Evaluating Answers to Reading Comprehension Questions in Context: Results for German and the Role of Information Structure

Detmar Meurers Ramon Ziai Niels Ott Janina Kopp
Seminar für Sprachwissenschaft / SFB 833
Universität Tübingen
Wilhelmstraße 19 / Nauklerstraße 35
72074 Tübingen, Germany
{dm,rziai,nott,jkopp}@sfs.uni-tuebingen.de

Abstract

Reading comprehension activities are an authentic task including a rich, language-based context, which makes them an interesting real-life challenge for research into automatic content analysis. For textual entailment research, content assessment of reading comprehension exercises provides an interesting opportunity for extrinsic, real-purpose evaluation, which also supports the integration of context and task information into the analysis.

In this paper, we discuss the first results for content assessment of reading comprehension activities for German and present results which are competitive with the current state of the art for English. Diving deeper into the results, we provide an analysis in terms of the different question types and the ways in which the information asked for is encoded in the text.

We then turn to analyzing the role of the question and argue that the surface-based account of information that is given in the question should be replaced with a more sophisticated, linguistically informed analysis of the information structuring of the answer in the context of the question that it is a response to.

1 Introduction

Reading comprehension exercises offer a real-life challenge for the automatic analysis of meaning. Given a text and a question, the content assessment task is to determine whether the answer given to a reading comprehension question actually answers the question or not. Such reading comprehension exercises are a common activity in foreign language teaching, making it possible to use activities which are authentic and for which the language teachers provide the gold standard judgements.

Apart from the availability of authentic exercises and independently motivated gold standard judgements, there are two further reasons for putting reading comprehension tasks into the spotlight for automatic meaning analysis. Firstly, such activities include a text as an explicit context on the basis of which the questions are asked. Secondly, answers to reading comprehension questions in foreign language teaching typically are between a couple of words and several sentences in length – too short to rely purely on the distribution of lexical material (as, e.g., in LSA, Landauer et al., 1998). The answers also exhibit a significant variation in form, including a high number of form errors, which makes it necessary to develop an approach which is robust enough to determine meaning correspondences in the presence of errors yet flexible enough to support the rich variation in form which language offers for expressing related meanings.

There is relatively little research on content assessment for reading comprehension tasks and it so far has focused exclusively on English, including both reading comprehension questions answered by native speakers (Leacock and Chodorow, 2003; Nielsen et al., 2009) and by language learners (Bailey and Meurers, 2008). The task is related to the increasingly popular strand of research on Recognizing Textual Entailment (RTE, Dagan et al., 2009) and the Answer Validation Exercise (AVE, Rodrigo et al., 2009), which both have also generally targeted English.

Proceedings of the TextInfer 2011 Workshop on Textual Entailment, EMNLP 2011, pages 1–9,
Edinburgh, Scotland, UK, July 30, 2011.

The RTE challenge abstracts away from concrete tasks to emphasize the generic semantic inference component and it has significantly advanced the field under this perspective. At the same time, an investigation of the role of the context under which an inference holds requires concrete tasks, for which content assessment of reading comprehension tasks seems particularly well-suited. Borrowing the terminology Spärck Jones (2007) coined in the context of evaluating automatic summarization systems, one can say that we pursue an extrinsic, full-purpose evaluation of aspects of textual inference. The content assessment task provides two distinct opportunities to investigate textual entailment: On the one hand, one can conceptualize it as a textual inference task of deciding whether a given text T supports a particular student answer H. On the other hand, if target answers are provided by the teachers, the task can be seen as a special bi-directional case of textual entailment, namely a paraphrase recognition task comparing the student answers to the teacher target answers. In this paper, we focus on this second approach.

The aim of this paper is twofold. On the one hand, we want to present the first content assessment approach for reading comprehension activities focusing on German. In the discussion of the results, we will highlight the impact of the question types and the way in which the information asked for is encoded in the text. On the other hand, we want to discuss the importance of the explicit language-based context and how an analysis of the question and the way a text encodes the information being asked for can help advance research on automatic content assessment. Overall, the paper can be understood as a step in the long-term agenda of exploring the role and impact of the task and the context on the automatic analysis and interpretation of natural language.

