3.1 General information about Project A4

3.1.1 Title: Comparing Meaning in Context

3.1.2 Research area(s): 104 Sprachwissenschaften (Computational Linguistics)

3.1.3 Principal investigator(s)
Meurers, Detmar, Prof. Dr.
*29.11.1967, German

Universität Tübingen
Seminar für Sprachwissenschaft
Wilhelmsstr. 19
72074 Tübingen

Tel.: +49 (0)7071-29-75927
E-Mail: detmar.meurers@uni-tuebingen.de

Do any of the above mentioned persons hold fixed-term positions? no

3.1.4 Legal issues

<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 trials</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

Project A4 investigates the nature and interaction of context and sentential meaning in an authentic language-based task: teacher assessment of answers to reading comprehension questions. To determine whether an answer successfully addresses a question on a text, the project analyzes how the answer can be compared to target answers to determine whether the meaning is sufficiently similar given the question and text context. As empirical basis, the project collected the CREG corpus consisting of 36,000 answers, each with an explicit context consisting of a reading comprehension question and the reference text that the question is about. The project developed the short-answer assessment systems CoMiC and CoSeC and showed them to be highly accurate for this real-life task, evaluated with reference to a gold standard independently established by language teachers. In terms of advancing the linguistic modeling driving the analysis, the project in the second phase established that reliable expert annotation of Focus as a key notion of Information Structure can be achieved through an incremental annotation process making use of the explicit context. The relevance of the information structural analysis was externally validated by showing that the Focus annotation improves the quality of automatic answer assessment. Further independent evidence for the role and relevance of the operationalization of Focus was provided through crowd sourcing experiments establishing that untrained native speakers can reliably identify Focus in specific types of authentic, contextualized data. Turning to other aspects of Information Structure, the project established that integrating a distributional-semantic approximation of Givenness into the meaning comparison mechanism further improves the quality of automatic answer assessment.

In the proposed third phase of the SFB, the project will continue the perspective of the second phase and advance it so that i) the meaning comparison underlying the content assessment integrates the breadth of characteristics of the task context for which the answers were produced, ii) the content assessment approach supports provision of feedback in the real-life educational application developed in the transfer project T1, and iii) the information structural analysis successfully developed for explicit question-answer con-
texts is generalized to be applicable to language produced in less explicit contexts. In terms of the task factors to be considered, the nature of the reference text (regarding its linguistic complexity and topic) and the characteristics of the student answers (such as the variability of the forms used) will be investigated in a way connecting the work to domain adaptation as a current research strand in computational linguistics. The project will extend the short-answer assessment system in order to analyze the practical effectiveness of incorporating different task-related information sources for meaning comparisons. To synergistically support the knowledge transfer into the T1 project, information about the meaning comparison underlying the binary assessment will be exposed to facilitate the provision of feedback in the real-life educational setting. Continuing the successful information structural analysis and annotation of authentic language data, we will explore how it can be generalized and applied to language data lacking an explicit question context. Building on the annotation scheme we developed for manually determining the Questions under Discussion (QUD) of every sentence in a text (Riester, Brunetti & De Kuthy submitted), we will develop a computational approach incrementally generating questions for every sentence, which can then be used as guidance for annotating focus, using both the manual and the automatic focus annotation approach developed in the second phase. An analysis in terms of explicit QUDs as a hinge between the information structure of the sentence and the discourse directly addresses the principles and limits of context integration. Harvesting the results of explicitly modeling the nature and context of authentic language use, the project empirically and conceptually contributes to the research goal of the A area to spell out the conditions under which meaning composition and context integration takes place.

3.3 Project progress to date

3.3.1 Report and current state of understanding

Project A4 investigates the analysis of meaning in context, where the functional demands of the task that the language is used for shape the perspective of the analysis carried out. In the first two phases of the SFB, the project successfully identified levels of linguistic representation which can be reliably identified in authentic data to robustly compare the meaning of sentences and text fragments in context. On the one hand, this includes different levels of representation of the answers to the reading comprehension questions, from surface representations to deeper linguistic representations. On the other hand, the functional demands of the task were made explicit through an information structural analysis distinguishing focused and given information to identify where the answer has the potential to address the question. The quality of the conceptual approach was confirmed by showing quantitative gains in the automatic answer assessment system developed by the A4 project, using as reference the gold standard assessment independently determined by teachers evaluating which student answers successfully answered a question and which did not.

