3.1 General information about project A4

3.1.1 Title: Comparing Meaning in Context

3.1.2 Research areas: Computational Linguistics

3.1.3 Principal investigator

Prof. Dr. Detmar Meurers
* 29.11.1967, German
Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen
Wilhelmstr. 19
72074 Tübingen,
Tel.: +49(0)7071-29-73963
E-Mail: detmar.meurers@uni-tuebingen.de

Do the above mentioned persons hold permanent positions? yes

3.1.4 Legal issues

This project includes

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>research on human subjects or human material.</td>
</tr>
<tr>
<td>2.</td>
<td>clinical studies.</td>
</tr>
<tr>
<td>3.</td>
<td>experiments involving vertebrates.</td>
</tr>
<tr>
<td>4.</td>
<td>experiments involving recombinant DNA.</td>
</tr>
<tr>
<td>5.</td>
<td>research involving human embryonic stem cells.</td>
</tr>
<tr>
<td>6.</td>
<td>research concerning the Convention on Biological Diversity.</td>
</tr>
</tbody>
</table>

3.2 Summary

How can meaning be analyzed and compared in realistic situations, in which ill-formed language or individual differences in situative or world knowledge complicate or even preclude a complete linguistic analysis? This question captures the underlying motivation of Project A4 Comparing Meaning in Context. The computational linguistic project A4 investigates which linguistic representations can be obtained effectively and robustly for comparing the meaning of sentences and text fragments in context, and how the meaning comparison can be realized on this basis. To make these general questions concrete in a specific real-life task, the project focuses on reading comprehension exercises, which are commonly used in current foreign language teaching. Given a reading text and a comprehension question asked about that text, teachers evaluate whether a response given by a student actually answers the question – in other words, whether the learner answer conveys the same meaning as the target answer the teacher envisaged. The project thus focuses on an authentic task in which meaning is explicitly compared, and for which the meaning assessment performed by the teachers can serve as a gold-standard against which the computational analysis can be evaluated.

Building on the Corpus of Reading Comprehension Exercises in German (CREG) and the meaning comparison architectures CoMiC and CoSeC established in the first phase, in the second phase the project will i) investigate the information structuring of the learner answer given the question, and its impact on meaning comparison, ii) explore the relation between the question and how the requested information is encoded in the reading text, and iii) enrich the nature of the representations which are aligned in meaning comparison, to support many-to-many alignments, and to go beyond semantic term alignment by investigating the use of logical inference for a more general treatment of negation.
3.3 Project development

3.3.1 Report and state of the art

Empirical basis: The Reading Comprehension Corpus in German (CREG)

Given the importance of a real-life, task-based corpus as empirical basis of the project, we started by developing the WEb-based Learner COrpus Machine (WELCOME), a web-based application supporting our project subcontractors at The Ohio State University and Kansas University in collecting a structured corpus from the reading comprehension exercises used in their regular German courses at all levels (Meurers, Ott & Ziai, 2010).

Using WELCOME, we collected the Reading Comprehension Corpus in German (CREG), the first data collection of its kind. It consists of reading comprehension questions, the reading material these questions are about, the target answers provided by the teachers, and the learner answers. Each learner answer was evaluated in terms of meaning by two teachers hired by our US partners to make explicit whether a given student response actually answers the question or not. Corpus collection continued during the entire first phase of the project and as of July 2012 the corpus contains 159 texts, 1,470 questions with 1,916 target answers, and 35,211 student answers. The structure of the corpus and the collection process are documented in Ott, Ziai & Meurers (2012). WELCOME was essential for the creation of this highly structured corpus, which required distributed data collection in a real-life setup where the teachers as the domain experts in language teaching are not computer or corpus experts. The CoMiC corpus is freely available for research under a Creative Commons by-nc-sa license, as is the WELCOME application. In support of a broad usefulness and relevance of CREG, we also collected a corpus from a control group of German native speakers.

Computational approaches

Turning to the meaning comparison approaches explored by the project, the project pursued two routes. The first approach, implemented in the CoMiC system, explores the use of a wide range of representations of the learner and target answers, from the surface forms to deeper levels of linguistic modeling. The second approach, realized in the CoSeC (Comparing Semantics in Context) system, performs meaning comparison on the basis of an underspecified semantic representation robustly derived from the learner and target answers.

