Short Answer Assessment in Context: The Role of Information Structure

Dissertation
zur
Erlangung des akademischen Grades
Doktor der Philosophie
in der Philosophischen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Ramon Ziai

aus

Straßburg

2018

Gedruckt mit Genehmigung der Philosophischen Fakultät
der Eberhard Karls Universität Tübingen

Dekan: Prof. Dr. Jürgen Leonhardt

Hauptberichterstatter:	Prof. Dr. Detmar Meurers
Mitberichterstatter:	Prof. Dr. Manfred Stede

Tag der mündlichen Prüfung: 13. Juli 2017

Universitätsbibliothek Tübingen, TOBIAS-lib

Acknowledgments

Writing an acknowledgments section involves recalling many people from different periods of time and how they contributed to one's own work or life. While I think I picked at least a representative (if not complete) set of people below, I apologize in advance for everyone I forgot to name here.

I begin by thanking my advisor and academic mentor, Detmar Meurers. Detmar's unique combination of enthusiasm and ability of seeing and drawing connections between seemingly different strands of research is primarily responsible for the fact that I remained in Tübingen for my PhD. Over the years, I have benefited greatly from his way of thinking and working, and all the creative discussions we had, not to mention the funding he acquired in connection with the SFB 833, which made this thesis and other work possible in the first place.

I also thank Manfred Stede, whom I basically ambushed after one of his classes at ESSLLI 2016 in order to convince him to be my external reviewer and defense examiner. Within five minutes, he had agreed (this should be a Guinness record for academic straight-forwardness), making my life easier in a stressful time and providing me with a thorough and critical new perspective on my thesis. In connection with the defense, I am also very grateful to Jonas Kuhn, Fritz Hamm and Harald Baayen for taking time out of their busy schedules to examine my thesis work and enable me to graduate.

Moving on to other important collaborators, I am very grateful to Niels Ott and Kordula De Kuthy for working with me in the different stages of the A4 project of SFB 833. Niels shares a significant amount of my history at the SfS in Tübingen, collaborating with me on a number of student projects, programming marathons and later, research undertakings. Kordula's contribution in A4 based on her expertise in theoretical linguistics and information structure made sure there was a solid theoretical and empirical foundation for me to count on. Concerning the latter, I also want to express gratitude towards Arndt Riester for the interesting discussions we had on defining and annotating focus.

Annotation projects need trained annotators, and I count myself lucky to have been able to rely on the work of Kordula De Kuthy, Tobias Kolditz, Zarah Weiß, Cornelius Fath, Heike Cardoso, Stefanie Wolf and Philip Schulz in different stages of focus annotation. In a similar vein, any project needs competent

administration, and here I am very grateful to Beate Starke for continuously making all of our lives easier.

No thesis should be submitted without someone else reading it first, and thankfully, I had the help of a number of colleagues here: Adriane Boyd and Kordula De Kuthy read multiple chapters of an earlier version, providing invaluable feedback, and Björn Rudzewitz, Willy Cheung and Piklu Gupta also made very helpful suggestions.

In terms of writing motivation, I want to thank my "lunch gang" for continuously asking about my progress, even (and especially) when I did not want to be asked: Jochen Saile, Sowmya Vajjala, Serhiy Bich, Petra Augurzky and Adriane Boyd, you gave me a fixed point in the day where I could get my head out of work and talk about other important matters, such as music, kids, culture and who made the best lunch choice.

I thank all the people sharing the same floor with me for the random friendly conversations about shared suffering, recent sucesses, upcoming vacation and other things. Special thanks go out to my fellow PhD companions Sowmya Vajjala, Serhiy Bich, and all the fellow PhD students from the SFB 833. I also thank my ex-colleagues Niels Ott and Holger Wunsch for sharing the occasional beer with me to talk about life, the universe and everything.

Moving outside of work, I am very grateful to all the musicians I have had the pleasure of playing with. They are too numerous to list exhaustively, but let me mention the bands 'Secret Asian Men', 'Tapas', 'Yellow Moon Orchestra' and 'Bazaar'. Special thanks go to my friends Thorsten Hans and Dirk Wassen for fun times and open ears. Music is closely related to dance, and I also want to express my gratitude to all those that improved my life by sharing great dances and good times with me at Salsa and Kizomba parties, especially to my dance partner and friend Elena Ciuffreda, who always had an open ear for me. I also thank Melissa Castillo for good times and tasty cupcakes and Sureia Ionescu for funny, relaxed evenings; among many other people.

Finally, I want to thank the people who apart from myself suffered most under my stress during the final writing period: Vassilena Terziyska deserves the main credit here for being there for me despite going through a challenging time herself. Above all, I also want to express my heartfelt gratitude to my parents Petra and Tamim, and my brother David, for everything.

Summary

Short Answer Assessment (SAA), the computational task of judging the appropriateness of an answer to a question, has received much attention in recent years (cf., e.g., Dzikovska et al. 2013; Burrows et al. 2015). Most researchers have approached the problem as one similar to paraphrase recognition (cf., e.g., Brockett & Dolan 2005) or textual entailment (Dagan et al., 2006), where the answer to be evaluated is aligned to another available utterance, such as a target answer, in a sufficiently abstract way to capture form variation. While this is a reasonable strategy, it fails to take the explicit context of an answer into account: the question.

In this thesis, we present an attempt to change this situation by investigating the role of Information Structure (IS, cf., e.g., Krifka 2007) in SAA. The basic assumption adapted from IS here will be that the content of a linguistic expression is structured in a non-arbitrary way depending on its context (here: the question), and thus it is possible to predetermine to some extent which part of the expression's content is relevant. In particular, we will adopt the Question Under Discussion (QUD) approach advanced by Roberts (2012) where the information structure of an answer is determined by an explicit or implicit question in the discourse.

We proceed by first introducing the reader to the necessary prerequisites in chapters 2 and 3. Since this is a computational linguistics thesis which is inspired by theoretical linguistic research, we will provide an overview of relevant work in both areas, discussing SAA and Information Structure (IS) in sufficient detail, as well as existing attempts at annotating Information Structure in corpora. After providing the reader with enough background to understand the remainder of the thesis, we launch into a discussion of which IS notions and dimensions are most relevant to our goal. We compare the *given/new* distinction (information status) to the *focus/background* distinction and conclude that the latter is better suited to our needs, as it captures requested information, which can be either given or new in the context.

In chapter 4, we introduce the empirical basis of this work, the Corpus of Reading Comprehension Exercises in German (CREG, Ott, Ziai & Meurers 2012). We outline how as a task-based corpus, CREG is particularly suited to the analysis of language in context, and how it thus forms the basis of our

efforts in SAA and focus detection. Complementing this empirical basis, we present the SAA system CoMiC in chapter 5, which is used to integrate focus into SAA in chapter 8.

Chapter 6 then delves into the creation of a gold standard for automatic focus detection. We describe what the desiderata for such a gold standard are and how a subset of the CREG corpus is chosen for manual focus annotation. Having determined these prerequisites, we proceed in detail to our novel annotation scheme for focus, and its intrinsic evaluation in terms of inter-annotator agreement. We also discuss explorations of using crowd-sourcing for focus annotation.

After establishing the data basis, we turn to the task of automatic focus detection in short answers in chapter 7. We first define the computational task as classifying whether a given word of an answer is focused or not. We experiment with several groups of features and explain in detail the motivation for each: syntax and lexis of the question and the answer, positional features and givenness features, taking into account both question and answer properties. Using the adjudicated gold standard we established in chapter 6, we show that focus can be detected robustly using these features in a word-based classifier in comparison to several baselines.

In chapter 8, we describe the integration of focus information into SAA, which is both an extrinsic testbed for focus annotation and detection per se and the computational task we originally set out to advance. We show that there are several possible ways of integrating focus information into an alignment-based SAA system, and discuss each one's advantages and disadvantages. We also experiment with using focus vs. using givenness in alignment before concluding that a combination of both yields superior overall performance.

Finally, chapter 9 presents a summary of our main research findings along with the contributions of this thesis. We conclude that analyzing focus in authentic data is not only possible but necessary for a) developing context-aware SAA approaches and b) grounding and testing linguistic theory. We give an outlook on where future research needs to go and what particular avenues could be explored.

Zusammenfassung

Short Answer Assessment (SAA), die computerlinguistische Aufgabe mit dem Ziel, die Angemessenheit einer Antwort auf eine Frage zu bewerten, ist in den letzten Jahren viel untersucht worden (siehe z.B. Dzikovska et al. 2013; Burrows et al. 2015). Meist wird das Problem analog zur Paraphrase Recognition (siehe z.B. Brockett & Dolan 2005) oder zum Textual Entailment (Dagan et al., 2006) behandelt, indem die zu bewertende Antwort mit einer Referenzantwort verglichen wird. Dies ist prinzipiell ein sinnvoller Ansatz, der jedoch den expliziten Kontext einer Antwort außer Acht lässt: die Frage.