3.3.1.1 Advancing meaning comparison and connecting it to related research

Having identified the assessment of short-answer reading comprehension exercises as a valuable real-life testbed for computational approaches analyzing and comparing meaning in context, the CREG (Corpus of Reading Comprehension Exercises in German) corpus (Ott, Ziai & Meurers 2012) with its 36,000 student answers collected by the project in two large German programs in US universities over four years provided a representative empirical test case. The core system developed for the meaning comparison approaches explored by the project is CoMiC (Comparing Meaning in Context; Meurers, Ziai, Ott & Kopp 2011b), an alignment-based system operating in three stages. First, it annotates linguistic units (words, chunks and dependencies) in student and target answers on various levels of abstraction. Second, it finds alignments of linguistic units between student and target answers based on this annotation. Finally, it classifies the student answer based on number and type of alignments, using a supervised machine learning setup. Figure 1 shows a short example from the CREG corpus with alignments between the target answer (TA) and the student answer (SA) used for meaning comparison in CoMiC.

The quality of the NLP analyses on which the system is based was investigated by the project particularly with an eye on the quality of dependency parsing for learner language (Ott & Ziai 2010, Krivanek & Meurers 2014). In addition to the German data, we also evaluated the system for English answer assessment (Meurers, Ziai, Ott & Bailey 2011a). Complementing the multi-level CoMiC approach starting from the surface representations, a second content assessment system was developed in order to explore an approach with closer ties to formal semantics. This CoSeC system (Comparing Semantics in Context; Hahn & Meurers 2012, 2014) performs meaning comparison on the basis of an underspecified semantic representation robustly derived from the student and target answers.
Ott, Ziai, Hahn & Meurers (2013) compared the CoMiC and CoSeC systems with shallow bag-of-word approaches for short-answer assessment. They showed that the two systems using various levels of linguistic abstraction outperform shallow features in “unseen question” scenarios, where the particular question has not been seen before, whereas shallow approaches fare better in “unseen answer” scenarios, where the machine learner saw a range of answers to the same question as part of the training data. Linguistic analysis thus pays off for tasks where the test data is sufficiently different from the training material, i.e., knowledge about language then beats knowledge about the task – though an ideal system would manage to combine both (which we return to in the planned work). As a step in this direction, Rudzewitz (2015) advanced the performance of short-answer assessment accuracy by combining general linguistic and task-specific characteristics to weight the alignments for meaning comparisons in the CoMiC system. He showed that it is possible to fully automatically determine which aligned elements are more important than others, resulting in significantly improved accuracy when these weightings are expressed in the machine learning features for meaning comparisons. This pilot study demonstrated the benefit of making use of both language-wide properties and task-specific characteristics for short-answer assessment – a perspective we plan to pursue further in the third phase of the project.

To broaden our perspective on meaning comparison and connect the research to related research strands in computational linguistics, following up on Ziai, Ott & Meurers (2012), we applied and adapted the CoMiC system to a range of different but related domains. This allows us to determine which information sources are effective in the specific context of short-answer assessment, and which properties are general to meaning comparisons across domains.

We applied the CoMiC system to community question answering in Rudzewitz & Ziai (2015). In contrast to traditional short-answer assessment, no target answer is available. Instead, questions are aligned with answers, giving us the opportunity to develop fine-grained question-answer features re-usable for short-answer assessment. The web-specific highly variable nature of the data allowed us to make the system more robust for real-life applications. Rudzewitz (2016b) extended this work by systematically testing which information sources are effective for meaning comparisons in cases where the language under consideration shows a variable degree of comparability. The study, which used question answering as a test case, showed that in contrast to prior approaches taken in the project, it is not necessarily beneficial to enforce alignments between language that is of limited comparability, but the nature of the variability provides a valuable and effective information source for meaning comparisons.

To tackle answer selection in a multiple choice setting for the 2015 CLEF shared task (Rodrigo et al. 2015), Ziai & Rudzewitz (2015) adapted the CoMiC system to automatically select answers. In this scenario, no target answer is available either, but the potentially correct answer is aligned to a segment of a corresponding reading comprehension text. This study confirms that it in principle is possible to move away from teacher-specified target answers to an automatic approach that infers target hypotheses from the text. To validate the general applicability and extensibility of our approach, Rudzewitz (2016a) explored plagiarism detection as another area involving the identification of meaning equivalence. He extended the CoMiC system with features traditionally used for authorship attribution and showed that the CoMiC system yields highly competitive results for the task of plagiarism detection and readily supports integration of a range of authorship attribution features. Going full circle, he also investigated the performance of the extended feature set for short-answer assessment again and established that features traditionally used for authorship also pick up on characteristics relevant for content assessment and improve results there.

3.3.1.2 Information Structure of answers in the reading comprehension context

One of our central themes in the second phase of the project has been to investigate the role of Information Structure in Content Assessment. The central hypothesis is that the information structuring of the answer

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>What are the objections people have about Hamburg?</em></td>
<td>The stink of fish and fuel at the quays.</td>
</tr>
<tr>
<td></td>
<td><strong>SemType</strong></td>
</tr>
<tr>
<td></td>
<td>SemType</td>
</tr>
<tr>
<td></td>
<td>Fisherr</td>
</tr>
<tr>
<td></td>
<td><strong>SemType</strong></td>
</tr>
<tr>
<td></td>
<td>SA: Der Geruch von Fisch und Schiffsdiesel beim Hafen.</td>
</tr>
</tbody>
</table>

**Figure 1:** Alignments between student and target answer.
according to a question (in our case, the reading comprehension question) provides a partitioning of meaning that is beneficial in assessing the content of the answer.