CoMiC. As proposed, we started realizing the CoMiC approach with a reimplementation and extension our English Content Assessment Module (Bailey & Meurers, 2008). The resulting CoMiC-EN system (Meurers, Ziai, Ott & Bailey, 2011a) is based on the general UIMA NLP architecture (Ferrucci & Lally, 2004). The key component of CoMiC-EN is the automatic stand-off annotation of learner and target answers using exchangeable modules, the so-called annotators. The modular architecture readily supported the development of the multi-level CoMiC-DE system (Meurers, Ziai, Ott & Kopp, 2011b), the first content assessment system for German. Figure 1 exemplifies the different levels of linguistic modeling and the knowledge needed to perform the meaning comparison between learner and target answers on that level. All of these have been explored in the project, except for logical inference.

Figure 1: Levels of linguistic modeling for meaning comparison
Some of the annotators we developed for the CoMiC systems are wrappers around state-of-the-art NLP tools, whereas others were created by the project from scratch. The automatic annotation on the different levels of linguistic abstraction is used to create a set of alignment links between the words, chunks and dependency triples of each student answer and the corresponding target answer of the exercise. The overall alignment configuration for each pair is then classified as a correct or incorrect meaning comparison in a machine learning step, for which the project explored Memory-based Learning (TIMBL, Daelemans et al., 2007) in Meurers, Ziai, Ott & Bailey (2011a) and Support Vector Machines (LibSVM, Chang & Lin, 2011) in Ziai, Ott & Meurers (2012). In the content assessment experiment described in Meurers et al. (2011b), the CoMiC-DE system yields a state-of-the-art performance of 84.4% accuracy for the binary classification distinguishing semantically correct and incorrect answers, on a balanced subset of the CREG corpus.

**CoSeC.** Complementing the multi-level CoMiC approach starting from the surface representations, with the second system, CoSeC, we wanted to explore an approach with closer ties to formal semantics. Using an explicit semantic formalism in principle can support the precise representation of meaning differences, and it makes it possible to directly represent information structure as a structuring of semantics representations, in line with the structured meaning approach (Krifka, 2007). We decided on using Lexical Resource Semantics (LRS, Richter & Sailer, 2004), given that underspecified semantic formalisms avoid the costly computation of all readings and provide access to the building blocks of the semantic representation, while the dominance constraints provide the information about their composition.

As described in Hahn & Meurers (2011), LRS representations can be derived automatically from dependency structure using a two-step approach. First, the dependency structure is transformed into a completely lexicalized syntax-semantics interface representation, which abstracts away from some form variation at the surface. These representations are then mapped to LRS representations. The approach is robust in that it always results in an LRS structure, even for ill-formed sentences. CoSeC then aligns the LRS representations of the target answer and the student answer to each other and also to the representation of the question. The alignment approach takes into account local criteria, in particular semantic similarity, as well as global criteria measuring the extent to which the alignment preserves structure at the level of variables and dominance constraints. Consider the pair of examples in (1), based on data from the CREG corpus.

(1) a. *Der Text ist von einer Frau geschrieben.*
   ‘The text was written by a woman.’

b. *Ein Mann hat den Text geschrieben.*
   ‘A man wrote the text.’

Figure 2 shows the LRS dominance graphs for the semantic subparts of the underspecified semantics. Example (1a) is shown on the left and (1b) on the right, with dashed alignments between them (where locally compatible, e.g., note that *woman* is not aligned to *man*). Besides the actual semantic terms, the graphs contain the meta-variables A, ..., M, which indicate where formulas can be plugged in to obtain resolved representations, as typical for underspecified semantic formalisms.

![Figure 2: An alignment between the semantic representations of (1a) and (1b)](image)

Overall meaning comparison is done based on a set of numerical scores computed from alignments and their quality. In the meaning assessment experiment described in Hahn & Meurers (2012), the best accuracy achieved by CoSeC is 86.3%, outperforming CoMiC-DE on the same CREG subcorpus. CoSeC thus confirmed that an approach based on formal semantic representations can be competitive for content assessment in real-world contexts. The approach includes a basic version of information structuring, which contributes 5.4% to this result.

---

1For simplicity, we show grammatical function labels instead of semantic role labels.
Dependency parsing of learner language. Both CoSeC and CoMiC rely on dependency parsing so that we wanted to study more closely what effect the learner language input has on state-of-the-art parsers such as MaltParser (Nivre et al., 2007), which generally are trained on native language corpora. We found that learner language can be parsed with labeled attachment scores between 79% and 86%, depending on the quality of the part-of-speech tagging. However, specific properties of learner language, such as instances of sentences missing verbs, were shown to be highly problematic (Ott & Ziai, 2010). In another parsing experiment on the CREG corpus (Krivaneck & Meurers, 2011), we showed that a rule-based dependency parsing approach using a hand-crafted lexicon (WCDG, Foth & Menzel, 2006) more reliably identified the main head-argument relations than a statistical parser (MaltParser), which performed better on adjunct relations.