In der vorliegenden Arbeit wird ein Ansatz dargestellt, diesen Stand der Forschung zu ändern, indem die Rolle der Informationsstruktur (IS, siehe z.B. Krifka 2007) im SAA untersucht wird. Der Ansatz basiert auf der grundlegenden Annahme der IS, dass der Inhalt eines sprachlichen Ausdrucks auf einer bestimmte Art und Weise durch seinen Kontext (hier: die Frage) strukturiert wird, und dass man daher bis zu einem gewissen Grad vorhersagen kann, welcher inhaltliche Teil des Ausdrucks relevant ist. Insbesondere wird der Question Under Discussion (QUD) Ansatz (Roberts, 2012) übernommen, bei dem die Informationsstruktur einer Antwort durch eine explizite oder implizite Frage im Diskurs bestimmt wird.

In Kapitel 2 und 3 wird der Leser zunächst in die relevanten wissenschaftlichen Bereiche dieser Dissertation eingeführt. Da es sich um eine computerlinguistische Arbeit handelt, die von theoretisch-linguistischer Forschung inspiriert ist, werden sowohl SAA als auch IS in für die Arbeit ausreichender Tiefe diskutiert, sowie ein Überblick über aktuelle Ansätze zur Annotation von IS-Kategorien gegeben. Anschließend wird erörtert, welche Begriffe und Unterscheidungen der IS für die Ziele dieser Arbeit zentral sind: Ein Vergleich der *given/new*-Unterscheidung und der *focus/background*-Unterscheidung ergibt, dass letztere das relevantere Kriterium darstellt, da sie erfragte Information erfasst, welche im Kontext sowohl gegeben als auch neu sein kann.

Kapitel 4 stellt die empirische Basis dieser Arbeit vor, den Corpus of Reading Comprehension Exercises in German (CREG, Ott, Ziai & Meurers 2012). Es wird herausgearbeitet, warum ein task-basiertes Korpus wie CREG besonders geeignet für die linguistische Analyse von Sprache im Kontext ist, und dass es daher die Basis für die in dieser Arbeit dargestellten Untersuchungen zu SAA

und zur Fokusanalyse darstellt. Kapitel 5 präsentiert das SAA-System CoMiC (Meurers, Ziai, Ott & Kopp, 2011b), welches für die Integration von Fokus in SAA in Kapitel 8 verwendet wird.

Kapitel 6 befasst sich mit der Annotation eines Korpus mit dem Ziel der manuellen und automatischen Fokusanalyse. Es wird diskutiert, auf welchen Kriterien ein Ansatz zur Annotation von Fokus sinnvoll aufbauen kann, bevor ein neues Annotationsschema präsentiert und auf einen Teil von CREG angewendet wird. Der Annotationsansatz wird erfolgreich intrinsisch validiert, und neben Expertenannotation wird außerdem ein Crowdsourcing-Experiment zur Fokusannotation beschrieben.

Nachdem die Datengrundlage etabliert wurde, wendet sich Kapitel 7 der automatischen Fokuserkennung in Antworten zu. Nach einem Überblick über bisherige Arbeiten wird zunächst diskutiert, welche relevanten Eigenschaften von Fragen und Antworten in einem automatischen Ansatz verwendet werden können. Darauf folgt die Beschreibung eines wortbasierten Modells zur Fokuserkennung, welches Merkmale der Syntax und Lexis von Frage und Antwort einbezieht und mehrere Baselines in der Genauigkeit der Klassifikation klar übertrifft.

In Kapitel 8 wird die Integration von Fokusinformation in SAA anhand des CoMiC-Systems dargestellt, welche sowohl als extrinsische Validierung von manueller und automatischer Fokusanalyse dient, als auch die computerlinguistische Aufgabe darstellt, zu der diese Arbeit einen Beitrag leistet. Fokus wird als Filter für die Zuordnung von Lerner- und Musterantworten in CoMiC integriert und diese Konfiguration wird benutzt, um den Einfluss von manueller und automatischer Fokusannotation zu untersuchen, was zu positiven Ergebnissen führt. Es wird außerdem gezeigt, dass eine Kombination von Fokus und Givenness bei verlässlicher Fokusinformation für bessere Ergebnisse sorgt als jede Kategorie in Isolation erreichen kann.

Schließlich gibt Kapitel 9 nochmals einen Überblick über den Inhalt der Arbeit und stellt die Hauptbeiträge heraus. Die Schlussfolgerung ist, dass Fokusanalyse in authentischen Daten sowohl möglich als auch notwendig ist, um a) den Kontext in SAA einzubeziehen und b) linguistische Theorien zu IS zu validieren und zu testen. Basierend auf den Ergebnissen werden mehrere mögliche Richtungen für zukünftige Forschung aufgezeigt.

Related Publications

Parts of this thesis also appear in the following peer-reviewed publications:

1. Ziai, Ramon & Detmar Meurers. 2018. Automatic Focus Annotation: Bringing Formal Pragmatics Alive in Analyzing the Information Structure of Authentic Data. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, New Orleans, LA: ACL. URL `http://www.sfs.uni-tuebingen.de/~rziai/papers/Ziai.Meurers-18.pdf`. To appear.

2. Ziai, Ramon, Kordula De Kuthy & Detmar Meurers. 2016. Approximating Givenness in Content Assessment Through Distributional Semantics. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM)*, 209–218. Berlin, Germany: ACL. URL `http://aclweb.org/anthology/S16-2026.pdf`

3. De Kuthy, Kordula, Ramon Ziai & Detmar Meurers. 2016b. Focus Annotation of Task-Based Data: Establishing the Quality of Crowd Annotation. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, 110–119. Berlin, Germany: ACL. URL `http://aclweb.org/anthology/W16-1713.pdf`

4. De Kuthy, Kordula, Ramon Ziai & Detmar Meurers. 2016a. Focus Annotation of Task-Based Data: a comparison of expert and crowd-sourced annotation in a reading comprehension corpus. In *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference (LREC)*, 3928–3934. Portorož, Slovenia. URL `http://www.lrec-conf.org/proceedings/lrec2016/pdf/1083_Paper.pdf`

5. Ziai, Ramon & Detmar Meurers. 2014. Focus Annotation in Reading Comprehension Data. In *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII, 2014)*, 159–168. COLING Dublin, Ireland: ACL. URL `http://aclweb.org/anthology/W14-4922.pdf`

6. Ziai, Ramon, Niels Ott & Detmar Meurers. 2012. Short Answer Assessment: Establishing Links Between Research Strands. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7) at NAACL-HLT 2012*, 190–200. Montreal. URL `http://aclweb.org/anthology/W12-2022.pdf`

7. Ott, Niels, Ramon Ziai & Detmar Meurers. 2012. Creation and Analysis of a Reading Comprehension Exercise Corpus: Towards Evaluating Meaning in Context. In Thomas Schmidt & Kai Wörner (eds.), *Multilingual Corpora and Multilingual Corpus Analysis* Hamburg Studies in Multilingualism (HSM), 47–69. Amsterdam: Benjamins. URL `https://benjamins.com/#catalog/books/hsm.14.05ott`

8. Meurers, Detmar, Ramon Ziai, Niels Ott & Janina Kopp. 2011b. Evaluating Answers to Reading Comprehension Questions in Context: Results for German and the Role of Information Structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, 1–9. Edinburgh. URL `http://aclweb.org/anthology/W11-2401.pdf`

9. Meurers, Detmar, Ramon Ziai, Niels Ott & Stacey Bailey. 2011a. Integrating Parallel Analysis Modules to Evaluate the Meaning of Answers to Reading Comprehension Questions. *IJCEELL. Special Issue on Automatic Free-text Evaluation* 21(4). 355–369. URL `http://www.inderscience.com/info/inarticle.php?artid=42793`

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

Funding

The research presented in this dissertation has been funded by the Deutsche Forschungsgemeinschaft through project A4 (‘Comparing Meaning in Context’) of Collaborative Research Center 833 (‘The Construction of Meaning’), during phase I (2009/2 – 2013/1) and phase II (2013/2 – 2017/1) of the project.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
CAM	Content Assessment Module
CAS	Common Analysis Structure
CL	Computational Linguistics
CoMiC	Comparing Meaning in Context
CREE	Corpus of Reading Comprehension Exercises in English
CREG	Corpus of Reading Comprehension Exercises in German
CRF	Conditional Random Field
DP	determiner phrase
IDF	Inverse Document Frequency
IS	Information Structure
NER	Named Entity Recognition
NLP	Natural Language Processing
POS	part of speech
PP	prepositional phrase
QAC	Question-Answer Congruence
QA	Question Answering
QUD	Question Under Discussion
RTE	Recognizing Textual Entailment
SAA	Short Answer Assessment
SLA	Second Language Acquisition
SVM	Support Vector Machine
TF	Term Frequency
TMA	Traditional Marriage Algorithm
UIMA	Unstructured Information Management Architecture
URL	Uniform Resource Locator
WELCOME	Web-based Learner Corpus Machine
XML	eXtensible Markup Language

1. Introduction

1.1. Motivation

Communication is arguably one of life's most important aspects and greatest challenges. We use language to communicate, which can be thought of as a system of signs that allows us to convey a potentially unlimited number of facts. However, these facts are not conveyed in isolation: there is usually some information required by a certain situation or question. For example, when someone says *It is four o' clock*, the situation could be that someone else asked a question such as *What is the time?*, where a time expression is required. In other words, humans use language to fulfill certain functional goals.