2 Data

The experiments described in this paper are based on the Corpus of Reading comprehension Exercises in German (CREG), which is being collected in collaboration with two large German programs in the US, at Kansas University (Prof. Nina Vyatkina) and at The Ohio State University (Prof. Kathryn Corl). German teachers are using the WEb-based Learner COrpus MachinE (WELCOME, Meurers et al., 2010) interface to enter the regular, authentic reading comprehension exercises used in class, which are thereby submitted to a central corpus repository. These exercises consist of texts, questions, target answers, and corresponding student answers. Each student answer is transcribed from the hand-written submission by two independent annotators. These two annotators then assess the contents of the answers with respect to meaning: Did the student provide a meaningful answer to the question? In this binary content assessment one thus distinguishes answers which are appropriate from those which are inappropriate in terms of meaning, independent of whether the answers are grammatically well-formed or not.

From the collected data, we selected an even distribution of unique appropriate and inappropriate student answers in order to obtain a 50% random baseline for our system. Table 1 lists how many questions, target answers and student answers each of the two data sets contains. The data used for this paper is made freely available upon request under a standard Creative Commons by-nc-sa licence.[1]

	KU data set	OSU data set
Target Answers	136	87
Questions	117	60
Student Answers	**610**	**422**
# of Students	141	175
avg. Token #	9.71	15.00

Table 1: The reading comprehension data sets used

3 Approach

Our work builds on the English content assessment approach of Bailey and Meurers (2008), who propose a Content Assessment Module (CAM) which automatically compares student answers to target responses specified by foreign language teachers. As a first step we reimplemented this approach for English in a system we called CoMiC (Comparing Meaning in Context) which is discussed in Meurers et al. (2011). This reimplementation was then adapted for German, resulting in the CoMiC-DE system presented in this paper.

The comparison of student answers and target answer is based on an alignment of tokens, chunks, and

[1] http://creativecommons.org/licenses/by-nc-sa/3.0/

dependency triples between the student and the target answer at different levels of abstraction. Figure 1 shows a simple example including token-level and chunk-level alignments between the target answer (TA) and the student answer (SA).

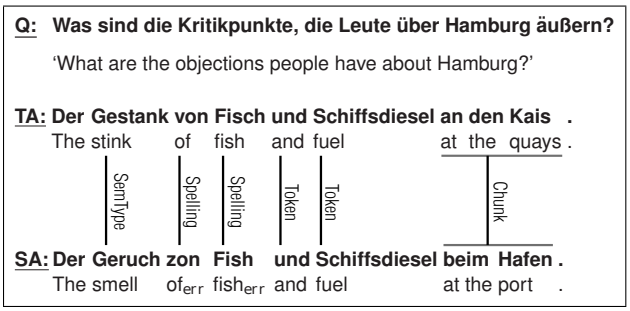

Figure 1: Basic example for alignment approach

As the example suggests, it is not sufficient to align only identical surface forms given that significant lexical and syntactic variation occurs in typical student answers. Alignment thus is supported at different levels of abstraction. For example, the token units are enriched with lemma and synonym information using standard NLP tools. Table 2 gives an overview of which NLP tools we use for which task in CoMiC-DE. In general, the components are very similar to those used in the English system, with different statistical models and parameters where necessary.

Annotation Task	**NLP Component**
Sentence Detection	OpenNLP http://incubator.apache.org/opennlp
Tokenization	OpenNLP
Lemmatization	TreeTagger (Schmid, 1994)
Spell Checking	Edit distance (Levenshtein, 1966), igerman98 word list http://www.j3e.de/ispell/igerman98
Part-of-speech Tagging	TreeTagger (Schmid, 1994)
Noun Phrase Chunking	OpenNLP
Lexical Relations	GermaNet (Hamp and Feldweg, 1997)
Similarity Scores	PMI-IR (Turney, 2001)
Dependency Relations	MaltParser (Nivre et al., 2007)

Table 2: NLP tools used in the German system

Integrating the multitude of units and their representations at different levels of abstraction poses significant challenges to the system architecture. Among other requirements, different representations of the same surface string need to be stored without interfering with each other, and various NLP tools need to collaborate in order to produce the final rich data structures used for answer comparison. To meet these requirements, we chose to implement our system in the Unstructured Information Management Architecture (UIMA, cf. Ferrucci and Lally, 2004). UIMA allows automatic analysis modules to access layers of stand-off annotation, and hence allows for the coexistence of both independent and interdependent annotations, unlike traditional pipeline-style architectures, where the output of each component replaces its input. The use of UIMA in recent successful large-scale projects such as DeepQA (Ferrucci et al., 2010) confirms that UIMA is a good candidate for complex language processing tasks where integration of various representations is required.