Research context

Considering our work in the context of related work done outside of the SFB, CoMiC and CoSeC are the first short answer assessment systems for German. So we here compare them to systems for English, where the state of the art is represented by approaches such as the one by Sukkarieh & Blackmore (2009) and Mohler et al. (2011). Sukkarieh & Blackmore report question-based accuracy figures averaging to 84.8%, which is on par with the results we reported above. A comparison to Mohler et al. (2011) is detailed in Ziai et al. (2012), which also explains how difficult it is to compare short answer assessment systems in general. Our article emphasizes the need for common data and shared tasks in this area, such as the proposed SEMEVAL 2013 task on Student Input Assessment. The Corpus of Reading Exercises in German (CREG) as the first data collection of its kind for German is also designed with this need for common data in mind, both in computational linguistics and in other areas, such as second language acquisition research.

To conclude, on the basis of the CREG corpus and the approaches to meaning comparison we developed in the first phase, the project is now able to tackle the following research questions broadening the analysis and deepening the linguistic modeling: How can the information structure of the learner answers be analyzed automatically, and how can these information structure distinctions support a more precise meaning comparison in context? How is the information requested by the reading comprehension questions encoded in the reading texts, and how does the nature of that encoding influence the realization of answers and how they are best analyzed? How can the linguistic representations used in our approaches be extended to support more advanced alignments, such as a semantics-based many-to-many alignment?

M.A. theses

Kassner, L. (2012). Harvesting paraphrases from recurring sequences in aligned bilingual corpora: From data-induced units to linguistically interpretable categories.


PhD theses


3.3.2 Own project-relevant publications

a) Selected on the basis of peer-review


Publisher site: http://benjamins.com/#catalog/books/hsm.14.05ott


3.4  Project outline

In the first phase, the project established the CREG corpus and the CoMiC and CoSeC meaning comparison architectures, which in the second phase will allow the project i) to investigate the information structuring of the learner answer given the question, and its impact on meaning comparison, ii) to explore the relation between the question and how the requested information is encoded in the reading text, and iii) to enrich the nature of the representations which are aligned in meaning comparison, to support many-to-many alignments, and to go beyond semantic term alignment by investigating the use of logical inference for a more general treatment of negation.

3.4.1  The information structuring of answers and its impact

In our early work on English (Bailey, 2008; Bailey & Meurers, 2008) we observed that answers frequently repeat words given in the question. In example (2), the target answer (TA) repeats “the moral question raised by the Clinton incident” from the question, whereas the first learner answer (LA1) does not. Yet both TA and LA1 essentially answer the question in the same way.

(2)  Q:  What was the major moral question raised by the Clinton incident?

TA: The moral question raised by the Clinton incident was whether a politician’s personal life is relevant to their job performance.

LA1: A basic question for the media is whether a politician’s personal life is relevant to his or her performance in the job.

LA2: The moral question raised by the Clinton incident was whatever.

The issue arising from the occurrence of such given material for a content assessment approach based on alignment is that in principle all alignments are counted, yet those for given material do not actually contribute to answering the question – as illustrated by LA2, which does not actually answer the question. We concluded that an answer should not be rewarded (or punished) for repeating material that is given in the question. In CoMiC, all given words are automatically marked up using the stand-off annotation available in the UIMA-based architecture and the marked up words are ignored in the alignment.

This givenness filter is based on a characterization of the material we want to ignore, motivated by the fact that it is easy to identify the material that is repeated from the question. While this is often effective, in general it is much too simplistic. First of all, identifying words repeated from the question as given is tied to surface representations, so it cannot capture when the same given information is expressed in other words. More generally, a form-based givenness filter falls short of the semantic notion of givenness coined by Schwarzschild (1999), which also includes co-reference and contextually accessible information as given. A binary classification identifying information given in the question also is not sufficient for identifying non-given (i.e., new) material in the answer that is unrequested by the question. Such unrequested new information is not needed and may even be harmful when determining whether the question has been answered. Example (3) from the CREG corpus illustrates this with two appropriate answers, TA and LA, which are essentially identical in terms of providing the same information needed to answer the question (shown un-
lined), and in terms of the given material repeated from the question (shown in bold), but the LA in addition contains a lot of background material.

(3) Q: *An was denken viele Menschen, wenn sie von Weißrussland hören?* ‘What do many people think of when they hear about Belarus?’

TA: *Sie denken an die Tschernobyl-Katastrophe von 1986.*

‘They think of the Chernobyl disaster of 1986.’


