (2)

Question: *Welche 2 Städte besucht Heike im Urlaub?*

Target: *Heike besucht Berlin und Eutin.*

Student: *Fahre ich manchmal nach Eutin.*

- Annotator Y: incorrect, missing concept
- Annotator K: correct, missing concept

⇒ Learner strategies used to answer questions are relevant, in particular lifting. E.g., the text here included:
  - *Im Sommerurlaub, fahre ich manchmal nach Eutin.*

Sources of disagreement

Examples (3)

Question: *Haben alle Zimmer eine Dusche?*

Target: *Nein, nicht alle Zimmer haben eine Dusche.*

Student: *Nein, alle Zimmer haben keine Dusche.*

- Annotator Y: correct, extra concept
- Annotator K: incorrect, blend

⇒ Amend annotation guidelines:
  - If ambiguous sentence has a correct reading, mark as correct and provide detailed assessment for that reading.
  - Take the full answer into account (e.g., not just the “Nein”).

Creation & Analysis: Reading Comprehension Corpus
Detmar Meurers, Niels Ott, Ramon Ziai
Background and Motivation
Comparing meaning in context
Data from authentic tasks
Compiling a Task-Based Learner Corpus
Corpus ingredients
Obtaining the Data
WELCOME Tool
Annotating content assessment
Examples
Analysis
Current corpus snapshot
Interannotator agreement
Sources of disagreement
Examples
Meaning assessment results
Conclusion
Appendix
SFB 833
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

A first look at meaning assessment

Analysis based on assessments where both annotators agree.

- binary (3114 total answers):
  - correct (adequate meaning): 75.7% (2.356)
  - incorrect (inadequate meaning): 24.3% (758)
- detailed classification of comparison (3108 total answers):
  - blend: 22.9% (711)
  - correct: 62.7% (1948)
  - extra concept: 1% (33)
  - missing concept: 13% (404)
  - non-answers: 0.4% (12)
- relating both assessments (2749 answers):

	identical	extra concept	missing concept	blend	incomp.
correct	93.2%	1.2%	5.3%	0.4%	0%
incorrect	0	0	1.8%	96.4%	1.8%

Creation & Analysis: Reading Comprehension Corpus
Detmar Meurers, Niels Ott, Ramon Ziai
Background and Motivation
Comparing meaning in context
Data from authentic tasks
Compiling a Task-Based Learner Corpus
Corpus ingredients
Obtaining the Data
WELCOME Tool
Annotating content assessment
Examples
Analysis
Current corpus snapshot
Interannotator agreement
Sources of disagreement
Examples
Meaning assessment results
Conclusion
Appendix
SFB 833
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

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. (→ open source)
- We use the corpus as empirical basis for our research on automatic meaning comparison in context.
  - task and annotation scheme supports good inter-annotator agreement
- More generally, the corpus will be available for SLA research on learner language development and linguistic research into language in context.

Creation & Analysis: Reading Comprehension Corpus
Detmar Meurers, Niels Ott, Ramon Ziai
Background and Motivation
Comparing meaning in context
Data from authentic tasks
Compiling a Task-Based Learner Corpus
Corpus ingredients
Obtaining the Data
WELCOME Tool
Annotating content assessment
Examples
Analysis
Current corpus snapshot
Interannotator agreement
Sources of disagreement
Examples
Meaning assessment results
Conclusion
Appendix
SFB 833
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

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 (BEA-3) at ACL'08*. Columbus, Ohio, pp. 107–115. URL http://aclweb.org/anthology/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.

Meurers, D., N. Ott & R. Ziai (2010). Compiling a Task-Based Corpus for the Analysis of Learner Language in Context. In *Proceedings of Linguistic Evidence*. Tübingen, pp. 214–217. URL http://purl.org/dm/papers/meurers-ott-ziai-10.html.