In order to determine the global alignment configuration, all local alignment options are computed for every mappable unit. These local candidates are then used as input for the Traditional Marriage Algorithm (Gale and Shapley, 1962) which computes a global alignment solution where each mappable unit is aligned to at most one unit in the other response, such as the one we saw in Figure 1.

On the basis of the resulting global alignment configuration, the system performs the binary content assessment by evaluating whether the meaning of the learner and the target answer are sufficiently similar. For this purpose, it extracts features which encode the numbers and types of alignment and feeds them to the memory-based classifier TiMBL (Daelemans et al., 2007). The features used are listed in Table 3.

Features	**Description**
1. Keyword Overlap	Percent of keywords aligned (relative to target)
2./3. Token Overlap	Percent of aligned target/learner tokens
4./5. Chunk Overlap	Percent of aligned target/learner chunks
6./7. Triple Overlap	Percent of aligned target/learner triples
8. Token Match	Percent of token alignments that were token-identical
9. Similarity Match	Percent of token alignments that were similarity-resolved
10. Type Match	Percent of token alignments that were type-resolved
11. Lemma Match	Percent of token alignments that were lemma-resolved
12. Synonym Match	Percent of token alignments that were synonym-resolved
13. Variety of Match (0-5)	Number of kinds of token-level alignments

Table 3: Features used for the memory-based classifier

4 Content Assessment Experiment

4.1 Setup

We ran our content assessment experiment using the two data sets introduced in section 2, one from Kansas University and the other from The Ohio State University. Both of these contain only records where both annotators agreed on the binary assessment (appropriate/inappropriate meaning). Each set is balanced, i.e., they contain the same number of appropriate and inappropriate student answers.

In training and testing the TiMBL-based classifier, we followed the methodology of Bailey (2008, p. 240), where seven classifiers are trained using the different available distance metrics (Overlap, Levenshtein, Numeric Overlap, Modified value difference, Jeffrey divergence, Dot product, Cosine). Training and testing was performed using the *leave-one-out* scheme (Weiss and Kulikowski, 1991) and for each item the output of the seven classifiers was combined via majority voting.

4.2 Results

The classification accuracy for both data sets is summarized in Table 4. We report accuracy and the total number of answers for each data set.

	KU data set	OSU data set
# of answers	610	422
Accuracy	**84.6%**	**84.6%**

Table 4: Classification accuracy for the two data sets

The 84.6% accuracy figure obtained for both data sets shows that CoMiC-DE is quite successful in performing content assessment for the German data collected so far, a result which is competitive with the one for English obtained by Bailey and Meurers (2008), who report an accuracy of 78% for the binary assessment task on a balanced English data set.

A remarkable feature is the identity of the scores for the two data sets, considering that the data was collected at different universities from different students in different classes run by different teachers. Moreover, there was no overlap in exercise material between the two data sets. This indicates that there is some characteristic uniformity of the learner responses in authentic reading comprehension tasks, suggesting that the course setting and task type effectively constrains the degree of syntactic and lexical variation in the student answers. This includes the stage of the learners in this foreign language teaching setting, which limits their exposure to linguistic constructions, as well as the presence of explicit reading texts that the questions are about, which may lead learners to use the lexical material provided instead of rephrasing content in other words. We intend to explore these issues in our future work to obtain a more explicit picture of the contextual and task properties involved.

Another aspect which should be kept in mind is that the scores we obtained are based on a data set for which the two human annotators had agreed on their assessment. We expect automatic classification results to degrade given more controversial data about which human annotators disagree, especially since such data will presumably contain more ambiguous cues, giving rise to multiple interpretations.