‘Foreigners thinking about Belarus think less of vacation but rather of the Chernobyl disaster of 1986. Back then, parts of a nuclear plant exploded and some areas of Belarus were polluted by the radioactivity.’

The limits of a basic given/new distinction also become apparent when considering alternative questions. For such questions, the answer has to select one of an explicitly given set of alternatives, as in (4).

(4) Q: *Ist die Wohnung in einem Neubau oder einem Altbau?* ‘Is the flat in a new building or an old building?’

TA: *Die Wohnung ist in einem Neubau.*

‘The flat is in a new building’

LA: *Die Wohnung ist in einem Neubau.*

‘The flat is in a new building’

The question asks whether the flat is in a new or in an old building, and both alternatives are explicitly given in the question. The learner answer picked the same alternative as the target answer – indeed, the two answers are identical. But the basic givenness filter excludes from alignment all the material repeated from the question. Hence the content assessment classification fails to identify the student answer as appropriate.

The current givenness filter clearly is limited, both in terms of its empirical success and in terms of the conceptual links to the linguistic research issues that it supports. In the second phase, we therefore plan to replace it with an actual analysis of the information structure of the answer in the context of the question. The task-based CREG corpus with the explicit reading comprehension questions and the explicit texts they refer to provide a rich empirical basis for this research strand.

Analyzing the explicit information about the question and the answer provided by the CREG corpus 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, we propose to identify which part of an answer constitutes the so-called focus answering the question. Here focus is to be understood as the minimal part of the response that represents a compatible alternative given the meaning of the question, in line with Roberts (1996) and also Rooth (1992)’s notion of question-answer congruence.

Work packages

Moving from the simple givenness filter to an analysis of information structure involves linking the conceptual distinctions and discussion of information structure in the literature to the data found in the CREG corpus. The notion of focus needs to be operationalized in a way that supports the reliable annotation of this distinction for the learner and target answers in the corpus, first manually (WP1) and then automatically (WP2). The resulting information structure annotation can then be integrated into the meaning comparison (WP3) to improve the quality of the answer assessment.

**WP 1: Focus annotation scheme development and gold-standard annotation.** While research on information structure has a long and varied history, the corpus annotation of information structural differences is only starting to receive more attention (Baumann et al., 2004; Ritz et al., 2008; Riester et al., 2010; Calhoun et al., 2010), notably including work done in the SFB 632 ‘Information Structure’. A general problem identified by this research strand is that it is difficult to reliably annotate information structural notions for texts lacking explicit task and context information; this includes news texts, which are implicitly based on time-specific world knowledge and inferences. This is where the CREG corpus – in which the explicit reading comprehension task providing access to the answer in the context of a concrete question asking about a specific text – provides an ideal experimental sandbox for exploring information structure in an explicit, authentic task.

In developing an information structure annotation scheme, we will explore two groups of cues in annotation: i) syntactic cues, such as the sentence topology (Höhle, 1986; Musan, 2002) and phrase boundaries in gen-
eral, and ii) semantic cues, based on the properties of the question. The latter will include a classification of questions by answer type, such as LOCATION or ENTITY, which allows focus annotation to be semantically constrained to expressions that are compatible with the answer type. This idea bears an interesting resemblance to a structured meaning analysis of questions as described in Krifka (2001), where such answer types are called ‘restrictions’ on the domain of the question meaning.

Using the annotation scheme thus developed, we plan to manually annotate at least 10% of the CREG corpus as a gold standard. Two annotators will be used for this subcorpus to determine inter-annotator agreement, followed by reconciliation and iterative improvement of the annotation guidelines.

One anticipated problem with focus annotation in connection with CREG is the noise introduced by non-well-formedness in learner language. To address this problem, we will follow Falko (Lüdeling et al., 2005; Reznicek et al., 2012) in first determining so-called ZH1 target hypotheses, which provide minimal correction of form errors in learner errors. We performed a case study, annotating a subset of CREG with such target hypotheses and obtained an exact agreement of 72% between annotators. Based on this experience and further refined annotation guidelines, it will thus be realistic to provide a ZH1 annotation for CREG.

Complementing the focus annotation by trained annotators, we will also investigate the use of crowdsourcing as offered by Amazon Mechanical Turk (AMT) or CrowdFlower, which was shown to be as effective as trained annotators for learner language annotation tasks (Evanini et al., 2010; Tetreault et al., 2010) and was also successfully used in textual entailment (Zeichner et al., 2012) and paraphrasing (Chen & Dolan, 2011). To turn corpus annotation subtasks into Human Intelligence Tasks (HITS) carried out by non-experts, our work will include the automatic generation of multiple-choice items to reduce annotation task complexity.