Focus annotation by experts
Considering Focus as a central notion of Information Structure (Krifka 2007), we examined what constitutes the focus in answers to reading comprehension questions. The rationale for this is that if the focus of the answer is known, Content Assessment can zoom in on the focused part of the answer and, as a consequence, classify the relevant content of the answer more accurately. Our first step was the development of an incremental annotation scheme for focus (Ziai & Meurers 2014) which takes the explicitly provided questions into account and precisely determines the extent of the focus by using meaning-based word substitution tests. In De Kuthy, Ziai & Meurers (2016a), we successfully applied this annotation scheme to a substantial part of the CREG corpus, obtaining inter-annotator agreement for focus annotation above the state-of-the-art for German, and state-of-the-art for focus annotation in general ($\kappa = .7$). Moreover, an extrinsic evaluation of focus annotation in the task of Content Assessment shows that focus has a significant impact on classification accuracy across different data sets.

Focus annotation by non-experts using crowd sourcing
Since manual focus annotation by experts is very time-consuming for large data sets, both for annotator training and the annotation itself, a second component of our work explored the use of crowd-sourcing for focus annotation. Crowd-sourcing as a way of collecting linguistically annotated data has been shown to work well for a number of tasks (cf., e.g., Finin et al. 2010, Tetraeult et al. 2010, Zaidan & Callison-Burch 2011). We investigated how systematically the untrained crowd can identify a meaning-based notion like focus in authentic data and which characteristics of the data and context lead to consistent annotation results. For our non-expert focus annotation studies (De Kuthy, Ziai & Meurers 2015, 2016a,b) we implemented a crowd-sourcing task for the CREG-5K data set. We used the crowd-sourcing platform CrowdFlower.com to collect focus annotations from crowd workers. CrowdFlower makes it possible to require workers to come from German speaking countries, a feature that other platforms such as Amazon Mechanical Turk do not readily support, and it has a built-in quality control mechanism which ensures that workers throughout the entire job maintain a certain level of accuracy.

The results of our crowd-sourcing annotation study reported in De Kuthy, Ziai & Meurers (2015, 2016a) show that majority voting on crowd worker judgments compared to an expert annotation can reach expert level for specific types of data (in particular who-, when-, and where-questions), whereas the crowd cannot handle why-questions, showing that those require more explicit instructions to approximate expert focus analysis. Our studies thus established that crowd sourcing focus annotation is promising a) for exploring the impact of different types of data and instructions on the annotation of authentic data and b) for the large-scale annotation of specific subsets of data.

In search for independently observable indicators for the quality of crowd judgments, we refined the crowd-sourcing approach by taking into account the variability within the set of crowd judgments. The approach presented in De Kuthy, Ziai & Meurers (2016b) is based on the idea that sentences with little variation in the annotation provided by the crowd are more reliably annotated, i.e., are of a higher quality. As indicator for the quality of crowd judgments we therefore spelled out a measure of crowd diversity, Consensus Cost, and investigated its usefulness both intrinsically, by relating it to the expert-based gold standard, and extrinsically, by integrating cost-based focus annotation data in the CoMiC system. For answers annotated with low Consensus Cost, the quality of annotation is generally high, with percentage agreement with the gold standard between 0.7 and 1.0. We thus concluded that there is a clear association between Consensus Cost and annotation quality. A low Consensus Cost can serve as a proxy for high annotation quality. For the extrinsic evaluation we were able to show that training data selection based on Consensus Cost is beneficial for the content assessment with the CoMiC system.

Automatic focus annotation
Finally, we implemented a first automatic approach to focus annotation (Ziai & Meurers submitted). We used syntactic and positional features of answer words in conjunction with surface question properties and basic givenness in a logistic regression model which classifies each word individually as either focus or background. The model reaches 77.3% accuracy in distinguishing focus from background on the word level, outperforming the majority baseline (58.1%) by almost 20%. In the extrinsic evaluation, we found that while automatic focus detection still introduces too much noise to be of use in “unseen answers” scenarios, it does already provide a benefit in “unseen question” scenarios.

A computational linguistic approximation of Givenness
Complementing our research strand on Focus, we also advanced the computational integration of Givenness as a second relevant dimension in Information Structure. We investigated how a distributional approximation of Givenness based on word vectors can improve meaning assessment (Ziai, De Kuthy & Meurers 2016).
Specifically, we trained a word2vec model on a large German corpus and treated Givenness as Semantic Relatedness between question and answer words, using estimated similarity thresholds as criteria for treating words as GIVEN. Quantitative results show a modest but systematic improvement over the surface approach to Givenness used in CoMiC so far. On the qualitative side, the approach uncovered interesting cases of where and how distributional semantics falls short of modeling the semantic notion of Givenness.

