Integrating Parallel Analysis Modules to Evaluate the Meaning of Answers to Reading Comprehension Questions

Detmar Meurers
Seminar für Sprachwissenschaft
Universität Tübingen
Wilhelmstraße 19
72072 Tübingen, Germany
dm@sfs.uni-tuebingen.de

Ramon Ziai, Niels Ott
Sonderforschungsbereich 833
Universität Tübingen
Nauklerstraße 35
72072 Tübingen, Germany
{rziai,nott}@sfs.uni-tuebingen.de

Stacey Bailey
Department of Linguistics
The Ohio State University
1712 Neil Avenue
Columbus, Ohio 43210, USA
stacey.m.bailey@gmail.com

Abstract:
Contextualized, meaning-based interaction in the foreign language is widely recognized as crucial for second language acquisition. Correspondingly, current exercises in foreign language teaching generally require students to manipulate both form and meaning. For Intelligent Language Tutoring Systems to support such activities, they thus must be able to evaluate the appropriateness of the meaning of a learner response for a given exercise.

We discuss such a content-assessment approach, focusing on reading comprehension exercises. We pursue the idea that a range of simultaneously available representations at different levels of complexity and linguistic abstraction provide a good empirical basis for content assessment. We show how an annotation-based NLP architecture implementing this idea can be realized and that it successfully performs on a corpus of authentic learner answers to reading comprehension questions. To support comparison and sustainable development on content assessment, we also define a general exchange format for such exercise data.

Keywords: content assessment, shallow semantic analysis, meaning comparison, textual entailment, intelligent computer-assisted language learning, ICALL, intelligent tutoring systems

Biographical notes: Detmar Meurers is a professor of Computational Linguistics at the University of Tübingen, Germany. Previously he was an associate professor at The Ohio State University, where he founded the ICALL research group focusing on intelligent tutoring systems, content assessment, and automatic input enhancement for language learners.

Ramon Ziai is a PhD candidate at the Collaborative Research Center 833 at the University of Tübingen, Germany. His main research interest and background is in computational linguistics. More specifically, he is interested in shallow semantic analysis and the question of how ill-formed input can be processed.

Niels Ott holds a BA and an MA in computational linguistics from the University of Tübingen, where he is a PhD candidate at the Collaborative Research Center 833. His main research interest are shallow semantic analysis, robust language processing, and text difficulty measures.

Stacey Bailey completed her PhD in 2008 on automated content assessment at The Ohio State University, where she was a member of the ICALL research group. She now works as a Senior Artificial Intelligence Engineer at The MITRE Corp.

1 Motivation

Research in second language acquisition and foreign language teaching and learning has established that contextualized, meaning-based interaction in the foreign language is a crucial component for successful second language acquisition (cf., e.g., Ellis, 2005). Correspondingly, exercises in current foreign language teaching generally require students to manipulate both form and meaning as, for example, is the case for reading and listening comprehension, summarization, or information gap activities. For Intelligent Language Tutoring Systems to provide feedback for such activities, it thus becomes crucial for such systems to go beyond the traditional form-focused analysis towards an evaluation that includes the meaning of a learner response for a given exercise.

In this article, we discuss such a content-assessment approach, focusing on answers to reading comprehension questions. Building on Bailey (2008) and Bailey & Meurers (2008), we further pursue the idea that a range of simultaneously available representations at different levels of complexity and linguistic abstraction constitute a valuable empirical basis for content assessment. We first describe the original approach (section 2), for which questions of the processing architecture and explicit data structures had not been a focus. We then motivate and describe our new, annotation-based NLP architecture for content assessment based on the UIMA framework and discuss how we used it to reimplement the approach (section 3). Evaluating the approach on a corpus of authentic learner answers to reading comprehension questions, we confirm that the approach successfully performs content assessment for real-life exercises (section 3.2).

To support comparison and sustainable development on content assessment, we also define a general exchange format for reading comprehension data and make the corpus available in this form (section 4). We conclude with a characterization of several research issues which we believe to be important for future development (section 5) such as a better integration of context information, refined diagnosis categories for meaning comparison, and improved adaptivity of analysis combining language processing strategies from shallow to deeper analysis.

2 Background: Content Assessment for Reading Comprehension

Our approach focuses on the evaluation of answers to reading comprehension questions. This kind of task has several properties that make it interesting for automatic content evaluation. First, it is a common, real-life activity in foreign language classrooms which means that developing a content assessment approach for such a task is of practical relevance and authentic learner data together with independent gold standard assessment by teachers is in principle available to develop and test an approach.

Second, student answers to reading comprehension questions can exhibit significant variation on lexical, morphological, syntactic and semantic levels so that performing content assessment by relying on simple string comparison of learner answers to a list of pre-stored answers is not a realistic option.