From this example, we can see that the functional goal imposes requirements on a) the content expressed in the question and b) what an acceptable answer to the question looks like. These contextual constraints make discourse processing easier, because the participants of a conversation can build up reasonable expectations about what kind of content someone is going to provide: a question such as *Did you see the game yesterday?* is not likely to be answered with *Berlin is the capital of Germany*, but rather with a simple *yes* or *no*. Forming expectations on utterances is especially necessary in settings where the forms used to express content deviate from standard language, as is for example the case with colloquial language and learner language. The latter is particularly interesting in this regard, because while learners usually know exactly what content they want to express, they often do not have sufficient command of the target language to produce a correct version of what they want to say. Knowing what kind of content they are supposed to provide can facilitate comprehending what they say even when the form is erroneous. As an example, consider a situation where a tourist on the street asks *How get train station?* and even though he should have said *How do I get to the train station?*, it is possible to determine what he wants to know.

In the case of meaning-focused tasks in language learning, teachers apply strategies in assessing the language a learner produces: because they know what the task requires and what kind of content the learner needs to provide, they are able to better interpret what the learner says. A typical example of such a setting is reading comprehension, where the learner needs to answer questions about a text in order to demonstrate their understanding of the text's content. One such task is depicted in Figure 1.1.

The text describes the situation of women in the music industry. The question given here asks what the success or failure of a female artist is based on, and the answer addresses that question by paraphrasing a part of the content given in the text on that topic.

We can see that there are several components in reading comprehension tasks. First, all relevant content is usually encoded in the text which serves as a knowledge base for the task. As we will see later, it is also possible for a reading comprehension task to be partly based on external knowledge, but for now we will focus on text-based questions. Second, the question asks for a specific piece of content based on the text by picking up a topic, in this case the success of female artists, and formulating an information requirement based on it. Third, the answer attempts to address the information requirement formulated by the question and to provide the necessary content.

If one were to assess the answer's correctness, one has several options. Either one validates the given answer's content against the facts in the text under the perspective introduced by the question. This approach, however, involves a certain amount of searching in the text in order to identify the necessary pieces of knowledge which the answer then needs to be compared to. The approach typically taken by language teachers is therefore to construct a reference answer to each question, and compare learner answers to that reference answer. Suppose the reference answer for the exercise in Figure 1.1 is *The success or failure of a female artist is based largely on her physical appearance and gendered performance style*, which appears verbatim in the text. Then the task of assessing whether the answer given in Figure 1.1 becomes one of comparing the following two answers:

(1) The success or failure of a female artist is based largely on her physical appearance and gendered performance style.

Text: Although the twentieth century saw the rise of women as professional musicians, the majority of composers and performers were, and still are, men. The music industry in the U.S. and Britain overwhelmingly reflects the values of a patriarchal society; the success or failure of a female artist is based largely on her physical appearance and gendered performance style. Blues, rock, and pop began as genres dominated by men, and thus included styles of dress, lyrics, and sound born of a male perspective. The history of these genres, then, is also a history of women seeking to locate their space within a predominantly masculine musical environment.

Women are always judged, in part, on their image, and it is through the manipulation of this image that some women artists have been able to push the boundaries of gender identity. Women have been able to enter popular genres of music either by playing with the aesthetics of masculinity, or by playing into a male expectation of femininity. Sexuality, therefore, is a tool women continue to use to shape and reshape their place within popular music.

Pushing boundaries is a balancing act, however, and a contradictory process. In order to gain access to the world of popular music, a female artist must at once be pleasing her audience, and, at the same time, remain true to herself as a woman. A desire to be too much "one of the guys" can lead to identity problems and ultimately to self-destruction. An artist's use of irony or parody may run the risk of being mistaken for genuineness, causing her to be objectified. Working within the limits of popular music has proven difficult and dangerous for women. But due to the professionalism and inventiveness of many female performers, the space for women in popular music is being expanded and redefined.

Question: According to the text, what is the success or failure of a female artist based on?

Answer: The success is based on the way she looks.

Figure 1.1.: Reading comprehension example from `https://www.800score.com/content/gre/rce2.html`, adapted

(2) The success is based on the way she looks.

We observe that (2) is shorter than (1) and does not mention the phrases *failure*, *female artist*, *her physical appearance* and *gendered performance style*. Instead, there is a new phrase, *the way she looks*, which in the context here can be seen as synonymous to *her physical appearance*. The other concepts are not present, but are all of them necessary in order to answer the question?

In order to answer that, let us take a closer look at (2). We can see that there is a part of it that picks up the question (*The success is based on*) and a part that introduces information not present in the question (*the way she looks*). The first part connects the answer thematically to the question, while the second attempts to provide the requested content. A similar structuring is in place for (1): *The success or failure of a female artist is based largely on* reprises the question and *her physical appearance and gendered performance style* provides the requested content. So it seems we can answer our earlier question whether all concepts in (1) are necessary with "no", since there is clearly a difference in relevance between requested content and repeated material (e.g. *failure*, *female artist*) with respect to the question.

Based on the example above, one could hypothesize that the important criterion for assessing the answer is whether it provides new content. After all, removing the repeated content in the answers above left us exactly with the relevant part. However, it turns out that this fails to account for new material which is not relevant to the question and repeated content that is relevant to the question (as in *or*-questions). Consider the following alternative answer to the question in Figure 1.1:

(3) In the U.S. and British music industry, which reflects the values of a patriarchal society, the success of a female artist depends on her lyrics and sound.

In (3), there is again a part that reprises question material (*the success of a female artist depends on*) and a part that addresses the question (*her lyrics and sound*), albeit providing the wrong content according to the reading text. However, there is now a large remaining part which is new: *In the U.S. and British music industry, which reflects the values of a patriarchal society,...* This part is clearly not requested by the question, so it can not count towards answering it. In order to systematically identify and exclude such material from answer comparison, we need to examine the question and its properties more closely.

Looking at the question (*According to the text, what is the success or failure of a female artist based on?*), let us recall that a question also exhibits a structuring in terms of content: there is a part that picks up a topic from the text and a part that formulates what information is required in connection with that topic. In *wh*-questions such as the one we are dealing with here, it is a reasonable assumption that the phrase starting with the *wh*-word is responsible for requesting information. Indeed, when considering both the *wh*-phrase *what* and the previously identified relevant answer parts *her physical appearance and gendered performance style* in (1) and *the way she looks*, there seems to be a close relationship between the two.

Based on the above observations and our concrete example, we can formulate a first version of our mission statement in the following way: given a question such as *According to the text, what is the success or failure of a female artist based on?* and an answer such as *The success is based on the way she looks*, we want to identify properties of the question and answer that allow us to characterize how exactly one can predict that the relevant part in the answer is *the way she looks*, in order to use this prediction to assess the appropriateness of the answer.

1.2. Research Questions and Goals

This thesis deals with three inter-dependent research questions:

1. How can question-answer relationships be analyzed systematically in real-life language data?
2. How can such an analysis support an automatic detection of requested content in answers?
3. What impact do analysis and detection of requested content have on determining the appropriateness of answers to questions?

In terms of the linguistic research areas concerned, the goal of this thesis is two-fold: On the one hand, we aim to advance **explicit linguistic modeling of Information Structure (IS) in authentic data** by operationalizing current IS theories and applying them to reading comprehension data, a real-life case of question-answer interactions supporting such operationalization. In our work, this involves both manual annotation and automatic detection of IS notions. On the other hand, we want to study the **impact of explicit information-structural modeling on the task of Short Answer Assessment (SAA)**, both for the purpose of externally grounding and evaluating our IS analysis and to investigate whether IS analysis has the potential to improve the state of the art in computationally assessing answers to questions.

1.3. Contributions

This thesis makes the following specific contributions:

1. **Annotation and analysis of focus**: we develop a new approach to annotating and analyzing focus in authentic data. The approach builds on current meaning-based views of focus and operationalizes them in an incremental annotation scheme, resulting in substantial inter-annotator agreement ($\kappa = .7$) on a data set involving non-wellformed language.

2. **Crowd annotation of focus:** we demonstrate that focus annotation is feasible using non-experts, i.e., ordinary speakers of a language, establishing the quality of crowd annotation both by comparing it to our expert annotation and by independently predicting it using a new measure we define.

3. **Automatic focus detection:** building on our successful focus annotation work, we present the first automatic focus detection approach for German to our knowledge. It combines a range of linguistically well-motivated features based on both questions and answers and reaches 78.1% accuracy (majority baseline : 58.1%) in predicting focus vs. background on the token level.

4. **Extrinsic evaluation in Short Answer Assessment:** we show that focus can be integrated into alignment-based SAA systems as a filter and perform an extrinsic evaluation of manual and automatic annotation within the CoMiC system, revealing that both manually and automatically determined focus have the potential to result in quantitative gains.

5. **Corpus collection software:** we present the WELCOME system to which we significantly contributed, a web-based application which enables distributed data entry for the purpose of creating richly structured reading comprehension corpora, and show how this system was used to collection a large German corpus.