3.3.1.3 Information source detection

The Corpus of Reading Comprehension in German (CREG, Ott, Ziai & Meurers 2012) used for training and testing in content assessment in addition to the teacher-provided target answers also includes the reading texts as the source of information that the student worked with to answer the reading comprehension questions. As a step towards analyzing the Information Structure of these texts, we explored when and how it is possible to identify the information that the question asks about directly in the reading text. We refer to the task of identifying the sentences that contribute to answering the reading-comprehension question as information source annotation. Building on the acquired experience in designing crowd-sourcing experiments, we set up a crowd-sourcing study on information source annotation in which crowd workers were presented with a text and a reading comprehension question from the CREG corpus and were asked to mark up to five sentences as the information sources. Complementing this manual annotation, we also explored a basic automatic method using a bag-of-words approach that selects the sentences in a text most closely resembling the reading comprehension question in terms of content words.

As an extrinsic evaluation of these crowd-annotated information sources, we employed the CoMiC system to compare the classification accuracy in Content Assessment based on information sources with the accuracy based on target answers. The accuracy results based on information sources obtained by the crowd workers and the automatic approach are slightly below those obtained with target answers. This indicates that successful content assessment of student answers can be also possible based on the information provided by the reading text. Given that the gold standard teacher assessment of the student answers was primarily based on a comparison of student answers and target answers, we are exploring in a separate crowd-sourcing experiment whether the decrease in performance may also be due to the gold standard favoring the target answers over information sources.

PhD dissertations and M.A. theses


3.3.2 Project-related publications by participating researchers

a) Selected on the basis of peer-review


3.4 Project plan

In the proposed third phase of the SFB, the project will continue the perspective of the second phase and advance it in three research strands so that i) the meaning comparison underlying the content assessment integrates the breadth of characteristics of the task context for which the answers were produced (section 3.4.1), ii) the content assessment approach supports provision of feedback in the real-life educational application developed in the transfer project T1 (section 3.4.2), and iii) the information structural analysis successfully developed for explicit question-answer contexts is generalized to be applicable to language produced in less explicit contexts (section 3.4.3). By spelling out context in form of concrete task contexts and explicit question-answer contexts, the project contributes to the common research agenda of the A area investigating the exact conditions under which meaning composition and context interact. The project empirically substantiates the foundational conclusion of the A area that information structure is part of the systematic composition of meaning and by making all components explicit highlights how it simultaneously satisfies lexical, syntactic, and discourse constraints delineating the limits of contextual adaptivity.

3.4.1 Advancing meaning comparison within the task context

The central research challenge tackled in this first research strand of the project is to capture and integrate all of the relevant information available within the specific task context. The issue how the characteristics of a writing task determine what is written is a topic of broader relevance in computer-based assessment (Quixal & Meurers 2016) as well as in the analysis and interpretation of learner corpora (Alexopoulou Michel, Murakami & Meurers submitted, Meurers & Dickinson 2017), so that the identification of task properties in the project will be informed by these analyses of task type and task complexity. Other publications specific to reading comprehension tasks also showed that different text types correlated with different student answer outcomes (Eason et al. 2012). Complementing the wider research context providing candidates for task properties to be considered, the WP will also build on the results already obtained in the pilot study of Rudzewitz (2015), which combined general linguistic and task-specific characteristics for short-answer assessment, and Rudzewitz (2016a) showing that features traditionally used for authorship-identification can contribute relevant information.

The project will advance our approach to short-answer assessment in two work packages: WP1 systematically analyzes which characteristics of text, questions, target and student answers and the relation between these are potentially relevant, making them explicit through automatic or manual annotation, and WP2 integrates the explicitly modeled task information to evaluate its impact on meaning comparison in an automatic short-answer assessment system. Importantly, explicitly identifying and weighting the effective task characteristics of short-answer assessment as a particular instantiation of meaning comparison also allows us to identify which properties of language and context are effective for meaning comparison in general, and which ones are specific to the particular domain of short-answer assessment.

Work Packages

WP 1: Identifying and annotating task properties

To facilitate systematic integration of task-related properties, this work package a) identifies the task properties which are potentially relevant in the context of content assessment of reading comprehension answers and b) develops automatic means or manual annotation of those characteristics in the CREG data to support their empirical analysis and evaluation.

While minimally a student answer can simply be compared to a target answer without requiring further information, we already established that additionally taking the reading comprehension question into account to structure the information in the student answer into given and focused material significantly improves content assessment results. Conceptually speaking, for meaning comparison to be reliable, it should be based on an explicit functional goal fixing the perspective and granularity of the meaning comparison. Complementing this functional context, the reading comprehension task offers further information impacting the nature of a stu-
dent answer and thereby content assessment evaluating it, including characteristics of the reading text, the question, and the set of student answers – both by themselves and in relation to one another. For example, what a student answer states and how it is expressed will be affected by the type of question, its linguistic complexity, and the readability of the text it asks about.