And third, it is possible to focus on the language-related aspects of content assessment by selecting reading comprehension questions which target information represented in a given text (as opposed to asking about world knowledge or personal experience relating to the text). For the type of reading comprehension questions we are focusing on it is possible for the teacher to specify target answers to which student answers are compared. Figure 1 shows an example reading comprehension exercise from the corpus collected by Bailey (2008).

QUESTION: *What are the methods of propaganda mentioned in the article?*

TARGET ANSWER: *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.*

STUDENT ANSWERS:

1. *A number of methods of propaganda are used in the media.*
⇒ Binary assessment: incorrect meaning
⇒ Detailed assessment: missing concept

2. *Bositive or negative labels.*
⇒ Binary assessment: incorrect meaning
⇒ Detailed assessment: missing concept

3. *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.*
⇒ Binary assessment: correct meaning
⇒ Detailed assessment: correct

Figure 1: Example from the English corpus collected by Bailey (2008)

The responses in this corpus were written by intermediate ESL students as part of their regular homework assignments. The students had access to their textbooks for all activities. The target answers were provided by the teachers, and two independent graders assessed the meaning of the student responses in relation to the target answers. The student answers were labelled with a binary assessment code (correct meaning vs. incorrect meaning) and a more detailed diagnosis (correct, missing concept, extra concept, blend, non-answer, alternate answer).

In order for content assessment to be able to deal with the significant variation in form between the target and the student answers, the Content Assessment Module (CAM) of Bailey & Meurers (2008) makes use of alignments between the student and target answers at different levels and using different types of linguistic abstraction. Figure 2 illustrates the basic idea. The different types of linguistic abstraction which are represented in parallel for each target and learner answer are illustrated in

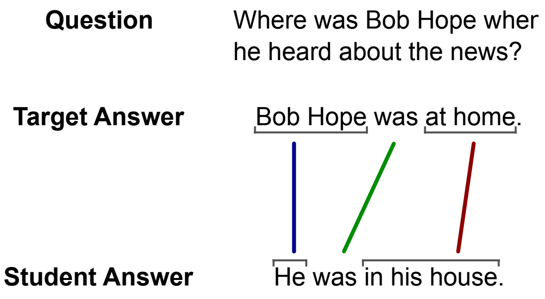

Figure 2: Basic alignment approach using token-level and chunk-level matching

Figure 3 and Figure 4 illustrates the different levels of complexity which are simultaneously available for establishing the alignment. The latter figure shows examples for cases where the local domain captured by a chunk or the functor-argument relation established by a dependency triple are needed to support better mappings than would result from the token level alone.

Alignment Type	**Example Match**
token-identical	*advertising — advertising*
lemma-resolved	*advertisement — advertising*
spelling-resolved	*campaing — campaign*
reference-resolved	*Clinton — he*
semantic similarity-resolved	*Initial — beginning*
specialized expressions	*May 24, 2007 — 5/24/2007*

Figure 3: Types of Alignment

Level	**Example**	**Alignment**
Tokens	*The explanation is simple.* *The reason is simple.*	*explanation* *reason*
Chunks	*A brown dog sat in a nice car.* *A nice dog sat in a car.*	*a brown dog* *a nice dog*

Dependency triples	*He knows the doctor.*
	John knows him.
		obj(knows, doctor)
		obj(knows, him)

Figure 4: Levels of Alignment

The general alignment-based approach is also pursued in several other application domains, such as automatic grading (e.g., Leacock, 2004; Pérez Marin, 2007), paraphrase recognition (e.g., Brockett & Dolan, 2005; Hatzivassiloglou et al., 1999), or recognition of textual entailment (RTE, e.g., Dagan et al., 2009). Particularly interesting for our discussion here are approaches in machine translation evaluation such as the METEOR metric (Banerjee & Lavie, 2005), which also make use of more abstract representations of tokens. There is an important difference, though, which is directly relevant under the NLP architecture perspective of this paper. In the original version of METEOR, abstract linguistic levels of representation are only considered in case an alignment cannot be found based on the surface-based token representation. In contrast, our CAM approach always bases content classification on a parallel representation at all levels (token, chunk, dependency) and all types of abstraction. Interestingly, the newer METEOR-next approach (Denkowski & Lavie, 2010) pursues a similar strategy, which further highlights the importance of parallel representations.
Based on this rich empirical basis of possible alignments, CAM selects a globally successful alignment configuration. It then extracts features based on the number and nature of the alignments and uses this evidence for a memory-based machine learner (TiMBL, see Daelemans et al. 2007). The full list of features used is given in Figure 5.