**WP 2: Automatic focus detection.** Based on the gold-standard annotation of focus in the CREG subcorpus resulting from WP 1, we will develop an automatic approach to identifying focus in answers. In a standard supervised machine learning setup, we plan to train a classifier using our annotated data to predict whether a given unit in an answer is part of the focus or not. As in our previous work employing machine learning (e.g., Ziai, Ott & Meurers, 2012; Vajjala & Meurers, 2012), different machine learning approaches and implementations such as the WEKA (Hall et al., 2009) toolkit and LIBSVM (Chang & Lin, 2011) will be explored. However, the real focus of this work package will be on identifying the linguistic features supporting focus classification.

The feature exploration will start with the syntactic and semantic cues mentioned in the introduction of the manual annotation task above. In automatically determining answer types of questions, we can build on work in Question Answering (e.g., Li & Roth, 2002; Croce et al., 2011) where good accuracy has been achieved using a fixed set of answer types. Syntactic features will be derived from automatic syntactic analysis based on our previous parsing experience with CREG (Ott & Ziai, 2010) in conjunction with automatically determined topological fields following, e.g., Becker & Frank (2002). Such an approach, involving the analysis of both questions and answers, can also be informed by related QA work such as Damjanovic et al. (2010).

While we anticipate that some more complex reading comprehension questions will pose significant problems for focus detection, based on our first analysis of question types in a CREG subset (Meurers et al., 2011b) and the above reference points, we expect the automatic approach to work well for most of the CREG corpus.

**WP 3: Integration and evaluation of information structuring in content assessment.** While information structure annotation is of general relevance, the more specific motivation for identifying focus in answers in this project is to improve the alignment of learner answers to target answers in the content assessment task. Employing an automatic focus detection approach, we can limit the alignment procedure to the focus of each answer, disregarding irrelevant extraneous information, yet still paying attention to given material when it is part of the focus. In pursuit of this goal, the third work package will integrate the focus annotation layer in the alignment process of our meaning comparison approaches. By developing a weighting scheme for individual alignments, focused material can, for example, be weighted more prominently in alignment in order to reflect its importance in assessing the answer.

### 3.4.2 The reading text as information source

The project so far has focused on the linguistic and computational analysis of the learner and target answers in the context of the question. At the same time, when we created the Corpus of Reading Comprehension in German (CREG), we intentionally included the reading texts in the corpus, so that we have access to the source of the information that the learner referred to in order to answer the question. The second broad area of research in the new project phase is designed to integrate the reading text into the content assessment approach.

The CoMiC and CoSeC systems that have been developed in the first phase both assess student answers to reading comprehension exercises automatically. In order to do so, both systems crucially make use of the target answers in CREG, which were provided by the teachers. Content assessment evaluates the quality of
the alignment between learner answer and the closest target answer. Yet, this is not possible for questions for which no target answers have been specified, or for learner answers that do not aim at expressing the meaning of one of the target answers envisaged by the teacher. Requiring teachers to provide target answers for reading comprehension questions that they want to grade with an automatic content assessment system is a low threshold compared to the design of detailed scoring rubrics for each question, as used by the English answer assessment system C-Rater (Sukkarieh & Blackmore, 2009). Nevertheless, it clearly is relevant to explore when and how it is possible to identify the information directly in the reading text that the question asks about. Even if this does not result in a content assessment approach that can operate without manually specified target answers, the information that is identifiable in the text and the way in which this information is expressed in the text in any case is directly relevant for analyzing the learner answers. It can also help us interpret why the learner included certain contents and used specific forms to express it. The CREG corpus collected in the first phase is an ideal empirical basis for pursuing this exploration. The task of identifying the information in the text that is needed to answer a given question is related to the task of automatic Question Answering (QA, cf., e.g., Dang et al., 2007), even though there also are some clear differences. QA traditionally uses a three-step strategy to find snippets of texts answering the question entered into a QA system by the user: i) question analysis, ii) document retrieval, and iii) answer extraction. Question analysis usually deals with the detection of question types, that is, the type of information asked for, e.g., a place, location or time. In some works, this also deals with topic or focus detection (Cabrio et al., 2007). Furthermore, question analysis also extracts query terms used in the subsequent step. Document retrieval uses search engine technology to find candidate documents containing possible answer snippets. Finally, answer extraction aims at finding relevant snippets of these documents that contain an answer that is to be presented to the user.