6. **SAA system:** we present CoMiC, an alignment-based SAA system to which we significantly contributed, which achieves state-of-the-art performance for both English and German. It also provides the basis for our integration of focus into SAA.

7. **Reading comprehension corpus:** we present CREG, the largest reading comprehension corpus publicly available, and one of the few data sets publicly available to SAA researchers in general. It contains more than 35,000 student answers and 1,600 target answers to over 1,500 questions on approximately 150 reading texts.

8. **Focus-annotated reading comprehension corpus:** our main focus annotation effort resulted in CREG-ExpertFocus, a data set of 4,177 answers to corresponding reading comprehension questions annotated with our focus scheme.

9. **Crowd-sourced focus annotation corpus:** Complementing the expert-annotated corpus, our experiments on crowd-sourcing focus annotation resulted in two annotated reading comprehension data sets with more than 5,500 and 3,300 answers, respectively.

1.4. Thesis Overview

In this section, we give an overview of the thesis in terms of the parts and chapters it contains.

Part I: Background

Part I provides the reader with all necessary background information, both from a linguistic and from a computational perspective.

Chapter 2 introduces the task and the field of Short Answer Assessment and discusses the challenges involved. We discuss related work and give an overview of some of the approaches that currently exist in the field. We also discuss how systems differ, taking special note of the systems which involve some notion of task context. We also present the data sets currently available to researchers.

Chapter 3 introduces the theoretical linguistic field of Information Structure. After giving a brief overview of the IS distinctions discussed in the literature, we discuss previous work on annotating these distinctions in corpora. We then zoom in on two dimensions: given/new and focus/background, before concluding that focus is most relevant for our research questions.

Part II: Our Empirical Basis and Experimental Sandbox

Part II introduces the corpus and the SAA system which form the basis for our efforts and experiments.

Chapter 4 introduces the Corpus of Reading Comprehension Exercises in German (CREG), from which we draw the data for both SAA and focus analysis. We motivate why such a richly structured task-based corpus is necessary for studying language in context. We outline how the corpus was collected and what exactly it contains in terms of quality and quantity. Finally, we characterize the subsets of CREG that are used in evaluation and other contexts.

Chapter 5 presents our SAA system CoMiC for assessing the meaning of answers, which provides the experimental sandbox for exploring the impact of focus information in SAA. We outline the conceptual basis of the system including its three stages (annotation, alignment, classification). We then show how the system was implemented using the Unstructured Information Management Architecture (UIMA) framework, enabling a straightforward transfer from English to German, and report the performance it achieves on different CREG subsets.

Part III: Focus

The final part of this thesis deals with our main contributions, in terms of annotating, detecting and integrating focus.

Chapter 6 describes our manual focus annotation effort. Pointing out the issues that arise when only relying on surface criteria in focus annotation, we develop an iterative approach built on meaning-based criteria, which consists of a) making the question form explicit, b) determining the set of alternatives and c) marking the extent of the focus in the answer, for which we introduce an explicit form-based test. We describe two rounds of the annotation effort resulting in substantial inter-annotator agreement. Finally, we explore crowd-sourcing as a way of obtaining focus annotation and describe a concrete crowd-sourcing experiment, showing that at least for some question types, the crowd reaches the level of expert annotations.

Chapter 7 deals with developing an automatic approach to focus identification built on the annotation results from chapter 6. We first discuss what constitutes observable linguistic evidence relevant for focus detection before discussing our resulting initial feature set and how we use it in a classifier to detect focus. After reporting initial promising results, we launch into a qualitative evaluation of typical classifier behavior. We then present related work in focus detection and finally discuss three extensions to our model before presenting the final improved results.

Chapter 8 finally tackles the integration of focus information into our CoMiC system for SAA. We discuss several ways in which focus could be used in the system architecture, before settling on the "focus as filter" approach where only linguistic units that are focused can be aligned. Having decided on the method, we provide a thorough evaluation of manually and automatically determined focus information within SAA. Results for manually annotated focus information show that focus significantly outperforms non-focus SAA system variants. Using automatically determined focus information, we obtain an improvement in the case where the system has not seen the questions in the test set before and thus can benefit also from noisy focus information.

Part IV: Conclusion

Chapter 9 concludes the thesis by first summarizing it in detail on a part-by-part basis. Based on the summary, we outline our contributions to the fields of Information Structure and Short Answer Assessment and end by discussing avenues for future research.

Part I.

Background

2. Computational Linguistics: Short Answer Assessment

In this chapter, we introduce the computational part of the background for this thesis: Short Answer Assessment (SAA), which is the main computational task we aim to improve. We first define what SAA is (section 2.1) and what challenges relate it to and separate it from related computational linguistic tasks (section 2.2). We then briefly discuss how SAA systems are evaluated (section 2.3). In line with the research questions of this thesis, we then define the criterion of context-awareness for distinguishing SAA approaches that make use of the context in classifying answers from those that do not (section 2.4), before using this distinction to structure our overview of approaches[1](section 2.5). Finally, we list the most important data sets that are available to date (section 2.6).

2.1. What is Short Answer Assessment?

As the name implies, Short Answer Assessment refers to the task of evaluating an answer of limited length in some way. The **length** of the answers varies across data sets, task types and student populations, but can roughly be specified as 1–3 sentences. The term **answer** implies that what is being evaluated was uttered in response to an explicit question or prompt. In addition, by answer we mean a natural language answer to a natural language question. Finally, **assessment** usually means dealing out some kind of judgment, and because questions usually ask for content, it is the meaning of the answer that we are primarily interested in assessing, not its form.

However, it is hard for both humans and machines to evaluate meaning without being able to compare it to some reference. Therefore, an additional

[1] The overview partly draws on Ziai et al. (2012).

component is needed which constitutes the context that the questions are about, such as a picture, an audio recording or a text. In this thesis, we concern ourselves mainly with questions about texts, because here the content is linguistically encoded. Pictures and audio, on the other hand, require non-linguistic means of extracting knowledge.

In addition, since it greatly facilitates the evaluation task, one often constructs explicit reference answers that candidate answers can be compared to. This is done both to reduce the complexity of assessment, since one does not have to search the text, and to make assessment more consistent, since one makes explicit what a correct answer should look like. The example in Figure 2.1 illustrates the SAA setting.

Text: Failing to check the facts
There have been some embarrassing examples where major newspapers and TV networks have published false information because reporters have not checked for accuracy. One such example was the publication of a report of the death of the elderly comedian, Bob Hope. A U.S. Congressman apparently misheard someone talking about Bob Hope. He stood up in Congress and announced the death of the comedian. This was then picked up and published widely in the media. When reporters called Mr. Hope's home to follow up the story, his daughter was very surprised and assured them that he was at the moment happily eating his breakfast.

Question: Where was Bob Hope when he heard about the news?

Ref. answer: Bob Hope was at home.

Learner answer: He was in his house.

Figure 2.1.: Example text, question, reference and learner answer in a reading comprehension task (Bailey & Meurers, 2008).

The example shows a fragment of a text about the comedian Bob Hope, who was wrongly pronounced dead by a U.S. Congressman. The question asks about Bob Hope's location at the time of the statement. The reference and learner answers both express the fact that he was at home, yet do so in different ways: where the reference answer uses the full name, the learner answer uses a pronoun to refer to Bob Hope. Moreover, in this particular context, *in his house* appears to be synonymous with *at home*.

2.2. Challenges

The example above demonstrates the key challenge of SAA: there are multiple ways of expressing appropriate content, all of which should ideally be accounted for. Depending on how open the question is, the theoretical **space of acceptable answers in terms of content** (cf., e.g., Quixal & Meurers 2016) can theoretically be infinite. In practice, content evaluation of completely open questions, such as essay questions, does not occur because even humans need to know what a correct answer should look like in order to rate a candidate answer. However, even in more constrained settings, a question may be answered correctly in different ways, as example (4) illustrates.

(4) Q: The text describes the health insurance system in the US. How does it compare to the one in your home country?

A1: In Germany, the situation is as follows...

A2: I come from Spain, and we have better health insurance than the Americans.

But even if the content is equivalent, this has to be recognized first. We are dealing with natural language expressions, and one prevalent and challenging characteristic of natural language expressions is **form variation**. It can occur on the lexical level, where the variation consists of a different word choice, or on the syntactic level, where a different construction is chosen.

(5) Q: What did Peter do at the party after he drank too much beer?

A1: He *started* to sing.

A2: He *began* to sing.

(5) is an example for lexical variation, where the verb *started* is substituted for the verb *began*. Both words are synonymous with respect to the given situation, so either is acceptable in the answer. Note that synonymy is heavily dependent on context, e.g. *to expire* and *to die* are synonymous when talking

about people ('Sadly, he died/expired') but not with respect to passports ('My passport *died/expired').

(6) Q: What happened to Mark in the street today?
A1: *He was hit by* a car.
A2: A car *hit him*.

(6) is an example of active-passive alternation as one instance of syntactic variation. In this case, A1 uses the passive voice (*He was hit by*...), whereas A2 uses the active voice (... *hit him*). Both express exactly the same content using the same lexical material.