4.3 Evaluation by question type

The overall results include many different question types which pose different kinds of challenges to our system. To develop an understanding of those challenges, we performed a more fine-grained evaluation by question types. To distinguish relevant subcases, we applied the question classification scheme introduced by Day and Park (2005). This scheme is more suitable here than other common answer-typing schemata such as the one in Li and Roth (2002), which tend to focus on questions asking for factual knowledge.

Day and Park (2005) distinguish five different question forms: yes/no (question to be answered with either yes or no), alternative (two or more yes/no questions connected with or), true or false (a statement to be classified as true or false), *who/what/when/where/how/why* (*wh*-question containing the respective question word), and multiple choice (choice between several answers presented with a question, of any other question type). In addition, they introduce a second dimension distinguishing the types of comprehension involved, i.e., how the information asked for by the question can be obtained from the text: literal (questions that can be answered directly and explicitly from the text), reorganization (questions where information from various

parts of the text must be combined), inference (questions where literal information and world knowledge must be combined), prediction (prediction of how a story might continue), evaluation (comprehensive judgement about aspects of the text) and personal response (personal opinion or feelings about the text or the subject).

Out of the five different forms of question, our data contains questions of all forms except for the multiple choice category and the true or false category given that we are explicitly targeting free text responses. To obtain a more detailed picture of the *wh*-question category, we decided to split that category into its respective *wh*-words and added one more category to it, for *which*. Also, we added the type "several" for questions which contain more than one question presented to the student at a time. Of the six comprehension types, our data contained literal, reorganization and inference questions.

Table 5 reports the accuracy results by question forms and comprehension types for the combined OSU and KU data set. The counts encode the number of student answers for which accuracy is reported (micro-averages). The numbers in brackets specify the number of distinct questions and the corresponding accuracy measures are computed by grouping answers by their question (macro-averages). Comparing answer-based (micro-average) accuracy with question-based (macro-average) accuracy allows us to see whether the results for questions with a high number of answers outweigh questions with a small number of answers. In general the micro- and macro-averages reported are very similar and the overall accuracy is the same (84.6%). Overall, the results thus do not seem to be biased towards a specific, frequently answered question instance. Where larger differences between micro- and macro-averages do arise, as for alternative, *when*, and *where* questions, these are cases with few overall instances in the data set, cautioning us against overinterpreting results for such small subsets. The 4.2% gap for the relatively frequent "several" question type underlines the heterogeneous nature of this class, which may warrant more specific subclasses in the future.

Overall, the accuracy of content assessment for *wh*-questions that can be answered with a concrete piece of information from the text are highest, with 92.6% for "which" questions, and results in the upper 80s for five other *wh*-questions. Interestingly, "who" questions fare comparatively badly, pointing to a relatively high variability in the expression of subjects, which would warrant the integration of a dedicated approach to coreference resolution. Such a direct solution is not available for "why" questions, which at 79.3% is the worst *wh*-question type. The high variability of those answers is rooted in the fact that they ask for a cause or reason, which can be expressed in a multitude of ways, especially for comprehension types involving inferences or reorganization of the information given in the text.

This drop between comprehension types, from literal (86.0%) to inference (81.5%) and reorganization (78.0%), can also be observed throughout and is expected given that the CoMiC-DE system makes use of surface-based alignments where it can find them. For the system to improve on the non-literal comprehension types, features encoding a richer set of abstractions (e.g., to capture distributional similarity at the chunk level or global linguistic phenomena such as negation) need to be introduced.

Just as in the discussion of the micro- and macro-averages above, the "several" question type again rears its ugly heads in terms of a low overall accuracy (77.7%). This supports the conclusion that it requires a dedicated approach. Based on an analysis of the nature and sequence of the component questions, in future work we plan to determine how such combinations constrain the space of variation in acceptable answers.

Finally, while there are few instances for the "alternative" question type, the fact that it resulted in the lowest accuracy (57.1%) warrants some attention. The analysis indeed revealed a general issue, which is discussed in the next section.