Summing up, the overall CAM approach consists of three phases:

1. **Annotation** uses NLP to enrich the student and target answers, as well as the question text, with linguistic information on different levels and types of abstraction.
2. **Alignment** maps elements of the learner answer to elements of the target response using the annotated information.
3. **Classification** analyzes the possible alignments and labels the learner response with a binary content assessment and a detailed diagnosis code.

Features	Description
1. Keyword Overlap	Percent of keywords aligned (relative to target)
2. Target Overlap	Percent of aligned target tokens
3. Learner Overlap	Percent of aligned learner tokens
4. Target Chunk	Percent of aligned target chunks
5. Learner Chunk	Percent of aligned learner chunks
6. Target Triple	Percent of aligned target triples
7. Learner Triple	Percent of aligned 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

Figure 5: Features used for machine learning of content assessment classification

3 An Annotation-based NLP Architecture for Content Assessment

3.1 Architecture Requirements and Solutions

The CAM approach sketched in the previous section provides a good starting point as far as the empirical and conceptual basis is concerned. But given its nature as a pilot study into content assessment, we did not focus on the NLP architecture and data structure choices. In order to push this strand of research further, on the practical side questions arise on how such an approach is best realized in a general NLP architecture. On the one hand, it should support modular experimentation and development of content assessment approaches such as for our current research on a content assessment prototype for German. It should also facilitate integration into current architectures motivated for ICALL system such as TAGARELA (Amaral, Meurers & Ziai, 2011). On the theoretical side, a number of research issues present themselves, such as an investigation of the role of the context and information structure on content assessment and a more dynamic integration of different levels of linguistic representation, which would also benefit from a general and flexible NLP architecture and explicit data structures considerations. For these practical and theoretical reasons, we pursue an architecture satisfying the following requirements:

- **Representations and alignment:** CAM only aligns tokens to tokens, chunks to chunks, etc. However, in general the same meaning can in principle be expressed by linguistic units of different complexity and type, e.g., the token *initially* could be aligned to chunk *in the beginning*. Thus, alignments between different representations should be more fully supported.
- **Marking contextual relevance of material:** Some parts of the student and target answer, such as material already given in the question (which we return to in section 5) or punctuation, should not be taken into account when doing a semantic comparison. The original CAM simply deleted such material from the answers, destroying syntactic structures and leaving the answers incoherent. A mechanism is needed which excludes the relevant units from alignment but otherwise leaves the answers intact.
- **Explicitness of data structures and modularity of analyses:** As it is not clear from the start which NLP tool will perform best for a given task, we need a way to make explicit the data structures we want to work with regardless of which particular tool will provide them. Moreover, new analysis components should be straightforward to add without interfering with the ones already present in the system.

On the basis of these requirements, we chose the Unstructured Information Management Architecture (UIMA, see Ferrucci & Lally 2004) as the basis for our new system architecture, CoMiC (Comparing Meaning in Context). As a framework meant for complex NLP applications, UIMA not only supports but enforces the idea of annotation-based processing. Using so-called referential annotation, information on the text is added throughout processing but the text itself is never changed. The repository for such accumulated information is the Common Analysis System (CAS, see Götz & Suhre 2004) which basically provides annotation indexes over the text. Annotations have to be explicitly declared in order to be put into such indexes; for example, to annotate tokens one must first define a type *Token*. Such types can be associated with features, or attributes, which can again be of any simple (string, integer, etc.) or complex type. Through the type systems, UIMA achieves an abstraction between the analysis results and the NLP tools that provide them. The type system is declared as meta-data outside of the programming language.

In CoMiC, each NLP tool we use (see Figure 6) is encapsulated as a UIMA Annotator that contributes a specific analysis result to the CAS. Figure 7 shows the overall CoMiC architecture. A UIMA Collection Reader takes care of reading in the corpus data and setting up the initial CAS before it is enriched with annotations. While such a variety of parallel analysis results would pose problems for most file-based annotation formats, they are not problematic for UIMA, because each type of annotation is put into a separate index and hence integrates well with other results. Before alignment takes place, givenness and punctuation filters take care of marking material that is not to be included in alignment. Thanks to the explicit data structures, this can simply be done by setting a Boolean feature on the type *Token* to a certain value. Alignment modules can then check this value and exclude unwanted material.

Annotation	**original CAM**	**CoMiC-EN**
Sentence Detection	MontyLingua	OpenNLP
Tokenization	MontyLingua	OpenNLP
Lemmatization	MontyLingua PC-KIMMO	morpha
Spell Checking	Edit distance, SCOWL word list	same
Part-of-speech Tagging	TreeTagger	same
Noun Phrase Chunking	CASS	OpenNLP
Lexical Relations	WordNet	same
Similarity Scores	PMI-IR	same
Dependency Relations	Stanford Parser	MaltParser

Figure 6: NLP tools used in the original CAM and the English CoMiC system.