Related Fields

Naturally, there are CL fields that share some of the challenges above with SAA. The fields of Recognizing Textual Entailment (RTE) and Paraphrase Recognition come to mind first, since they most directly share the problem of meaning comparison with SAA: In Paraphrase Recognition (cf., e.g., Brockett & Dolan 2005), the task is to detect whether two utterances *H* and *T* express the same meaning, whereas RTE (cf., e.g., Dagan et al. 2009) takes this problem one step further by attempting to also detect whether the meaning of *H* can be inferred from the meaning of *T*. A related task is Text Simplification (cf., e.g., Chandrasekar et al. 1996), where the meaning of *T* needs to be preserved in *H*, but expressed in less complex form. Finally, the field of query-based summarization (cf., e.g., Daumé III & Marcu 2006) also aims to preserve some of the meaning of *T* in *H*, focusing on the main points with respect to a specific query.

SAA differs from these fields in several aspects. First, it contextualizes meaning comparison by embedding it in a concrete task context. Where RTE and Paraphrase Recognition are designed to be solved out of the broader context of a concrete application, SAA explicitly provides a scenario that the result of meaning comparison is to be used for: judging the appropriateness of an answer to a question. For the latter, naturally occurring human gold standard judgments are typically available, for example in the form of teacher ratings. It thus becomes possible to evaluate meaning comparison extrinsically by measuring the performance of an SAA system against the human gold standard, instead of having to create an artificial gold standard for meaning comparison in isolation. This is also true for the nature of the gold standard ratings: since the criteria on which an answer in SAA is evaluated are based on real-world needs, for example the need to grade homework assignments, they tend to be less artificial than intrinsic evaluation criteria.

Second, SAA often presents the additional challenge of dealing with ill-formed input, depending on the domain in which it is performed. In second language learning settings, for example, one frequently has to deal with learner errors in both content and form which can increase the difficulty of meaning evaluation, since it adds a dimension of **ill-formed variation** (cf., e.g., Meurers & Dickinson 2017, sec. 2.2) to the list of challenges to tackle. (7) is an example from the Corpus of Reading Comprehension Exercises in English (CREE) (Bailey, 2008) showing ill-formed variation. The text on which the question is based describes the influence of violence in television programs on children.

(7) Q: What seems to be missing from this type of programming?

SA: Be watching violence we become less sensitive to it

The student answer (SA), besides not properly addressing the given question, also contains a form error: *Be* was used instead of *By*, which is especially problematic since both are valid words of English, but only the latter is grammatical in this sentence.

Finally, SAA offers the possibility of studying answers in context of explicit questions. From a linguistic point of view, this is interesting because it offers an authentic data source for testing theories of Information Structure (IS), a central theme of this thesis which we will discuss in greater detail from chapter 6 onwards.

2.3. Evaluating SAA Systems

In this section, we briefly discuss how SAA systems are evaluated. Two dimensions are relevant here: the **evaluation setting**, which pertains to the nature of the training and testing data, and the **evaluation metrics**, with which success is measured quantitatively.

2.3.1. Evaluation Settings

The evaluation setting controls what kind of data a system will be tested on. On the one hand, there are general considerations in machine learning tasks, such as the widely accepted fact that training and testing data need to be distinct. Moreover, many approaches use a development set in addition to the training and testing set, which is used for tuning and optimizing machine learning parameters. Others use the **cross-validation** approach (cf., e.g., Kohavi 1995) for evaluation, which does not employ a designated development or test set. Instead, the training data is split into n folds (parts), and a separate model is trained for each possible combination of $n - 1$ folds. For each model, the held-out fold is used as the test set. At the end, combination of all test runs results in a fair evaluation of the entire training set. In the extreme case, n is set to the number of training instances, which means a separate classifier is built for every training instance. This is known as the **leave-one-out** scheme (cf., e.g., Weiss & Kulikowski 1991).

Besides these general considerations, which mainly affect how much data is chosen for what partition, an important question is on what basis one separates data for SAA evaluation. The vast majority of approaches use a partitioning based on individual **answers**. Of course, this means that the system has to classify answers it has not seen before, but it is possible that answers to the same questions are in fact part of the training set. Depending on how many similar answers have been seen during training, this setting somewhat simplifies the SAA task in the sense that a system does not need to evaluate an answer to a question, but can rather compare unseen answers to the ones already seen, and apply the corresponding judgment.

One step further is a data partitioning on the basis of **questions**. In this scenario, all answers to the test set questions are held out, requiring systems to be sufficiently general so that their approach generalizes to new questions. It is now no longer sufficient to compare answers to ones previously seen. However, questions and answers may still be on the same topics as the ones in the training set.

Finally, in the most extreme test case one partitions the data based on **domains**. This means that neither the answers, nor the questions, nor even the particular knowledge source (e.g. a reading text) that questions are based on have been seen before by the system. This represents by far the hardest evaluation scenario, because the approach must generalize to completely new lexical material.

The only context we know of where all of these settings have been explored is task 7 of SemEval 2013, described in Dzikovska et al. (2013). Following their terminology, we will call the three settings **unseen answers**, **unseen questions** and **unseen domains** in the remainder of this thesis.

2.3.2. Evaluation Metrics

Where evaluation metrics are concerned, SAA does not differ substantially from other CL tasks. In general, the evaluation measures used depend on the nature of the predicted outcome: if the outcome is on a continuous scale (e.g. 0–5) as grades often are, a **regression** evaluation metric is commonly used, such as a **correlation coefficient**. A correlation coefficient such as Pearson's r (Pearson, 1895) measures the degree to which two variables x and y are co-dependent: how systematically does y increase (positive correlation) or decrease (negative correlation) when x increases? In the SAA task, this question can be rephrased as "How systematically do the predicted scores increase when the gold scores increase?". The bounds for correlation values are 0 (no correlation) and 1 (perfect correlation), with .5 commonly seen as evidence for substantial correlation.

If the outcome is a nominal value, such as a class label *correct* or *incorrect* in the case of many SAA systems, calculating correlation is neither possible nor would it make sense. Instead, one uses measures for **classification** tasks, which are built on the observed system predictions compared to the ground truth (gold labels). The quantities observed with respect to some class c (e.g. *correct*) are called **true positives** (TP), **true negatives** (TN), **false positives** (FP) and **false negatives** (FN). While the former two represent true predictions c and $\neg c$ with respect to the gold standard, the latter two represent false ones (errors). Table 2.1 provides an overview of all four quantities in their relationships. This type of table is also called **confusion matrix**. Note that in binary classification tasks with classes c and d, $\neg c$ is equivalent to d, so all necessary information can be expressed in just one matrix.

On the basis of these quantities, several measures can be computed. Among

Actual class	Predicted class c	$\neg c$
c	TP	FN
$\neg c$	FP	TN

Table 2.1.: Confusion matrix

the most commonly used ones in CL are **precision** and **recall**. While precision (equation 2.1) can be paraphrased as "out of all the times we predicted *c*, how often were we correct?", recall (equation 2.2) expresses the question "how many of the actual instances of *c* did we get right?".

$$precision = \frac{TP}{TP + FP} \tag{2.1}$$

$$recall = \frac{TP}{TP + FN} \tag{2.2}$$

Both precision and recall are useful measures for evaluating classification approaches, but represent partial views of the whole performance. In order to get an intuitive picture of the overall performance of a classifier, we need to incorporate all observed quantities. The **accuracy** measure (equation 2.3) does just that, answering the question "how many instances did we classify correctly?".

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.3}$$

However, accuracy values are to be taken with a grain of salt. In data sets with an uneven class distribution, it is relatively easy to get high accuracy by simply always predicting the majority class; this is also called the **majority baseline**. Therefore, accuracy figures should always be interpreted in relation to this baseline, not in isolation.

2.4. Context Awareness

As mentioned in the introduction to this chapter, we want to distinguish systems based on context awareness, which can be defined as the degree to which the context of a short answer task is used in evaluating the answer. As we will see, this dimension can be seen as a spectrum, but we will eventually treat it as a binary distinction.

At one end, there are systems which essentially treat SAA as a paraphrase recognition task between student and target answer, without using any external information beyond what the Natural Language Processing (NLP) components employed for recognizing paraphrases need. The next step is to consult general-purpose external information sources, such as lexical semantic networks and large corpora, from which world knowledge can be extracted by various means and used to establish meaning equivalences that could otherwise not be recognized. However, this step still does not use the explicit task context, such as the knowledge source or the prompt, in any way. Because of this, we will subsume systems that fall into either of these two classes under **context-unaware**. These systems are described in section 2.5.1.

On the other end, there are some approaches which do use the task context. This may take the form of using the explicit knowledge source in a beneficial way. More important for the work in this thesis are approaches that integrate the question into SAA. Such integration can take the form of calculating similarity features between question and student or target answer, essentially treating the question no different from an answer. A slightly more involved use of the question is to give special treatment to previously mentioned words when evaluating the answer, as done by some approaches. We will call such systems **context-aware**, and describe them in section 2.5.2.