The different task aspects can also be related to one another. For example, by relating text, question and target answer, the way in which the information that the question asks about is encoded in the text (i.e., the information source) and how this compares to the language forms used to express that information in the target answer can be taken into account. Questions requiring the student to pull together information from multiple information sources or that make it necessary to reformulate information in order to be an appropriate answer (e.g., by turning passive into active voice, changing pronominal reference, case marking, or word order) will result in more varied answers, requiring deeper analysis for meaning comparison, than others. Complementing the analysis of the characteristics of the reading comprehension task itself, one can also analyze the nature of the set of student answers produced so far for this task to study and confirm the interplay between the characteristics of the task, the nature of the language it prompts the students to produce, and the analysis techniques and weighting of its components best suited to compare the meaning of student and target answers.

While much of the task information referred to above can be automatically derived from the CREG corpus, for some aspects this requires a dedicated effort. To be able to take into account the relation between question and reading text, we need to be able to identify the information sources in the CREG corpus. As presented in section 3.3.1.3, we already explored crowd-sourcing and automatic identification of information sources. To adequately represent the rich potential interaction between question types and possible distribution of the requested information in the text, the CREG-5K data set annotated with information sources so far will be extended in two ways: We will annotate half of the entire CREG corpus using crowd sourcing and then use this more representative basis to extend our basic computational approach to more reliably identify information sources automatically.

**WP 2: Advancing and generalizing meaning comparison using task properties**

The identification and annotation of task properties enables a multi-faceted analysis of their impact on meaning comparison and its degree of domain specificity. Based on the empirical basis provided by WP1, we can quantitatively and qualitatively evaluate for short-answer assessment which characteristics of the question, the target answer, the student answer, and the text are effective on their own or in relation to one another.

For example, in a pilot study we showed (Rudzewitz 2015) that the relation between student answers and the reading text can be exploited for short answer assessment by increasing the weight of words in the answer that also occur in the reading text. Consider the CREG example (1).

(1) [...] Beim Bummel durch den weitläufigen Park, vorbei am Palmehaus und der Orangerie kann man allen Stress und alle Sorgen vergessen und einfach die Schönheit der Umgebung genießen. [...] ‘Strolling through the extensive park along the palm house and the orangery, one can forget all the stress and the worries and simply enjoy the beauty of the surroundings.’

Q: Wie kann man sich in Pillnitz erholen und den Stress vergessen?

how can one self in Pillnitz relax and the stress forget

TA: Man kann die Schönheit genießen.

one can the beauty enjoy

SA: Man kann erholen und den stress vergessen vorbei am Palmenhaus und der Orangerie und einfach die Schönheit der Umgebung genießen.

one can relax and the stress forget around at the palm house and the orangery and simply the beauty of the surroundings enjoy

The alignment of the word Schönheit (beauty) between the student and the target answer is of greater relevance than the alignments of other words in the answers since Schönheit also occurs in the reading text that the question is asking about. To take this into account, one can enrich the alignment relations on which meaning comparison is based with weights encoding the relevance of the alignment in the task context. The machine learning step performing the content assessment based on features derived from the alignments can then be informed by the weights. Even though the percentage of aligned words in the student answer is low, we showed in Rudzewitz (2015) that an approach that applies the described task weighting in contrast to the baseline correctly predicts the label for the answer because it is informed about the higher importance of task-related words.

We will test the effect of the different task factors in the automatic short-answer assessment system CoMiC. This analysis will provide an answer to the question whether a meaning comparison system can make more accurate predictions regarding the semantic appropriateness of textual content produced by students when it takes into account different task characteristics.
Complementing the efforts to integrate task context information into the meaning comparison as such, the task context can also be used to advance the quality of the underlying NLP analysis. Student writing such as the data in the CREG corpus contains a lot of noise such as typos and other spelling errors. As shown by Keiper et al. (2016), it is possible to improve the NLP analysis for such data by taking into account the context: POS tagging of the CREG student answers was significantly improved by normalizing the spelling based on which words occur in the reading text. Based on the task information made explicit by WP1, we will integrate and extend such task-based normalization of student answers into the content assessment to improve the analysis of the student answers and the alignments between student and target answers. Finally, based on the information-source annotated CREG corpus resulting from WP1, it becomes possible to evaluate the student answers based on these information sources – in contrast to the evaluation based on the target answers performed for the first gold-standard assessment. We will conduct crowd-sourcing experiments to compare student answers assessment based on the information sources with an assessment based on the target answers and with the original teacher assessment of the student answers in the CREG corpus. On the conceptual side, this will provide important insights into the dependence of content assessment on the reference used for comparison. It also stands to generalize the approach to tasks in which no explicit target answers have been provided, significantly extending the applicability of the approach.