For our purposes in CoMiC and CoSeC, a couple of differences to this strategy employed in QA present themselves. Firstly, there is no need to go into document retrieval since the document in which the requested information must be found is explicitly given by the reading comprehension task. Secondly, given that our corpus includes target answers, our analysis and evaluation can make use of this set of model answers for each question. And thirdly, QA is an application domain, whereas the project pursues the dual goal of advancing the understanding of meaning analysis in context and of testing that understanding in a content assessment system. So we are interested in exploring the identification and annotation of the relevant pieces of information in the text, both manually and automatically. QA automatically annotates texts with semantic labels, such as answer types compatible with specific semantic type information (see, e.g. Prager et al., 1999). We aim at making use of more broadly applicable question types as defined in the reading comprehension literature (Day & Park, 2005), with question types used in QA serving as a supplementary set of clues provided to the system. Compared to a QA setting, the more explicit nature of our task setting will also enable a more fine-grained annotation of the source and nature of the information in the reading text and its relation to the question.

We also intend to explore answer validation as carried out in the Answer Validation Exercise (Rodrigo et al., 2009). In this shared task, answers given by QA systems were automatically evaluated using texts supporting the answers. While this is clearly related, our task is placed in a different setup, with the texts from the reading comprehension exercises being substantially longer compared to the small fragments of supporting text used in answer validation. Another related perspective is the one of Deep Read (Hirschmann et al., 1999) aiming at answering reading comprehension questions, and ABCs (Wellner et al., 2006), aiming at automatic text understanding using first-order logical representations. For evaluating their systems, both teams tested the performance of their approaches on reading comprehension exercises. In a similar vein, performance in reading comprehension has been used as one of the evaluation metrics at the Question Answering for Machine Reading Evaluation task (QA4MRE, Peñas et al., 2011). While all these approaches are grounded in the QA paradigm with its different requirements and perspectives, clearly the integration of their insights is likely to be useful for our goal of exploring the connection between the reading comprehension questions, the place where the requested information is expressed in the text and how it is encoded, and the target answers concisely encoding the relevant information.

Work packages

WP 4: Question form and information source annotation. In order to develop and evaluate an automatic approach identifying the relation between a question and the units in the reading text contributing to the answer, this relation must first be identified and annotated manually. The annotation scheme to be developed and applied in this work package will support the mark-up of the parts of the text that contribute to answering the question – we will refer to this as information source annotation. In the most basic case, we envisage information source annotation as a binary classification of each sentence in the reading text specifying whether the sentence contains information requested by a specific question or not. In a reading comprehension task with multiple questions, each sentence of the reading text thus is annotated with a list of the question IDs that it provides information for. For each question, the information source annotation thus provides a coarse-grained characterization of the parts of the text that we must look at to answer the question.
Developing a more fine-grained information source annotation will require i) supporting smaller information source units than full sentences, and it should ii) include a classification of how the information conveyed by that source contributes to the answer. As a first step towards a more fine-grained information source annotation, questions will be classified using a scheme derived from work on reading comprehension task design (Day & Park, 2005). They distinguish several forms of questions, among them yes/no-questions, different flavors of wh-questions; and they identify different types of comprehension required from the student answering the question, such as literal transfer of material, re-organization of several parts, or inferences to be drawn. In a pilot study, we already successfully applied this classification scheme to a subset of the CREG data containing 177 questions and 1032 student answers (Meurers et al., 2011b). The question form annotation will serve as a guide to the annotation of the markables in the text that contain the required information. The decision on which linguistic units to annotate will be based on the question forms and comprehension types. For example, for who-questions asking for information that can be transferred literally from the text, noun phrase chunks will be prioritized as markables.

On this basis, we will manually annotate the question form and the information source for 20% of the questions and corresponding texts in CREG. Following standard practice, the annotation will be individually carried out by two annotators and inter-annotator agreement will be calculated to provide feedback on annotation scheme design and the annotation manual – and, more generally, on the question, which aspects of question form and information source annotation can reliably be performed given the information included in the corpus. A third person will serve as a judge in cases where there is no agreement between the annotators. As in the case of the focus annotation in WP 1, we will also investigate the use of crowdsourcing of this type of annotation, which seems feasible for such a meaning-based task. For example, for the first step of information source annotation (the binary sentence classification task introduced at the beginning of the WP 4 description above) the Human Intelligence Task will present the workers with sentences from the reading texts along with questions from CREG and ask them to select which of the sentences are needed to answer the question. For question form annotation, on the other hand, one can show the question to be classified and ask the worker to make a multiple choice selection between different reading comprehension question types (derived from Day & Park 2005).