2.5. Overview of Systems

In this section, which is partly based on Ziai et al. (2012), we first give an overview of selected SAA systems in general, before discussing the context-aware systems in more detail. While the general part is not and not meant to be exhaustive, we do aim at describing all existing context-aware systems due to their relevance for our work. A much more exhaustive survey on SAA can be found in Burrows et al. (2015).

2.5.1. General Overview

This general overview is meant to showcase the breadth of the field in an exemplary manner. It also describes some of the pioneering approaches in the field, before bigger data sets became available and enabled the use of robust supervised machine learning approaches.

WebLAS

One of the earlier systems is WebLAS, presented by Bachman et al. (2002). A human task creator feeds the system with scores for model answers. Regular expressions are then created automatically from these model answers. Since each regular expression is associated with a score, matching the expression against a student answer yields a score for that answer. Bachman et al. (2002) do not provide an evaluation study based on data.

CarmelTC

Another earlier system is CarmelTC by Rosé et al. (2003). It has been designed as a component in the Why2 tutorial dialogue system (VanLehn et al., 2002). Even though Rosé et al. position CarmelTC in the context of essay grading, it may be considered an SAA system: in their data, the average length of a student response is approximately 48 words, and the responses are scored primarily for content. Their system is designed to perform text classification on single sentences in the student responses, where each class of text represents one possible model response, plus an additional class for 'no match'. They combine decision trees operating on an automatic syntactic analysis, a Naive Bayes text classifier, and a bag-of-words approach. In a 50-fold cross-validation experiment with one physics question, six classes and 126 student responses, hand-tagged by two annotators, CarmelTC reaches an F-measure value of 0.85. They do not report a baseline. Concerning the quality of the gold standard, they report that conflicts in the annotation have been resolved.

C-Rater

C-Rater (Leacock & Chodorow, 2003) is based on a paraphrase recognition approach. It employs hand-crafted correct answer models consisting of essential points formulated in natural language. Later work (Sukkarieh & Stoyanchev, 2009) also addresses semi-automatic model building, in which manual holistic scoring of answers is avoided. Leacock & Chodorow present two pilot studies, one of them dealing with reading comprehension. From 16,625 student answers with an average length of 43 words, they drew a random sample of 100 answers to each of the seven questions. This sample was scored by one human judge using a three-way scoring system (full credit, partial credit, no credit). Their system achieved 84% agreement with the gold standard. Information about the distribution of the scoring categories is given indirectly: A baseline system that always assigns the most frequent score would have achieved 47% accuracy.

IAT

Information extraction templates form the core of the Intelligent Assessment Technologies system (IAT, Mitchell et al. 2003. These templates are created manually in a special-purpose authoring tool by exploring sample responses. They allow for syntactic variation, e.g., filling the subject slot in a sentence with different equivalent concepts. The templates corresponding to a question are then matched against the student answer. Unlike other systems, IAT additionally features templates for explicitly invalid answers. They tested their approach with a progress test that has to be taken by medicine students. Approximately 800 students each worked on 270 test items. The automatically graded responses then were moderated: Human judges streamlined the answers to achieve a more consistent grading. This step already had been done before with tests graded by humans. Mitchell et al. state that their system reaches 99.4% accuracy on the full data set after the manual adjustment of the templates via the moderation process. Summarizing, they report an error of "between 5 and 5.5%" in inter-grader agreement and an error of 5.8% in automatic grading without the moderation step, though it is not entirely clear which data these statistics correspond to. No information on the distribution of grades or a baseline system is provided.

Oxford System

The Oxford system (Pulman & Sukkarieh, 2005) is another one to employ an information extraction approach. Again, templates are constructed manually. Motivated by the necessary robustness to process language with grammar mistakes and spelling errors, they use shallow analyses in their pre-processing. In order to overcome the problem of manually constructing templates, they also investigated machine learning techniques. To learn the templates, Pulman & Sukkarieh (2005) also annotated student answers with respect to "the part of the answer that deserves a mark", which bears an interesting resemblance to the focus annotation we describe in chapter 6. However, the automatically generated templates were outperformed by the manually created ones. Furthermore, they state that manually created templates can be equipped with messages provided to the student as feedback in a tutoring system. For evaluating their system, they used factual science questions and the corresponding student answers from GCSE tests. 200 graded answers for each of nine questions served as a training set, while another 60 answers served as a test set. They report that their system achieves an accuracy of 84%. With inconsistencies in the human grading removed, it achieves 93%. However, they do not report on the level of inter-grader agreement or on a baseline for the information extraction experiments.

Atenea

Pérez et al. (2005) present the Atenea system, a combined approach that makes use of Latent Semantic Analysis (LSA, Landauer et al. 1998 and n-gram overlap. While n-gram overlap supports comparing target responses and student responses with differing word order, it does not deal with synonyms and related terms. Hence, they use LSA to add a component that deals with semantic relatedness in the comparison step. As a test corpus, they collected nine different questions from computer science exams. A tenth question "[consists] of a set of definitions of 'Operating System' obtained from the Internet." Altogether, they gathered 924 student responses and 44 target responses written by teachers. Since their LSA module had been trained on English but their data were in Spanish, they chose to use Altavista Babelfish to translate the data into English. They do not provide information about the distribution of scores and about inter-grader agreement. Atenea achieves a Pearson's correlation of $r = 0.554$ with the scores in the gold standard. The system was later adapted more to a tutoring setting by modeling learner aspects and using them to suggest specific task progressions or topics, and also by including the possibility of self-assessment. This newer version is known as Willow (Perez-Marin & Pascual-Nieto, 2011).

Makatchev & VanLehn (2007)

The approach by Makatchev & VanLehn (2007) enters the landscape from the direction of Artificial Intelligence (AI). It is related to CarmelTC and its data set, but follows a different route: target responses are manually encoded in first-order predicate language. Similar logic representations are constructed automatically for student answers. They explore various strategies for matching these two logic representation on the basis of 16 semantic classes. In an evaluation experiment, they tested the system on 293 "natural language utterances" with 10-fold cross-validation. The test data are skewed towards the 'empty' label that indicates that none of the 16 semantic labels could be attached. They do not report on other properties of the data set such as number of annotators or number of questions to which the student answers were given. Their winning configuration yields an F-measure value of 0.4974.

Nielsen et al. (2009)

With their facets system, Nielsen et al. (2009) establish a connection to the field of Recognizing Textual Entailment (RTE, Dagan et al. 2009. In a number of friendly challenges, RTE research has spawned numerous systems that try to automatically answer the following question: Given a text and a hypothesis, is the hypothesis entailed by the text? Short answers assessment can be seen as a RTE task in which the target response corresponds to the text and the student response to the hypothesis. Nielsen et al. base their system on what they call facets. These facets are meaning representations of parts of sentences. They are constructed automatically from dependency and semantic parses of the target responses. Each facet in the target response is then looked up in the corresponding student response and equipped with one of five labels[2] ranging from unaddressed (the student did not mention the fact in this facet) to expressed (the student named the fact). This step is taken via machine learning. From a tutoring system in real-life operation, they gathered responses from third- to sixth-grade students answering questions for science classes. Two annotators worked on these data, producing 142,151 facets. Furthermore, all facets were looked up in the corresponding student responses and annotated accordingly, using the mentioned set of labels. The best result of the Facets System is 75.5% accuracy on one of the held-out test sets. With 10-fold cross-validation on the training set, it achieves 77.1% accuracy. The majority label baselines are 51.1% and 54.6% respectively. Providing this more fine-grained analysis of facets that are searched for in student responses, Nielsen et al. claim to "enable more intelligent dialogue control" in tutoring systems.

2.5.2. Context-aware Systems

As we have seen from the examples in the previous section, many systems do not model answer context in a general way, instead treating SAA much like a paraphrase modeling task. When we ask ourselves why this is the case, one possible reason in particular comes to mind.

From a practical point of view, it likely does not pay off to design a general way of e.g. modeling questions if the target application only needs to handle a relatively small number of different questions. Indeed, the overview in Burrows et al. (2015, Table 5) lists 37 approaches which on average (median) handle only nine questions. Only eight approaches were developed for a number of questions greater than 50, and four of these eight approaches use the corpus presented in this thesis. Conversely, the median number of answers tackled by systems is 1,029, showing that often many answers per question are available. With such answer/question ratios, it is likely that question-specific approaches perform much better than generalized, context-aware ones. This is further demonstrated by the fact that Tandalla (2012) won the recent ASAP-SAS challenge by developing specific regular expressions for each question. As we describe in section 2.6.2, there are only ten questions in the ASAP-SAS data set,

[2] In human annotation, they use eight labels, which are grouped into five broader categories as used by their system.

but more than a thousand training answers for each.

However, such approaches do not allow for any cross-task generalizations or insights, such as the role of information structure in SAA we are interested in, and they will also be less practical if one only has a few training answers per question, but a large number of questions in total. In the following, we will therefore review the few systems that we know of which do make general use of context in some way.

IndusMarker

In their IndusMarker system, Siddiqi et al. (2010) use question-specific pattern matching to grade short answers to various question types on the topic of Object Oriented Programming in computer science. Patterns are specified in the so-called Question Answer Markup Language (QAML), an XML-based format, and can be defined for the word or phrase level. For their evaluation, the authors collected answers from 225 students to six tests with a total of 87 questions. For each question, 25 answers were used to develop the patterns necessary for scoring, while the remaining 200 answers were used for testing.