3.4.2 Putting content assessment into practice: Supporting synergy with the Transfer Project T1

The transfer project T1 of the SFB, which started in October 2016, is a collaboration with the Diesterweg Verlag that develops an online workbook for the 7th grade Gymnasium and evaluates its success in terms of the learning outcomes supported by the individual, incremental feedback that the online workbook will offer. For analyzing the reading and listening comprehension exercises, the transfer project is designed to integrate the content assessment approach developed by the A4 project. In line with the reviewer and DFG decision to replace one of the positions in T1 with an expectation of more synergy with the A4 researchers, the A4 project plans to actively support the link between basic research and application. To be able to support that link, two components will be provided by this second research strand of the proposal: First, we will adapt and train the content assessment approach for the English workbook tasks. Second, information on the meaning comparison itself needs to be made accessible to be able to provide feedback rather than only obtain a binary judgment distinguishing successful from unsuccessful answers.

Work Packages

WP 3: Adapting and training content assessment for English workbook tasks

In addition to the systems developed for German, the A4 project has successfully applied the CoMiC system to English data in several contexts (Meurers et al. 2011a, Ziai et al. 2012, Ott et al. 2013, Rudzewitz & Ziai 2015, Ziai & Rudzewitz 2015, Rudzewitz 2016a), in part to connect our research to the broader content assessment and textual entailment research strands in computational linguistics typically organized and evaluated on English. Building on this experience, we will collaborate with T1 in adapting and retraining the content assessment system for the exercise types found in the Diesterweg workbook – a strand of work which allows us to directly put into practice the insights into domain adaptation we are gaining in the first research strand in WP 1 and 2.

WP 4: Identifying the nature of the difference as basis for feedback

The content assessment systems developed so far provide a binary judgment distinguishing successful from unsuccessful answers. To be able to provide feedback to students on what is wrong in an unsuccessful submission together with hints on how to improve it, it is necessary for the system to provide access to information on the meaning comparison and how it was focused by the information structuring resulting from the question context. The content assessment system therefore needs to be modified to provide information on the nature of the alignments and the way they become relevant through the information structuring imposed by the question. This information will make it possible to provide feedback on the nature of the information missing from a student answer. Complementing information on present and missing alignments between student and learner answer, for reading comprehension exercises we also plan to support application of the information source detection approach to guide the learner’s attention to the parts of a text that are relevant to answer a question by visually highlighting the information sources. In addition to supporting a rich synergy between A4 and T1, making the detailed alignment and information source information explicitly accessible and displayable will also serve an important function as debugging information for A4 itself, providing the kind of specifics needed for a detailed qualitative analysis of the system performance. Such analyses will be particularly relevant in this phase of A4 considering the first research strand’s goal of making use of a wider range of task information, bringing with it the need to quantitatively and qualitatively study the impact of the different types of information becoming available.
3.4.3 Generalization of the Information Structure approach

In the second phase of the A4 project, we were able to show that content assessment greatly benefits from an information structural analysis of both the target and the student answers. One of the crucial properties of the CREG corpus, the reading comprehension corpus used for the training and testing of our content assessment system CoMic, is that it contains explicit question answer pairs, both for the student answers and the target answers provided by the teachers. Based on insights in formal pragmatics, that the Information Structure of a sentence can be identified in relation to an explicit (or implicit) Question under Discussion (QUD), we developed an incremental annotation scheme (Ziai & Meurers 2014) which explicitly takes questions into account and determines the extent of the focus by using meaning-based word substitution tests.

This enabled us to identify for every student and target answer in the CREG corpus which part of the 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 (2012). Content Assessment can then zoom in on the focused part of the answer and, as a consequence, classify the relevant content of the answer more accurately. This integration of focus information into the CoMic system resulted in a substantial improvement of the content assessment, extrinsically validating the approach.

Another important result of this research is that we were able to show that reliable focus annotation in authentic data is feasible, even for somewhat ill-formed learner language, if one has access to explicit questions and takes them into account in an incremental annotation scheme. We demonstrated the effectiveness of the approach by reporting substantial inter-annotator agreement for our annotated data (De Kuthy et al. 2016a). Before, only few attempts at systematically identifying focus in authentic data had been made (Ritz et al. 2008, Calhoun et al. 2010). They generally ran into significant problems trying to reach good inter-annotator agreement, as they tried to identify focus in newspaper text or other data types where no explicit questions are available, making the task of determining the Question under Discussion, and thus reliably annotating focus, particularly difficult.

To continue with our successful information structural analysis and annotation of authentic data we propose to extend our approach to data that do not contain an explicit question context. A necessary first step here will be to identify the implicit questions under discussion and make them explicit.