WP 5: Automatic information source detection. The gold standard annotation in WP 4 will be used for training and testing in supervised machine learning experiments. This task is divided into two parts: Firstly, we will aim at performing the question classification according to Day & Park (2005) automatically, exploring techniques analogous to the detection of question types in question answering. This can be done independently from the following step, since WP 4 provides manual annotation on which the next research steps can be started. Secondly, we will aim at detecting the markables serving as information sources. As a starting point, we will again use sentences, then chunks of verb phrases and noun phrases (cf., e.g., Tjong Kim Sang & Buchholz 2000). We also envisage going beyond ordinary surface markables to allow dependency triples to be identified as information sources. The required dependency annotation of the reading texts (i.e., edited, native language texts) can readily be obtained using a state-of-the-art dependency parser, such as the MaltParser (Nivre et al., 2007). For any type of markables, a binary classifier can then be trained to determine whether the markable contributes to the answer of a given question or not. Just as in WP 2, we will experiment with different machine learning methods – though the main subject of research will be the exploration and characterization of features that reliably support such a classification.

WP 6: Integration into CoMiC and CoSeC. In the final step of this second broad area of research in the new project phase, we will integrate the module dealing with the reading texts into the content assessment systems. We will explore and evaluate both uses of the automatic information source detection: as a complement to target answers and as a replacement of target answers.

3.4.3 Enhancing meaning representation and comparison

While the first two research areas in the proposal focus on i) the information structuring of the answer given the question, and ii) the information source identification in the reading text given the question, in the third research area we intend to advance the representation of the learner and target answers and the methods used for the alignment and its evaluation in the CoSeC architecture.

Distributional semantics. Currently, the semantic representations used in the CoSeC approach represent lexically expressed predicates by lemma names. Lexical semantics is modeled by matching predicates that are recognized as semantically similar in the alignment step using GermaNet (Hamp & Feldweg, 1997). However, such resources generally do not allow comparison across different parts of speech, have limited coverage due to being hand-constructed, and they typically do not represent relations between units larger than words. To avoid these shortcomings, we plan to augment CoSeC with corpus-based distributional measures of similarity.
**N-to-m mappings.** The meaning alignment approach in CoSeC currently only supports links between single elements. This limits the coverage of non-compositional types of semantic similarity, as exemplified by the following pair of sentences.

(5) a. *Sie diskutieren ein Thema.*
   they discuss a topic
   'They discuss a topic.'

b. *Das Thema kommt zur Sprache.*
   the topic comes to the speech
   'The topic comes up.'

The figure below shows excerpts from the LRS representations of the two sentences together with an alignment. While *thema(x1)* and *thema(y1)* can be matched by a single link, the semantic relatedness between the remaining elements cannot be decomposed into a set of links between semantically related atomic elements. To account for such non-compositional links, the project will explore the use of many-to-many alignment links.

![LRS representation](image)

Automatically finding many-to-many links involves three parts. First, relevant candidates for such links have to be identified. Second, the correct mapping of the entities involved has to be identified. In the example above, the object of *diskutieren* (to discuss) should be linked to the subject of *zur Sprache kommen* (to come up). And third, the alignment algorithm has to be extended to find the globally best alignment including n-to-m mappings.

**Logical inferences.** While the level of representation used by CoSeC is a semantic formalism, the comparison of these representations performed by CoSeC is mostly syntactic in nature; yet, a semantic formalism would naturally support logical inferences. As an instance of this, the impact of negation currently is syntactically handled by special rules. A more principled and general approach that deals with a range of polarity and monotonicity phenomena occurring in natural language is the model of natural logic presented by MacCartney & Manning (2007, 2009). We plan to explore this approach on the basis of our semantic representations, with the goal of integrating it into the final classification step of CoSeC.

**Parameter optimization.** The CoSeC alignment algorithm uses several numerical parameters, for instance weights for the different local and global criteria, and constant costs for certain types of alignment links. In the current approach, they are optimized using grid search. However, grid search is inefficient or imprecise, depending on the parameter setting. The main obstacle to efficiently optimizing the parameters is that the accuracy expressed as a function of the parameters is not differentiable. The project will adapt the more efficient and precise optimization algorithm described by Och (2003) for Statistical Machine Translation to CoSeC. Och’s algorithm performs exact line search along randomly determined directions in the parameter space, exploiting the fact that the error rate in the model is a piecewise constant function of the parameters.