5 From eliminating repeated elements to analyzing information structure

Bailey (2008, sec. 5.3.12) observed that answers frequently repeat words given in the question. In her corpus example (1), the first answer repeats "the moral question raised by the Clinton incident" from the question, whereas the second one reformulates this given material. But both sentences essentially answer the question in the same way.[2]

[2]Independent of the issue discussed here, note the presuppo-

	Comprehension type							
	Literal		Reorganization		Inference		Total	
Question type	Acc.	#	Acc.	#	Acc.	#	Acc.	#
Alternative	0	1 (1)	–	0	66.7 (58.3)	6 (3)	57.1 (43.8)	7 (4)
How	85.7 (83.3)	126 (25)	83.3 (77.8)	12 (3)	100	7 (1)	86.2 (83.3)	145 (29)
What	87.0 (87.6)	247 (40)	74.2 (71.7)	31 (4)	83.3 (83.3)	6 (1)	85.6 (86.1)	284 (45)
When	85.7 (93.3)	7 (3)	–	0	–	0	85.7 (93.3)	7 (3)
Where	88.9 (94.4)	9 (3)	–	0	–	0	88.9 (94.4)	9 (3)
Which	92.3 (90.7)	183 (29)	100.0	14 (5)	83.3 (83.3)	6 (2)	92.6 (91.6)	203 (36)
Who	73.9 (80.2)	23 (9)	94.4 (88.9)	18 (3)	–	0	82.9 (82.4)	41 (12)
Why	80.5 (83.3)	128 (17)	57.1 (57.9)	14 (3)	84.4 (81.1)	32 (4)	79.3 (79.7)	174 (24)
Yes/No	–	0	100.0	5 (1)	–	0	100.0	5 (1)
Several	82.1 (85.6)	95 (13)	68.4 (75.1)	38 (5)	75 (74.3)	24 (2)	77.7 (81.9)	157 (20)
Total	86.0 (86)	819 (140)	78.0 (80.7)	132 (24)	81.5 (76.8)	81 (13)	84.6 (84.6)	1032 (177)

Table 5: Accuracy by question form and comprehension types following Day and Park (2005). Counts denoting number of student answers, in brackets: number of questions and macro-average accuracy computed by grouping by questions.

(1) What was the major moral question raised by the Clinton incident?

   a. The moral question raised by the Clinton incident was whether a politician's person life is relevant to their job performance.

   b. A basic question for the media is whether a politician's personal life is relevant to his or her performance in the job.

The issue arising from the occurrence of such given material for a content assessment approach based on alignment is that all alignments are counted, yet those for given material do not actually contribute to answering the question, as illustrated by the (non)answer containing only given material "The moral question raised by the Clinton incident was whatever." Bailey (2008) concludes that an answer should not be rewarded (or punished) for repeating material that is given in the question and her implementation thus removes all words from the answers which are given in the question.

While such an approach successfully eliminates any contribution from these given words, it has the unfortunate consequence that any NLP processes requiring well-formed complete sentences (such as, e.g., dependency parsers) perform poorly on sentences from which the given words have been removed. In our reimplementation of the approach, we therefore kept the sentences as such intact and instead made use of the UIMA architecture to add a givenness annotation to those words of the answer which are repeated from the question. Such given tokens and any representations derived from them are ignored when the local alignment possibilities are computed.

While successfully replicating the givenness filter of Bailey (2008) without the negative consequences on other NLP analysis, targeting given words in this way is problematic, which becomes particularly apparent when considering examples for the "alternative" question type. In this question type, exemplified in Figure 2 by an example from the KU data set, the answer has to select one of the options from an explicitly given set of alternatives.

Q: Ist die Wohnung in einem Neubau oder einem Altbau?
'Is the flat in a new building or in an old building?'

TA: Die Wohnung ist in einem Neubau .
The flat is in a new building

SA: Die Wohnung ist in einem Neubau
The flat is in a new building

Figure 2: "Alternative" question with answers consisting entirely of given words, resulting in no alignments.

The question asks whether the apartment is in a new or in an old building, and both alternatives are explicitly given in the question. The student picked the same alternative as the one that was selected in the target answer. Indeed, the two answers are identical, but the givenness filter excludes all material from alignment and hence the content assessment

sition failure arising for this authentic reading comprehension question – as far as we see, there was no "major moral question raised by the Clinton incident".

classification fails to identify the student answer as appropriate. This clearly is incorrect and essentially constitutes an opportunity to rethink the givenness filter.