We will base our work on the methodology for a combined analysis of data in terms of both discourse and Information Structure proposed in Riester, Brunetti & De Kuthy (submitted), integrating an explicitly spelled out notion of QUDs. In this work, the necessary steps of an analysis procedure based on QUDs are described and the method is demonstrated on authentic data taken from a German spoken-language corpus. This includes the formulation of pragmatic principles that support the analysis of the discourse structure, formulation of adequate QUDs, and finally the analysis of the Information Structure of individual utterances in the discourse in relation to the QUDs.

Work Packages

WP 5: QUD gold standard annotation

For our information structural analysis of a broader range of authentic data, we will explore focus annotation with QUDs by stepwise relaxing the strict task structure we relied on in reading comprehension data. One type of authentic data that still offers some explicit questions as guidance are interviews and panel discussions evolving around a common topic. The data we will use for our combined analysis of QUDs and Information Structure will therefore come from a spoken language corpus compiled from a publicly broadcasted panel discussion, the Stuttgart 21 corpus. This corpus consists of transcripts of 50 hours recorded speech from the public mediation panel on building of a train station (German), which we collected and syntactically annotated. In addition, parts of the Stuttgart 21 corpus have already been annotated with QUDs in a pilot study and were used as one of the empirical sources for the development of the QUD guidelines of Riester, Brunetti & De Kuthy (submitted). An example from the Stuttgart 21 corpus including the QUD and information structural analysis is shown in (2).

(2) $Q_s$: {What can you do at these 16 platforms?} 
$A_s$: und [an diese 16 Bahnsteiggleisen]. können eben, [wenn Sie keine Doppelbelegung machen wollen, das wäre mir neu,] können Sie maximal 16 Züge gleichzeitig abstellen.[~]~

which would be to me new can you maximally 16 trains simultaneously park

The planned comprehensive QUD and information structural analysis of our Stuttgart 21 corpus will proceed in two steps: First, two trained annotators will do a QUD analysis according to the QUD guidelines of Riester, Brunetti & De Kuthy (submitted). For the informational structural analysis we will then use the incremental annotation scheme of Ziai & Meurers (2014), which by explicitly taking questions into account gives instructions of how to determine the extent of the focus by using meaning-based word substitution tests.
WP 6: Question Generation

Since the information structural analysis of authentic data with spelling out QUDs for every single sentence is a very time consuming effort, we will investigate how this annotation process can at least be partially automated. A first step is to automatically generate all possible questions for a single utterance, from which the contextually appropriate QUD can be selected to provide guidance for annotating focus, both manually and using the automatic focus annotation approach developed in the second phase.

Automatically generating questions is a challenging task, since it involves a number of computational linguistic challenges, such as parsing, co-reference resolution and transformation of syntactic structures and requires a deep linguistic knowledge of the syntactic properties of questions of the relevant language, in our case German. A variety of question generation systems have been developed, mostly for educational purposes, such as the assessment of vocabulary knowledge, assisting students in reading (Mazidi & Nielsen 2015), vocabulary learning (Brown et al. 2005, Mostow et al. 2004) or assessment of reading comprehension (cf. Le et al. 2014). While such application based systems are very often designed to only generate a particular, task-specific set of questions, there are also a number of application-neutral, explorative systems which try to generate as many different questions as possible. In terms of methods for question generation, most of the approaches involve a form of transformation, be it based on shallow templates created manually (Liu et al. 2010) or learned from data (Curto et al. 2012), syntax-based transformation rules (Heilman 2011), or transformations based on semantic representations (Mannem et al. 2010, Yao & Zhang 2010, Yao et al. 2012, Chali & Hasan 2012) – with some also integrating discourse cues (Agnarwal et al. 2011). While question generation systems exist for several languages, the majority was developed for English and there is very little research on question generation for German. The only existing system for German (Gütl et al. 2011), focuses on the extraction of concepts from text and reports little on how the questions are actually constructed. To explore the possibilities and particular challenges of question generation in German, we piloted the development and implementation of a transformation-based question generation system for German (Kolditz 2015). The system takes text as input, selects potential answer phrases (NPs, PPs and embedded clauses) based on a syntactic analysis of the input sentences, replaces them with matching question phrases and transforms the syntactic representations of the respective declarative sentences into questions. In the evaluation of his system, Kolditz was able to show that the majority of questions generated for each sentence are of high quality for the limited set of targeted question types, potential answer phrases, and syntactic environments in which they can occur.