While Siddiqi et al. (2010) do not use contextual features in their approach, they do provide a detailed question type taxonomy in the evaluation of their system. Breaking down system performance according to each of the 16 question types, they report the highest accuracy for 'true/false' questions (100%), where only a binary decision is required by the student, and the lowest for 'contrast' questions (84.1%), where the task is to name differences of particular concepts. However, besides the more complex task the latter question type also corresponds to a much higher average answer length (20.4 words) in comparison to the 'true/false' questions (1.2 words), which shows that variation in answers differs greatly with respect to question types.

Mohler et al. (2011)

One slightly context-aware approach is described by Mohler et al. (2011). Student responses and target responses are annotated using a dependency parser. Based on that, subgraphs of the dependency structures are constructed in order to map one response to the other. These alignments are computed using machine learning. Dealing with subgraphs allows for variation in word order between the two responses that are to be compared. In order to account for meaning, they combine lexical semantic similarity with the aforementioned alignment. They make use of several WordNet-based measures and two corpus-based measures, namely Latent Semantic Analysis and Explicit Semantic Analysis (ESA, Gabrilovich & Markovitch 2007). For evaluating their system, Mohler et al. collected the data set we describe in section 2.6.1 and made it publicly available. The system achieves $r = 0.518$ and a Root Mean Square Error of 0.978 as its best result.

The context-awareness here consists of a processing step the authors call "question demoting", which means that all words present in the question are removed from both the reference answer and the student answer, so that repeating words from the question is not rewarded. This is essentially the same technique applied by the CAM system (Bailey & Meurers 2008, see section 5.1) and in CoMiC, which we describe in chapter 5. However, the technique is not related to IS research in any way by Mohler et al. (2011), despite having clear connections to the notion of givenness, which we introduce in chapter 3.

CoSeC

Hahn & Meurers (2012, 2013) present the CoSeC-DE approach based on Lexical Resource Semantics (LRS, Richter & Sailer 2003. In a first step, they create LRS representations from POS-tagged and dependency-parsed data. These underspecified LRS representations of student responses and target responses are then aligned. Using A* as heuristic search algorithm, a best alignment is computed and equipped with a numeric score representing the quality of the alignment of the formulae. If this best alignment scores higher than a threshold, the system judges student response and target response to convey the same meaning. The alignment and comparison mechanism does not utilize any linguistic representations other than the LRS semantic formulae. These semantic representations abstract away from surface features, e.g., by treating active and passive voice equally. Hahn & Meurers claim that that "[semantic representations] more clearly expose those distinctions which do make a difference in meaning." They evaluate the approach on a subset of the Corpus of Reading Comprehension Exercises in German (CREG) (see chapter 4) containing 1,032 learner responses and report an accuracy of 86.3%.

In terms of context awareness, CoSeC uses the same mechanism employed for answer alignment to recognize meaning equivalence between parts of the question and parts of the student and target answer. Hahn & Meurers (2012) argue, in the same vein as we do in this thesis, that IS categories are relevant and helpful in classifying short answers. However, they have to approximate the relevant IS notion, focus (see chapter 3), by treating it as content not already expressed in the question, which is essentially the same idea pursued in the CoMiC system (see chapter 5), but transferred to the abstraction level of semantic instead of surface realizations.

Horbach et al. (2013)

Horbach et al. (2013) present an interesting approach that makes use of the reading text in classifying short answers, an idea that was also envisaged in the A4 project of SFB 833[3]. The idea is to locate the source of information in the reading text than an answer draws its content from, and then use that additional context for SAA in a beneficial way.

To achieve this goal, Horbach et al. (2013) first performed an annotation study of a part of the CREG-1032 corpus (see section 4.2.3), where the annotation task was to link each student and target answer to the sentence in the reading text that most directly represents its content. The percentage agreement obtained for this task was 74%.

The basis for the SAA experiments is an alignment-based approach modeled closely after the CoMiC system we describe in chapter 5. This approach is used both for answer-to-answer alignment in SAA and for automating the identification of the closest sentence in the reading text, for which the manual annotation described above serves as a gold standard.

Equipped with an SAA system and the aforementioned annotation results, Horbach et al. (2013) build and evaluate several models: i) the standard answer-based CoMiC baseline, ii) a simple text-based model which only compares whether the source sentences for both target and student answer are the same, iii) a model consisting of the baseline and four text-based features encoding the relationship of source sentences, and between source sentences and answers,

[3] `http://purl.org/icall/comic`

and iv) a combination of i) and iii). For all models with source sentence identification, both manual and automatic identification was tested.

Results put the answer-based model augmented with text-based features ahead (83.7%) of both the baseline (81.7%) and the combined (81.0%) models in leave-one-out testing using a k-nearest-neighbor algorithm with $k = 5$. Interestingly, the corresponding model based on manual annotation performs worse (82.7%). On the whole, the results show that using the reading text as further evidence is beneficial. This is further demonstrated by the fact the simple text-based model reaches 76.2% without any answer-based features.

Rudzewitz (2015)

Rudzewitz (2015) presents a recent approach on using the task context, specifically the question and the text, based on the CoMiC system (see chapter 5) and the CREG corpus (see chapter 4). Based on the word alignments between student and target answer in CoMiC, the central idea of the approach is to weight alignments based on both syntactic and contextual properties.

As far as the task-independent syntactic properties are concerned, Rudzewitz (2015) used part-of-speech classes, grouping them into nominal, verbal, adjective/adverb and others. For the contextual features, two weighting sources were explored: one based on binary indicator features of surface question forms such as *who* and *what*, and one based on TF-IDF scores (Salton & McGill, 1983) calculated for each word across reading texts. Rudzewitz (2015) also explored possible combinations of these weighting approaches.

The approach was tested on several subsets of the CREG corpus, since their characteristics differ (see section 4.2.3). Results show that each weighting method has the potential to improve over the baseline, with the combination of all three performing best across different data sets.

SemEval 2013 Task 7 Systems

Task 7 of SemEval 2013 (Dzikovska et al., 2013), entitled "The Joint Student Response Analysis And 8th Recognizing Textual Entailment Challenge", is the most recent shared task in SAA and the only one aimed at a tutoring setting, where students interact with a system and have to answer a wide variety of questions. To tackle this variety, several participating systems made use of the question context in building their models, which is why we describe them here. Also, these are some of the most recent approaches in SAA using a publicly available data set.

The main task was this: given a question and one or more reference answers, classify the student answer into one of 'correct', 'partially_correct_incomplete', 'contradictory', 'irrelevant' and 'non_domain'. There were also 3-way and 2-way versions of the task, where several of the non-correct labels were collapsed. Participants had to submit predictions for three testing scenarios, 'unseen answers', 'unseen questions' and 'unseen domains', which we already described in section 2.3. The data set is publicly available and we describe it further in section 2.6.3.

The complexity of the shared task's setup was quite high: 5-way, 3-way and 2-way subtasks on two sub-corpora with several evaluation measures in different testing scenarios. Overall, the approach by ETS (Heilman & Madnani, 2013) was ahead in most of the comparisons. They used a combination of word overlap features, word and character *n*-grams and text similarity features in a feature stacking approach (Wolpert, 1992). They also used the domain adaptation technique by Daume III (2007) where multiple copies of a feature are used to adapt the system to a different testing scenario. Although their approach worked very well overall, we do not describe it in more detail here because the use of context was not explicitly modeled, but rather left to the machine learning approach in the form of learning a different set of feature weights.

In the following discussion, we simplify the evaluation situation somewhat by limiting ourselves to the 'unseen questions' (uQ) and 'unseen domains' (uD) testing scenarios, since due to the lack of previously seen similar answers they represent the greatest need for generalizing the use of context. Moreover, we only report results on the 5-way task, because it is the only one that all relevant systems took part in, and we only report one evaluation measure, accuracy. To make the accuracy figures interpretable we give the majority baseline for the testing scenarios in Table 2.2.

Last but not least, we should mention that we also took part in the shared task (Ott, Ziai, Hahn & Meurers, 2013) and performed competively, particularly in

Data set	Maj. baseline ('correct')
Beetle uQ	42.0%
SciEntsBank uQ	41.1%
SciEntsBank uD	42.0%

Table 2.2.: Majority baseline for 5-way 'unseen questions' and 'unseen domains'

the 'unseen answers' scenario. However, since our contribution is a combination of the CoSeC system discussed above, the CoMiC system discussed in chapter 5 and a bag-of-words approach, we do not discuss it here.

Lexical Baseline The SemEval 2013 Task 7 lexical baseline, described by Dzikovska et al. (2012), pursues a word overlap strategy between student answers and reference answers, and between student answers and questions, in order to provide a stronger baseline system than the majority baseline. Four similarity metrics are computed, all based on word overlap using the `Text::Similarity` Perl package[4]: i) the raw number of overlapping words, ii) the F1 score (average of precision and recall), iii) the Lesk score (Lesk 1986, used originally to compute semantic similarity of two words via their definitions) and iv) the cosine score.