The givenness filter is based on a characterization of the material we want to ignore, which was motivated by the fact that it is easy to identify the material that is repeated from the question. On the other hand, if we analyze the reading comprehension questions more closely, it becomes possible to connect this issue to research in formal pragmatics which investigates the information structure (cf. Krifka, 2007) imposed on a sentence in a discourse addressing an explicit (or implicit) question under discussion (Roberts, 1996). Instead of removing given elements from an answer, under this perspective we want to identify which part of an answer constitutes the so-called focus answering the question.[3]

The advantage of linking our issue to the more general investigation of information structure in linguistics is readily apparent if we consider the significant complexity involved (cf., e.g., Büring, 2007). The issue of asking what constitutes the focus of a sentence is distinct from asking what new information is included in a sentence. New information can be contained in the topic of a sentence. On the other hand, the focus can also contain given information. In (2a), for example, the focus of the answer is "a green apple", even though apples are explicitly given in the question and only the fact that a green one will be bought is new.

(2) You've looked at the apples long enough now, what do you want to buy?

a. I want to buy a green apple.

In some situations the focus can even consist entirely of given information. This is one way of interpreting what goes on in the case of the alternative questions discussed at the end of the last section. This question type explicitly mentions all alternatives as part of the question, so that the focus of the answer selecting one of those alternatives will typically consist entirely of given information.

As a next step we plan to build on the notion of focus characterized in (a coherent subset of) the information structure literature by developing an approach which identifies the part of an answer which constitutes the focus so that we can limit the alignment procedure on which content assessment is based to the focus of each answer.

6 Related Work

There are few systems targeting the short answer evaluation tasks. Most prominent among them is *C-Rater* (Leacock and Chodorow, 2003), a short answer scoring system for English meant for deployment in Intelligent Tutoring Systems (ITS). The authors highlight the fact that *C-Rater* is not simply a string matching program but instead uses more sophisticated NLP such as shallow parsing and synonym matching. *C-Rater* reportedly achieved an accuracy of 84% in two different studies, which is remarkably similar to the scores we report in this paper although clearly the setting and target language differ from ours.

More recently in the ITS field, Nielsen et al. (2009) developed an approach focusing on recognizing textual entailment in student answers. To that end, a corpus of questions and answers was manually annotated with word-word relations, so-called "facets", which represent individual semantic propositions in a particular answer. By learning how to recognize and classify these facets in student answers, the system is then able to give a more differentiated rating of a student answer than "right" or "wrong". We find that this is a promising move in the fields of answer scoring and textual entailment since it also breaks down the complex entailment problem into a set of sub-problems.

7 Conclusion

We presented CoMiC-DE, the first content assessment system for German. For the data used in evaluation so far, CoMiC-DE performs on a competitive level when compared to previous work on English, with accuracy at 84.6%. In addition to these results, we make our reading comprehension corpus freely available for research purposes in order to encourage more work on content assessment and related areas.

In a more detailed evaluation by question and com-

[3] The information structure literature naturally also provides a more sophisticated account of givenness. For example, for Schwarzschild (1999), givenness also occurs between hypernyms and coreferent expressions, which would not be detected by the simple surface-based givenness filter included in the current CoMiC-DE.

prehension type, we gained new insights into how question types influence the content assessment tasks. Specifically, our system had more difficulty classifying answers to "why"-questions than other question forms, which we attribute to the fact that causal relations exhibit more form variation than other types of answer material. Also, the comprehension type "reorganization", which requires the reader to collect and combine information from different places in the text, posed more problems to our system than the "literal" type.

Related to the properties of questions, we showed by example that simply marking given material on a surface level is insufficient and a partitioning into focused and background material is needed instead. This is especially relevant for alternative questions, where the exclusion of all given material renders the alignment process useless. Future work will therefore include focus detection in answers and its use in the alignment process. For example, given a weighting scheme for individual alignments, focused material could be weighted more prominently in alignment in order to reflect its importance in assessing the answer.

Acknowledgements

We would like to thank two anonymous TextInfer reviewers for their helpful comments.

References

Stacey Bailey, 2008. Content Assessment in Intelligent Computer-Aided Language Learning: Meaning Error Diagnosis for English as a Second Language. Ph.D. thesis, The Ohio State University. http://osu.worldcat.org/oclc/243467551.