For the material not excluded, alignment is done on the token, chunk and dependency levels, as in the original CAM. This works by first collecting candidate alignments for each element and then using the Traditional Marriage Algorithm (TMA, see Gale & Shapley 1962) to select the globally optimal alignment configuration. While we do not align tokens with chunks at the moment, we have included this possibility by defining a common supertype for both in the UIMA type system, enabling us to abstract over the two if necessary.

When all alignments have been determined and the TMA has selected the optimal configuration, a UIMA CAS Consumer uses the alignment information in the CAS to extract features for training or calling the classifier, for which we use TiMBL (Daelemans et al., 2007) as in the original CAM. At this point, UIMA-based processing ends and the feature configurations are written to a simple text file that the TiMBL program can read.

3.2 Results

For the purpose of comparing CoMiC-EN to the original CAM approach, we evaluated it against the same original data-set, which is described in section 4.1 in more detail. The memory-based learner TiMBL was trained on the 311 student and target answers from the development set and evaluated against the 255 student and target answers from the test set. We used the following distance measures with TiMBL: Cosine Distance, Dot Product, Weighted Overlap, Levenshtein Distance, Euclidean Distance, Modified Value Difference, Jeffrey Divergence and Numeric Overlap. Instead of relying on any single one of them, the best choice was automatically selected according to a majority voting of the distance measures for each data record.

The results obtained are summarized in Figure 8.

	CAM	**CoMiC-EN**
Development Set		
Binary Classification Detailed Classification	87% 79%	87.6% 78.7%
Test Set		
Binary Classification Detailed Classification	88% –	88.4% 79.0%

Figure 8: Evaluation results of the original CAM and CoMiC-EN

We report two numbers for both the development set and the test set: *Binary Classification* refers to the accuracy achieved in the task of deciding whether a student answer was correct or incorrect. *Detailed Classification* refers to the accuracy in predicting the correct detailed assessment: correct, missing concept, extra concept, blend, or non-answer. Both classification tasks were carried out using the 13 features of Figure 5.

As aimed for, the performance of CoMiC-EN using the new architecture reaches the same high level as the original CAM implementation. There are slight differences, which are to be expected given that, as we saw in Figure 6, different NLP tools were used for five of the nine annotators. But in an architecture making use of such a wide range of parallel representations for the alignments, the specific choice of NLP tools does not seem to be crucial to the performance of the overall approach.

We are not aware of a directly comparable content assessment system for answers to exercises written by language learners. Considering the 85% accuracy reported for a related content assessment task performed by the C-rater system (Leacock, 2004) on answers written by native English speakers suggests that the results of the CoMiC-EN system are competitive with the state of the art. For sustainable progress on short answer content assessment it clearly is important, though, to make results of different approaches more directly comparable. As a step in that direction we are making the CoMiC-EN corpus available. In the next section, we characterize the corpus and define a general format for reading comprehension activities in order to facilitate exchange and comparison of different approaches to this real-life task.