Extending Bailey & Meurers (2008)

- Current work focuses on German instead of English:
  - richer variation in forms and word order
- Current annotation scheme supports
  - detailed classification of meaning differences for both binary subcases (instead of only for inappropriate ones).
  - dynamic addition of alternate answers as targets
- Corpus currently being collected is significantly larger (currently 6 times, planned 50 times), which is crucial for investigating
  - detailed classification of meaning differences
  - identification of islands of compositionality
  - role of givenness in meaning assessment
  - impact of task strategies
  - different context types, e.g.,
    - different question types
    - different encodings of requested information in text

Creation & Analysis: Reading Comprehension Corpus
Detmar Meurers, Niels Ott, Ramon Ziai
Background and Motivation
Comparing meaning in context
Data from authentic tasks
Compiling a Task-Based Learner Corpus
Corpus ingredients
Obtaining the Data
WELCOME Tool
Annotating content assessment
Examples
Analysis
Current corpus snapshot
Interannotator agreement
Sources of disagreement
Examples
Meaning assessment results
Conclusion
Appendix
SFB 833
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

More complex examples

Question: Warum darf der hessische Apfelwein nicht mehr Wein genannt werden?

Target: Wo „Wein“ drauf steht, muss ein Getränk aus Trauben drin sein.

Learner Responses:

- Viele Leute wollen dass Apfelwein bleibt Apfelwein.
- Ein neues EU-Gesetz würde voraussetzen, dass Apfelwein nicht mehr ”wein” heissen darf.
- Wegen ein neues EU-Gesetz: Wo ”Wein” drauf steht, muss ein Getränk aus Trauben drin sein. Daher dürfte der gute alte Apfelwein in der Zukunft nicht mehr ”Apfelwein” heißen.

Creation & Analysis: Reading Comprehension Corpus
Detmar Meurers, Niels Ott, Ramon Ziai
Background and Motivation
Comparing meaning in context
Data from authentic tasks
Compiling a Task-Based Learner Corpus
Corpus ingredients
Obtaining the Data
WELCOME Tool
Annotating content assessment
Examples
Analysis
Current corpus snapshot
Interannotator agreement
Sources of disagreement
Examples
Meaning assessment results
Conclusion
Appendix
SFB 833
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

More complex examples (2)

Question: Wofür ist der Aufsichtsrat verantwortlich?

Target: Der Aufsichtsrat ist für die Bestellung, den Abruf und die Überwachung des Vorstandes verantwortlich. Außerdem ist er verantwortlich für die langfristige Planung, z.B. für die Verwendung des Gewinns der AG.

Learner Responses:

- Der Aufsichtsrat ist fuer die Bestellung verantwortlich.
- Der Aufsichtsrat beschäftigt sich mit der Bestellung, dem Abruf, der Überwachung des Vorstandes und der langfristigen Planung.

Creation & Analysis: Reading Comprehension Corpus
Detmar Meurers, Niels Ott, Ramon Ziai
Background and Motivation
Comparing meaning in context
Data from authentic tasks
Compiling a Task-Based Learner Corpus
Corpus ingredients
Obtaining the Data
WELCOME Tool
Annotating content assessment
Examples
Analysis
Current corpus snapshot
Interannotator agreement
Sources of disagreement
Examples
Meaning assessment results
Conclusion
Appendix
SFB 833
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

More complex examples: multiple targets

Question: Der Text sagt, dass die Wasserqualität in Salzburg sehr gut ist. Wie begründet der Text diese Behauptung?

Targets:

- Das Wasser aus den Grundwasserwerken ist schon von Natur aus so gut, dass es weder aufbearbeitet noch desinfiziert werden muss.
- Salzburgs Trinkwasser wird laufend kontrolliert, rund 2.400 bakteriologische und chemische Kontrollen garantieren, dass die Salzburg AG ein erstklassiges Lebensmittel ins Haus liefert.

Learner Responses:

- Die Wasserqualität ist sehr gut in Salzburg, weil es 90 Prozent davon Grundwasser aus den Bergen ist.
- Die Wasserqualität in Salzburg ist sehr gut, weil das Trinkwasser bei rund 2.400 bakteriologische und chemische Kontrollen kontrolliert ist.