Based on these promising pilot results we will extend the empirical reach of Kolditz’s question generation system to cover a broader range of question types and syntactic configurations for which questions can be formulated. Crucially, the nature and formulation of the question for a given sentence we need to be able to ask is constrained by the characteristic of what constitutes potential QUDs. In addition to the material present in a particular sentence, our question generation approach thus also needs to take the immediate context, in which the sentence occurs, into account. Following the Q-Givenness principle made explicit in Riester, Brunetti & De Kuthy (submitted), QUDs can only consist of given (or at least highly salient) material. We therefore plan to incrementally generate questions for a given text sentence by sentence, so that constantly evolving sets of given material can be explicitly modeled. The reading comprehension questions and their information sources identified in the CREG corpus as described in section 3.3.1.3 will serve as a gold standard for evaluating the empirical coverage of our question generation system. The generalizability then will be validated using data from the Stuttgart 21 Corpus, which we already collected, syntactically annotated, and piloted manual QUD annotation for.

For our QUD analysis the challenge will be to select the most appropriate QUD for a sentence from the list of questions that can be generated for that sentence. We plan to narrow down the set of appropriate questions by setting up a crowdsourcing study in which crowd workers select the most appropriate questions for a sentence in a given context. We can then test whether the obtained set of potential QUDs supports reliable manual focus annotation both by experts, using our focus annotation scheme developed in Ziai & Meurers (2014), and non-experts, via crowdsourcing focus annotation as described in De Kuthy et al. (2015). Finally, using the automatic focus annotation approach developed in the second phase, we will test whether the QUDs also support automatic focus classification in authentic data.
3.4.4 Timeline

2017/2
- WP 1: Identifying and annotating task properties
- WP 2: Advancing and generalizing meaning comparison using task properties
- WP 3: Adapting and training content assessment for English workbook tasks
- WP 5: QUD gold standard annotation: Expert annotation
- WP 6: Question Generation: formulation of principles guiding the generation

2018
- WP 1: Annotation of task properties
- WP 1: Identifying information sources automatically
- WP 2: Generalizing meaning comparison
- WP 2: Evaluation of student answers based on information sources
- WP 3: Training content assessment

2019
- WP 4: Identifying the nature of the differences as basis for feedback
- WP 5: Continuation of expert annotation
- WP 6: Implementation of the question generation approach

2020
- WP 1: Annotation of task properties
- WP 2: Integration of task-based normalization of student answers into content assessment
- WP 4: Generation of feedback
- WP 6: Experimental evaluation of the annotated and generated questions

2021/1
- WP 2: Testing and evaluation of the integration of task factors into content assessment
- WP 4: Evaluation of feedback
- WP 6: Experimental evaluation of the generated questions

3.5 Role within the Collaborative Research Centre

In the second phase, fruitful collaboration developed around the methods of analysis used and the empirical validation of information structural concepts. In terms of methods, A3 and A4 actively discussed the use of different distributional semantic approaches and dependency parsing for German and shared resources such as trained models. With project B8 (Stolterfoht), specializing in experimental work on information structural constraints on word order phenomena, a collaboration was established in which we investigated the empirical evidence for focus projection. In De Kuthy & Stolterfoht (to appear), we were able to show that, contrary to the assumption made in the literature, focus accent patterns produced in broad focus contexts are not perceived as ambiguous between a broad and a narrow focus interpretation. This investigation deepened our understanding of the possible extents of focus units in an utterance, which is directly relevant for our annotation of focus in the CREG corpus.

The three research areas envisaged for the third phase in A4 provide a rich basis for collaboration. On the side of the computational methods used, we will continue the collaboration with A3 on the distributional modeling of meaning, where both projects have an interest in modeling at lexical and phrasal levels using the quickly developing state-of-the-art methods from computational linguistics. On the linguistic side, opportunities for collaboration abound with respect to the envisaged generalization of Information Structure to less restricted contexts. Information Structure (and specifically focus) is a central theme in the work of project A7 (Winkler). The theory-driven experiments in A7 and the corpus-based work in A4 will support a focused research exchange, especially in connection with the annotation of QUDs in our WP 5 and the formulation of syntactic constraints necessary for question generation in WP 6. The specification of likely QUDs is also part of the proposed project plan of project B2 (Jäger/Franke) as a basis for their envisaged development of a suitable probabilistic pragmatics model that allows for reasoning about PSPs. We will therefore join forces with B2 in the planned QUD annotation of authentic data as part of our WP 5. Furthermore, we plan to conduct a number of experimental studies in collaboration with project Z2 (Beck/Kaup) in order to empirically validate the contextual appropriateness of the manually annotated and automatically generated QUDs. Z2 will support us in planning, designing and implementing these experiments and in analyzing the results.

As a common data basis for all projects working with authentic data, we compiled and annotated the Stuttgart 21 corpus, which consists of audio files and transcripts of 50 hours recorded speech from the public mediation panel on building a train station (German). We parsed the corpus with a regular and a topological field parser. The syntactically annotated corpus is available on the ANNIS3 http://purl.org/net/sfb833annis platform, which supports the search for specific syntactic constructions. The corpus is accessible to all projects of the SFB and has already been successfully used as data source for the cooperation on extraction phenomena between our project and A7.
3.6 Literature