**Work packages**

**WP 7: Distributional semantics.** The project will explore extending the CoSeC approach with PMI-IR (Turney, 2001) and vector-space models to inform local semantic similarity. More specifically, we will apply Latent Semantic Analysis (Landauer et al., 1998), Explicit Semantic Analysis (Gabrilovich & Markovitch, 2007) and syntactically enriched models (Erk & Padó, 2008; Mitchell & Lapata, 2008; Thater et al., 2010, 2011). The correct weighting of the numerical similarity scores resulting from the different techniques in the optimization criterion guiding the CoSeC alignment algorithm can be determined automatically on a development set. In order to escape from the imponderabilities of using commercial web search engines as required by approaches such as PMI-IR, the project in its current phase already makes use of very large offline corpora such as deWaC (Baroni et al., 2009) and COW (Schäfer & Bildhauer, 2012).

**WP 8: N-to-m mappings.** The project will investigate methods for automatic induction of inference rules as presented by Lin & Pantel (2001), Szpektor et al. (2004), and Dinu & Wang (2009), using the deWaC (Baroni et al., 2009) corpus. Furthermore, we will explore the use of distributional measures of similarity for phrases by measuring the distribution of chunks rather than words. Since these algorithms induce mappings at the level of surface representations, these have to be converted to mappings on semantic representations. This
will be done by extracting the semantic representations of the units related by the mappings using automatically generated LRS representations (Hahn & Meurers, 2011). Some of the algorithms used for finding semantically similar phrases, such as DIRT (Lin & Pantel, 2001), also induce variable bindings. For n-to-m mappings found by other techniques, the mapping can be induced using distributional semantics. Every multi-word unit has a set of argument slots which are defined by the dependency arcs whose heads belong to the multi-word unit. Given two multi-word units known to be semantically similar, the mapping of their argument slots will be determined based on the distributional similarity of their fillers, as in Lin & Pantel (2001). This generalizes the approach of Szpektor & Dagan (2009) to multi-word units.

We will add a factor measuring the quality of n-to-m links to the optimization criterion guiding the CoSeC alignment algorithm. To this end, candidates for n-to-m mappings have to be assigned a numerical score measuring their quality. Since the techniques for finding n-to-m mappings will use quantitative methods, we envisage making use of the scores assigned by these algorithms. Machine learning on a development set can be used to determine the right weighting of the scores resulting from the different algorithms in the overall CoSeC optimization criterion.

WP 9: Logical inferences. As an instance of a logical inference step that is directly relevant for meaning comparison, we will tackle semantic containment and monotonicity, which covers frequent phenomena such as quantifiers, hyponymy and negation. To this end, we will adapt the approach by MacCartney & Manning (2007), which operates on English surface strings, to LRS representations as derived from our German CREG data. We will automatically derive lexical entailments from GermaNet and then build rules recognizing downward-monotone and non-monotone expressions in German. The final entailment classification step will then be adapted to pairs of aligned LRS representations.

WP 10: Parameter optimization. The optimization criterion guiding the alignment approach depends on some of the parameters linearly and exponentially on some other parameters. By taking logarithms, the criterion can be transformed into a function depending linearly on the second group of parameters. The function and its logarithm can then be optimized separately using Och’s algorithm by iterating the following procedure with each function only for those parameters on which it depends linearly, leaving the other parameters fixed. First, the classification model is trained using the current parameter setting. Then, a direction in the parameter space is chosen randomly. For every sentence pair in the development set, the N best alignments are found efficiently. The piecewise constant objective function will be estimated, whose maximum along the line chosen can then be found efficiently.

3.4.4 Timeline

2013/2 WP 1: Focus annotation scheme development and gold-standard annotation
WP 4: Question form and information source annotation
WP 7: Distributional semantics

2014 WP 1, WP 4: annotation continued
WP 7: distributional semantics method exploration continued
WP 8: N-to-m mappings

2015 WP 1, WP 4: annotation continued
WP 2: Automatic focus detection
WP 5: Automatic Information source detection
WP 8: validation and evaluation of n-to-m mapping approach
WP 9: Logical inferences

2016 WP 2: validation and evaluation of automatic focus annotation
WP 3: Integration and evaluation of information structuring in content assessment
WP 5: validation and evaluation of automatic information source detection
WP 6: Integration into CoMiC and CoSeC
WP 9: continued exploration and evaluation of logical inference component
WP 10: Parameter optimization

2017/1 WP 3: validation and evaluation of information structuring component in content assessment
WP 6, WP 10: validation of system integration and exploration of performance

(Sections 3.5 to 3.7 have been intentionally omitted in the web publication of this proposal.)
3.8 Literature

http://rave.ohiolink.edu/etdc/view?acc_num=osu1204556485

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


http://worldbirmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2005-
journal/LanguageLearningAndError/multilevelerror.doc


Project A4