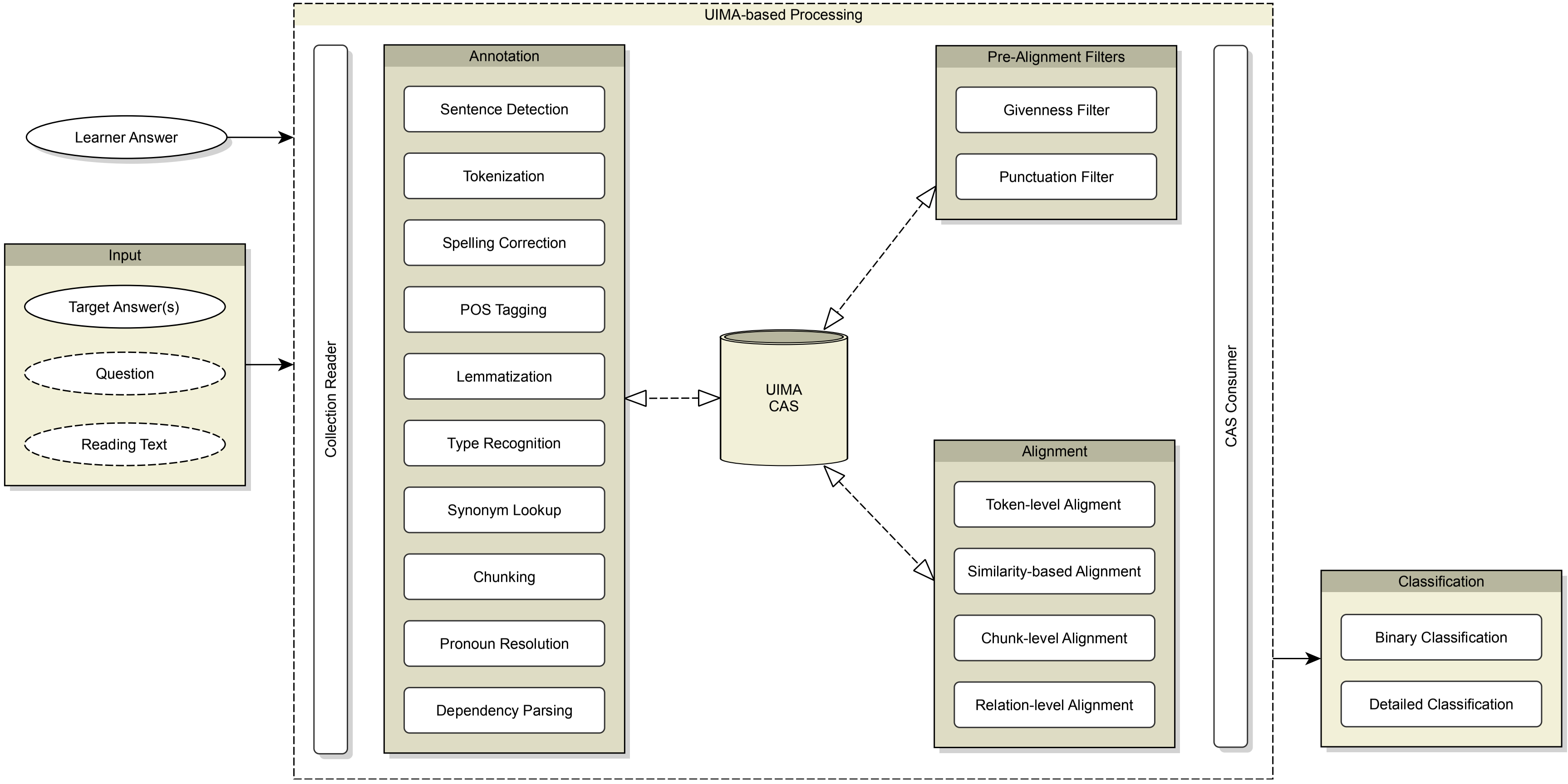

Figure 7: CoMiC architecture overview

4 The Corpus of Reading Comprehension Exercises in English (CREE)

4.1 Data

The English development and test corpus (Bailey & Meurers, 2008) consists of 566 responses by intermediate ESL students to short-answer comprehension questions. The responses were written as part of the regular homework assignments, where students had access to their textbooks, and typically are one to three sentences in length. Collection took place in two different classes at the same level – intermediate reading/writing course offered at The Ohio State University to students who need to improve their English to advance to regular college classes. Each course involved different teachers and students and each teacher created their own exercises, with some of the texts overlapping. The material from the first course was designated the development set, and that from the second course the test set. The development set contains 311 responses from 11 students answering 47 different questions, while the test set contains 255 responses from 15 students to 28 questions.

In order to support the comparison of the CoMiC-EN system with other approaches and architectures, the task and corpus on which the results described above were obtained needs to be accessible. As a step in this direction, we make the original English development and test corpus freely available[1] on request. To support this corpus exchange and obtain an explicit basis on which comparable exercise materials can be collected, we need an explicit data exchange format for such tasks, which we discuss in the next section.

4.2 Exchange Format

The CoMiC corpus exchange format is based on standard XML technology. It is designed to meet the requirements of the CREE corpus as well as those of our ongoing four year corpus creation effort CREG (Corpus of Reading comprehension Exercises in German), in which we are collecting a longitudinal learner corpus consisting of answers to reading comprehension questions written by American college students learning German (Meurers, Ott & Ziai, 2010). The structure of the format is illustrated in Figure 9.