Stacey Bailey and Detmar 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. http://aclweb.org/anthology/W08-0913.

Daniel Büring, 2007. Intonation, Semantics and Information Structure. In Gillian Ramchand and Charles Reiss (eds.), *The Oxford Handbook of Linguistic Interfaces*, Oxford University Press.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot and Antal van den Bosch, 2007. *TiMBL: Tilburg Memory-Based Learner Reference Guide, ILK Technical Report ILK 07-03. Version 6.0*. Tilburg University.

Ido Dagan, Bill Dolan, Bernardo Magnini and Dan Roth, 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii.

Richard R. Day and Jeong-Suk Park, 2005. Developing Reading Comprehension Questions. *Reading in a Foreign Language*, 17(1):60–73.

David Ferrucci, Eric Brown et al., 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3):59–79.

David Ferrucci and Adam Lally, 2004. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10(3–4):327–348.

David Gale and Lloyd S. Shapley, 1962. College Admissions and the Stability of Marriage. *American Mathematical Monthly*, 69:9–15.

Birgit Hamp and Helmut Feldweg, 1997. GermaNet – a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid. http://aclweb.org/anthology/W97-0802.

Manfred Krifka, 2007. Basic Notions of Information Structure. In Caroline Fery, Gisbert Fanselow and Manfred Krifka (eds.), *The Notions of Information Structure*, Universitätsverlag Potsdam, Potsdam, volume 6 of *Interdisciplinary Studies on Information Structure (ISIS)*.

Thomas Landauer, Peter Foltz and Darrell Laham, 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284.

Claudia Leacock and Martin Chodorow, 2003. C-rater: Automated Scoring of Short-Answer Questions. *Computers and the Humanities*, 37:389–405.

Vladimir I. Levenshtein, 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10(8):707–710.

Xin Li and Dan Roth, 2002. Learning Question Classifiers. In *Proceedings of the 19th International*

Conference on Computational Linguistics (COLING 2002). Taipei, Taiwan, pp. 1–7.

Detmar Meurers, Niels Ott and Ramon Ziai, 2010. Compiling a Task-Based Corpus for the Analysis of Learner Language in Context. In *Proceedings of Linguistic Evidence*. Tübingen, pp. 214–217. http://purl.org/dm/papers/meurers-ott-ziai-10.html.

Detmar Meurers, Ramon Ziai, Niels Ott and Stacey Bailey, 2011. Integrating Parallel Analysis Modules to Evaluate the Meaning of Answers to Reading Comprehension Questions. *IJCEELL. Special Issue on Automatic Free-text Evaluation*, 21(4):355–369. http://purl.org/dm/papers/meurers-ziai-ott-bailey-11.html.

Rodney D. Nielsen, Wayne Ward and James H. Martin, 2009. Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering*, 15(4):479–501.

Joakim Nivre, Jens Nilsson, Johan Hall, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov and Erwin Marsi, 2007. MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. *Natural Language Engineering*, 13(1):1–41.

Craige Roberts, 1996. Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics. In Jae-Hak Yoon and Andreas Kathol (eds.), *OSU Working Papers in Linguistics No. 49: Papers in Semantics*, The Ohio State University.

Álvaro Rodrigo, Anselmo Peñas and Felisa Verdejo, 2009. Overview of the Answer Validation Exercise 2008. In Carol Peters, Thomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas and Vivien Petras (eds.), *Evaluating Systems for Multilingual and Multimodal Information Access*, Springer Berlin / Heidelberg, volume 5706 of *Lecture Notes in Computer Science*, pp. 296–313.

Helmut Schmid, 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK, pp. 44–49.

Roger Schwarzschild, 1999. GIVENness, AvoidF and other Constraints on the Placement of Accent. *Natural Language Semantics*, 7(2):141–177.

Karen Spärck Jones, 2007. Automatic Summarising: The State of the Art. *Information Processing and Management*, 43:1449–1481.

Peter Turney, 2001. Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*. Freiburg, Germany, pp. 491–502.

Sholom M. Weiss and Casimir A. Kulikowski, 1991. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann, San Mateo, CA.