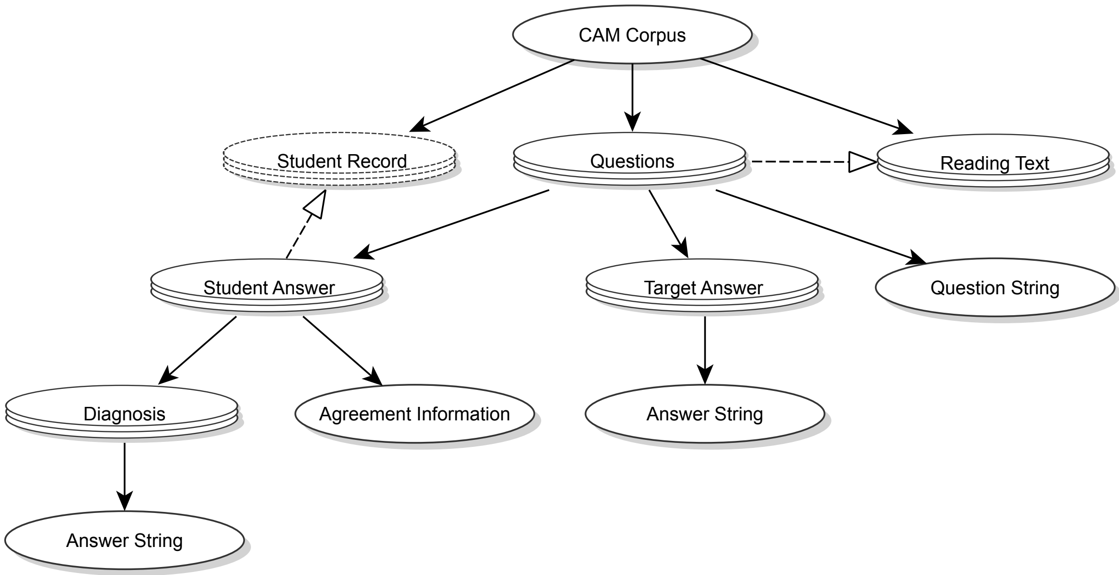

Figure 9: Structure of the CoMiC corpus exchange format

[1] The corpus will be made available under a Creative Commons by-nc-sa license.

Reading comprehension questions are the central element around which data are organized. Apart from the question string, each question contains a link to its corresponding reading text. Each question can be equipped with several target answers. Similarly, several student answers are attached to each question. Each student answer is linked to a student meta data record. Student answers are equipped with multiple diagnoses, each holding the assessment of one annotator. The string of the student answer is also stored in the diagnosis, since copying student answers from (potentially) handwritten submissions is already a step of interpretation. Additionally, each student answer can hold information about the agreement of the annotators. The current version of the format does not yet include the possibility to store records of student meta data, which we are considering for inclusion in a future version.

While the CREG corpus currently being collected makes use of all of these features, the CREE corpus stemming from the original corpus collection effort does not contain multiple target answers. For illustration, an excerpt of the CREE corpus in the XML format is depicted in Figure 10.

```
<?xml version="1.0" encoding="iso8859-1"?>
<?xml-stylesheet type="text/css" href="cam-corpus-web.css"?>
<CAMCorpus xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="http://purl.org/icall/cam-corpus/cam-exchange
-0.1.xsd">
<Texts/>
<Students/>
<Questions>
  <Question id="TU3CH6R32">
    <questionString>What are the methods of propaganda mentioned in
       the article?</questionString>
    <TargetAnswers>
      <TargetAnswer keywords="labels, positive, negative, visual, images,
        beautiful, famous, promote" question_id="TU3CH6R32">
        <answerText>The methods include labels, images, and beautiful or
          famous people promoting the idea or product.</answerText>
      </TargetAnswer>
    </TargetAnswers>
    <StudentAnswers>
      <StudentAnswer id="214" question_id="TU3CH6R32" student_id="SP0713">
        <diagnosis binary="N" detailed="MC" id="214">
          <answerText>One method is giving positive or negative labels to
            control audience supports.</answerText>
        </diagnosis>
      </StudentAnswer>
      <StudentAnswer id="219" question_id="TU3CH6R32" student_id="SP078">
        <diagnosis binary="Y" detailed="CA" id="219">
           <answerText> The methods of propaganda are labels, visual
             images, famous promoters, and to creat the
             impression.</answerText>
        </diagnosis>
      </StudentAnswer>
    </StudentAnswers>
  </Question>
</Questions>
</CAMCorpus>
```

Figure 10: XML example of the CoMiC corpus exchange format (reading texts omitted)

The CREE corpus makes use of the binary and detailed assessment scheme introduced in section 2. Systems other than CoMiC-EN may define and make use of other assessment schemes. Therefore, the diagnosis element with its XML attributes *detailed* and *binary* is likely to be too inflexible for the use across different assessment schemes. One convenient possibility to solve this problem would be to introduce a generic scheme of key-value pairs. Another possibility would be the division of XML namespaces (cf. Harold & Means, 2004, ch. 4). The latter option would be less convenient to implement but it would allow for automatic document validation by standard XML parsers.

5 Relevant issues for future work on content assessment

Building on the English CAM work and the annotation-based processing architecture we discussed in the previous sections, we are exploring several research issues as part of the SFB 833 project A4 “Comparing Meaning in Context: Components of a shallow semantic analysis”. As we consider these issues to be of general relevance for future development in content assessment, we briefly characterize them here.

Towards interpretation in context The Recognizing Textual Entailment task as a well-known generalization of several real-life tasks involving meaning comparison has been pointed out be problematic in lacking a context in which the evaluation takes place (cf., e.g., Manning, 2006). The reading comprehension task we propose to focus on provides an explicit context in form of the text, and the question asked about it. CAM currently takes this context into account for basic anaphora resolution for elements in the target and learner answers. But how about about other aspects of this context? How should information in the answers that in terms of the information structure (cf. Krifka, 2007) is *given* in the question be interpreted?

An example illustrating the issue is shown in Figure 11, where the target and learner responses contain different pieces of information which are *given* in the question. In a sense such material should not be compared when evaluating whether the learner response encodes the same meaning as the target response.

QUESTION: *What **was the** major **moral question raised by the Clinton incident**?*
TARGET ANSWER: ***The moral question raised by the Clinton incident was*** *whether a politician’s person life is relevant to their job performance.*
STUDENT ANSWER: ***A basic question for the media*** *is whether a politician’s personal life is relevant to his or her performance in the job.*

Figure 11: Example highlighting the distribution of *given* information

In the original CAM approach (Bailey & Meurers, 2008), we already mentioned in section 3 that words encoding *given* information were simply removed from the answers before comparing them – which in the ad hoc architecture and the plain text data structures used in the original CAM prototype was the only directly realizable option. Yet this only captures a rather limited notion of *givenness* directly attached to single words, and it destroys the overall structure of the sentences, which is needed for successful deeper linguistic analysis, such as dependency parsing. Furthermore it fails to make use of the *given* information as indicator that an answer actually is on target in answering a specific question – in contrast to the literature in Information Retrieval, which makes use of overlapping, *given* information between a query and a document in exactly this way. In sum, we consider a more comprehensive treatment of *given* information as an important research issue for work on content assessment.

Turning from the information *given* in the question to that requested in the question, it seems important to explore the nature of the questions and which task strategies they require. The targeted reading comprehension questions are similar in terms of the level of expected variation and explicitness of their activity models in that they support target answers. But such questions are not necessarily homogeneous. To tease apart question types that impact processing, we are investigating several features. The *learning goals* of a reading comprehension question differentiate targeted cognitive skills and knowledge (cf., e.g., Anderson & Krathwohl, 2001). With respect to the *knowledge sources*, we can distinguish implicit from explicit answer source (cf., e.g., Irwin, 1986; Pearson & Johnson, 1978). Regarding the *text type*, the rhetorical structure of the text has a clear impact on the ability to identify the information needed to answer a reading comprehension question (cf., e.g., Champeau de Lopez et al., 1997). And finally, one of the most concrete and relevant distinctions concerns the need for a classification of questions according to the type of answer they require, often referred to as *answer typing* (cf., e.g. Li & Roth, 2002).

In sum, an exploration of these relevant aspects of the context of the interpretation of the answers – the questions they answer and the text the questions are about – opens up important strands for future research on automatic content assessment.

Diagnosis categories Another strand concerns the question which diagnosis categories are appropriate and useful for content assessment. Content assessment in CoMiC currently distinguishes: correct, missing concept, extra concept, blend, and non-answer. Yet, in particular in light of the just mentioned work on answer typing, it seems clear that more detailed diagnosis categories could be developed, which more directly take into account what is known about the task and the context.

Adaptivity of analysis Given the high number of form errors in learner data – for example, in the CREE corpus a sentence on average contains more than two form errors – deep linguistic analysis and model construction often is not feasible. However, there often are well-formed "islands", in which a dedicated analysis is possible or even important. Such patterns include semantic units expected in the answer, e.g., as the result of answer typing, or specific linguistic constructions identified in the answer which require special treatment (e.g., negation). We intend to explore the identification of such patterns and other islands of compositionality, and how their analysis can adaptively be integrated into the overall architecture discussed in this paper. The overall aim is to discover which linguistic representations are effective and robust in a computational-linguistic comparison of the meaning of clauses and text fragments, and for what tasks and contexts such comparisons can effectively be calculated.

Related to this last point is the fact that our work in this paper and the published work on content assessment and related tasks such as the RTE challenge so far have almost exclusively focused on English. This raises the question how much the techniques which have been and are being developed are tuned to the specifics of English. Approaches which compare meaning based on representations close to the surface string clearly will benefit from the relatively fixed word order and limited morphological variation found in English compared to other languages. It thus will be important to explore languages other than English, such as the German data we are targeting with the CREG corpus, to explore the need for a flexible analysis regime adaptively comparing meaning at different depth of analysis and considering multiple representations in parallel.

6 Summary

In this article, we presented an annotation-based NLP architecture in which we realized a content-assessment approach which successfully evaluates the meaning of answers to authentic reading comprehension exercises. The work builds on the approach first explored in Bailey & Meurers (2008), with a focus on the parallel integration of multiple representations as the basis for content assessment, the NLP architecture and data structure needs arising from this focus, and the research issues and avenues which arise from it and for content assessment in general. We also defined a corpus exchange format and make our English reading comprehension corpus available in that format, which we hope will support sustained research on content assessment including a meaningful direct comparison of approaches on shared data sets for authentic tasks.

Building on the CoMiC approach discussed in this paper, we identified a number of important avenues for future research on automatic content assessment, which we are currently exploring in project A4 of the SFB 833. While being rooted and applicable to a task of clear practical relevance – evaluating the content of answers to reading comprehension questions as part of intelligent tutoring systems and language testing – our research in this domain ultimately aims to contribute to the general question how meaning comparison can take place in realistic situations, in which ill-formed language or differences in situative knowledge or world knowledge make a complete analysis difficult or impossible.

References

Amaral, L., D. Meurers & R. Ziai (2011). Analyzing Learner Language: Towards A Flexible NLP Architecture for Intelligent Language Tutors. *Computer-Assisted Language Learnig 24(1)* . http://purl.org/dm/papers/amaral-meurers-ziai-10.html

Anderson, L. W. & D. Krathwohl (eds.) (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman Publishers.

Bailey, S. (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.

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. http://purl.org/dm/papers/bailey-meurers-08.html

Banerjee, S. & A. Lavie (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*.

Brockett, C. & W. B. Dolan (2005). Support Vector Machines for Paraphrase Identification and Corpus Construction. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*. pp. 1–8.

Champeau de Lopez, C., G. Marchi & M. Arreaza-Coyle (1997). A Taxonomy: Evaluating Reading Comprehension in EFL. *English Teaching Forum* 35(2), 30–42.

Daelemans, W., J. Zavrel, K. der Sloot & A. van den Bosch (2007). *TiMBL: Tilburg Memory-Based Learner Reference Guide, ILK Technical Report ILK 07-03*. Induction of Linguistic Knowledge Research Group Department of Communication and Information Sciences, Tilburg University, P.O. Box 90153, NL-5000 LE, Tilburg, The Netherlands, version 6.0 ed.

Dagan, I., B. Dolan, B. Magnini & D. Roth (2009). Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering* 15(4), i–xvii.

Denkowski, M. & A. Lavie (2010). Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pp. 250–253.

Ellis, R. (2005). Instructed language learning and task-based teaching. In E. Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning*, Mahwah, NJ: Routledge, pp. 713–728.

Ferrucci, D. & A. Lally (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering* 10(3–4), 327–348. Gale, D. & L. S. Shapley (1962). College Admissions and the Stability of Marriage. *American Mathematical Monthly* 69, 9–15.

Götz, T. & O. Suhre (2004). Design and implementation of the UIMA Common Analysis System.

IBM Systems Journal 43(3), 476–489.

Harold, E. R. & W. S. Means (2004). *XML in a Nutshell*. Sebastopol, CA: O'Reilly, 3rd ed.

Hatzivassiloglou, V., J. Klavans & E. Eskin (1999). Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning. In *Proceedings of Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP'99)*. College Park, Maryland, pp. 203–212.

Irwin, J. W. (1986). *Teaching Reading Comprehension Processes*. Engelwood Cliffs, New Jersey: Prentice-Hall, Inc.

Krifka, M. (2007). Basic Notions of Information Structure. In C. Fery, G. Fanselow & M. Krifka (eds.), *The notions of information structure*, Potsdam: Universitätsverlag Potsdam, vol. 6 of *Interdisciplinary Studies on Information Structure (ISIS)*.

Leacock, C. (2004). Scoring Free-Responses Automatically: A Case Study of a Large-Scale Assessment. *Examens* 1(3).

Li, X. & D. Roth (2002). Learning Question Classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*. Taipei, Taiwan, pp. 1–7.

Manning, C. D. (2006). Local Textual Inference: It's hard to circumscribe, but you know it when you see it – and NLP needs it. Ms. Stanford University.

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. http://purl.org/dm/papers/meurers-ott-ziai-10.html

Pearson, P. D. & D. Johnson (1978). *Teaching Reading Comprehension*. New York: Holt, Rinehart and Winston.

Pérez Marin, D. R. (2007). Adaptive Computer Assisted Assessment of free-text students' answers: an approach to automatically generate students' conceptual models. Ph.D. thesis, Universidad Autonoma de Madrid.