The resulting eight scores were combined in a classifier that can then be trained on the given training set. It turned out that the lexical baseline was rather strong, and not always outperformed by all participants (Dzikovska et al., 2013). In fact, the overall best performing system by Heilman & Madnani (2013) explicitly incorporated the lexical baseline features as part of their own model. The accuracy scores for this baseline are given in Table 2.3.

Data set	Accuracy
Beetle uQ	48.0%
SciEntsBank uQ	41.3%
SciEntsBank uD	41.5%

Table 2.3.: Lexical baseline for 5-way 'unseen questions' and 'unseen domains'

[4] `http://search.cpan.org/dist/Text-Similarity/`

SoftCardinality Jimenez et al. (2013) present an approach called SoftCardinality, which is also based on text overlap, but takes a novel direction in computing a variant of classical set cardinality for complex structures. The basic elements to be compared are words, for which the authors define a character-based similarity function based on standard set cardinality and the Dice coefficient. Given this word similarity, the 'soft' cardinality of a sentence can then be computed, which in turn enables the authors to calculate sentence similarity and paragraph similarity.

The 42 features computed represent soft cardinalities calculated on different set-theoretic combinations and normalizations of student answer, target answer and question. Only basic pre-processing was done to the original text, up to the point where each stemmed word can be represented as character *n*-grams. Given the features and this representation, a decision tree model was trained for each subtask. The results are shown in Table 2.4. The system performed especially competitive on the SciEntsBank corpus, and obtained the best results in the 'unseen domain' scenario across all subtasks and systems.

Data set	Accuracy
Beetle uQ	45.1%
SciEntsBank uQ	52.5%
SciEntsBank uD	51.2%

Table 2.4.: SoftCardinality results for 5-way 'unseen questions' and 'unseen domains'

CNGL Bicici & van Genabith (2013) build their SAA system on top of a machine translation approach. The idea is to regard a combination of question, target answer and student answer as a translation problem with a source (e.g. the question), a target translation (e.g. the target answer) and a reference translation (e.g. the student answer). They use four different combinations, each representing different possible perspectives of how the task can be modeled within the translation framework.

Bicici & van Genabith (2013) then use a total of 283 features based on word *n*-grams and head-modifier dependencies to train translation models that identify translation acts between source and target and classify the student answer according to how well the source translates to the target, given the reference translation. The results are shown in Table 2.5. Despite the rather high complexity of the approach, the accuracies obtained are moderate, only outperforming the majority baseline for the Beetle corpus.

Data set	Accuracy
Beetle uQ	44.8%
SciEntsBank uQ	29.9%
SciEntsBank uD	27.4%

Table 2.5.: CNGL results for 5-way 'unseen questions' and 'unseen domains'

EHU-ALM Aldabe et al. (2013) also include the question in their calculation of various shallow and deep overlap measures. The features they use are based on i) text overlap following the lexical baseline, ii) lexical and graph similarity using WordNet (Miller, 1995) following the knowledge-based measures described by Mihalcea et al. (2006), corpus-based similarity measures (LSA, Landauer et al. 1998, and LDA, Blei et al. 2003), syntactic structure overlap and predicate-argument overlap (McCarthy et al., 2008). Except for the syntactic structure and predicate-argument overlap features, all of these are calculated between all possible combinations of question, target answer and student answer.

The features were combined using a Support Vector Machine (cf., e.g., Hearst et al. 1998) as the classifier. Results are shown in Table 2.6, placing the approach in the middle of the participant field. Besides standard training approaches, one interesting direction taken by Aldabe et al. (2013) is the partitioning of the training set into three different question types, namely *what*, *how* and *why* questions. Separate classifiers were then trained for each question type. However, the effect of this strategy was modest.

2.6. Overview of Available Data Sets

In this section, we give an overview of the publicly available data for SAA to date. This overview does not include our own empirical basis, the Corpus of Reading Comprehension Exercises in German (CREG), since it is part of our

Data set	Accuracy
Beetle uQ	42.2%
SciEntsBank uQ	46.8%
SciEntsBank uD	45.7%

Table 2.6.: EHU-ALM results for 5-way 'unseen questions' and 'unseen domains'

thesis research and hence presented in greater detail in chapter 4. Also, since our own SAA approach described in chapter 5 builds on the CAM approach by Bailey & Meurers (2008), we discuss the data CAM uses in section 5.1.

2.6.1. The Data Set by Mohler et al. (2011)

The data set used by Mohler et al. (2011) in the development and evaluation of their system consists of twelve introductory-level computer science assignments, two of which were examinations. 31 students wrote a total of 2,273 answers to the 80 questions in the assignments. Grading was done by two annotators (one TA and one of the authors of the paper) on a 0–5 integer scale. No explicit annotation guidelines beyond the grading scale were given. The gold standard to train and test the system on was then formed by taking the arithmetic mean of both annotator grades, instead of an adjudication process where differences are resolved in a principled manner. Given that the annotators only agreed in 57.7% of the cases (with a Pearson correlation of $r = 0.586$), this method affects more than 40% of the answers. Mohler et al. (2011) mention that "[t]he dataset is biased towards correct answers".

2.6.2. The ASAP-SAS data

The Automated Student Assessment Prize in Short Answer Scoring (ASAP-SAS)[5] was organized in 2012 by the Hewlett Foundation[6] as part of a larger effort to advance educational testing in the United States. The Kaggle platform was used to carry out the shared task, for which the best scoring systems were awarded prize money. The data comes from high school tests (mostly from

[5] https://www.kaggle.com/c/asap-sas

[6] http://www.hewlett.org/

10th grade) and was provided by several US states. It comprises ten questions, of which five come from biology or science classes, and five are from English language arts classes. For each question, about 2,100 to 3,000 responses are available, resulting in a total data set size of 27,367 responses, making this the largest English data set that is available for SAA. Each response was rated by two annotators on a 0–2 or 0–3 integer scale, depending on the question. Agreement was calculated using a weighted version of the Kappa statistic, ranging from .738 to .970, again depending on the particular question.

2.6.3. The SemEval 2013 Task 7 Data

Dzikovska et al. (2012) present the data set that was used for the Joint Student Response Analysis and 8th Recognizing Textual Entailment challenge at SemEval 2013. The data set draws on two sources, both coming from the science domain: the Beetle corpus collected and annotated as part of an evaluation of the Beetle II tutorial dialogue system (Dzikovska et al., 2011), and the SciEntsBank corpus consisting of student answers to questions in the Assessing Science Knowledge (ASK) project[7], whose annotation is described by Nielsen et al. (2008). The annotation scheme employed by Dzikovska et al. (2012) uses five categories (some which which were collapsed at sub-tasks of the challenge): 'correct', 'partially_correct_incomplete', 'contradictory', 'irrelevant' and 'non_domain'. These labels are designed so they fit both the SAA and the RTE setting, reflecting the joint nature of the challenge.

Since both sub-corpora had originally been annotated using other annotation schemes, this annotation had to be mapped to fit the labels listed above. In the case of the Beetle corpus, the original annotation (carried out with an agreement of $\kappa = .69$) was quite close and could be mapped straightforwardly. For the SciEntsBank corpus, the original annotation had been carried out on a much more fine-grained level, a dependency-based representation of sentences where individual "facets" (Nielsen et al., 2008) represent facts that students need to demonstrate their knowledge of. Annotator agreement on these facets was $\kappa = .73$. The facet-level annotation was converted to the response-level annotation required for the challenge using a set of rules, essentially projecting the facet-level annotations to the higher response level.

[7] http://bearcenter.berkeley.edu/project/assessing-science-knowledge-ask

As far as resulting corpus size is concerned, the Beetle corpus consists of 2,729 answers while the SciEntsBank corpus has 5,251 answers. Both data sets are skewed towards correct answers, with 42% such answers in the Beetle corpus and 40% in the SciEntsBank corpus.

Summary

In this chapter, we gave an introduction and an overview of the field of Short Answer Assessment (SAA). We first defined the task, which is to classify an answer to a question with respect to a knowledge source, and usually a reference/target answer. After that, we outlined the challenges of SAA, such as form variation, and mentioned the fields which share some of these challenges with SAA. We then briefly discussed evaluating SAA systems, focusing on evaluation settings and evaluation metrics. Since in this thesis we are most interested in whether systems make use of the context in evaluating answers, we briefly discussed how such **context-aware** systems differ from others.

Having equipped the reader with the preliminaries, we launched into an overview of SAA systems. We first discussed some earlier and more general approaches before giving a more comprehensive overview of the context-aware systems that we know of. From this survey, we can conclude that there is very little specific treatment of questions in SAA, with most systems incorporating it similarly to how they incorporate the reference answer. Most importantly, besides our own system and related ones (Bailey & Meurers, 2008; Hahn & Meurers, 2012), no approach has realized the connection between Information Structure and SAA.

Finally, we presented and characterized the publicly available data sets to date.

3. Information Structure: Focus as a Notion Perspectivizing Information in Answers

In this chapter, we give an overview of IS as the necessary basis for understanding the research we describe in later chapters. We will start by defining the research subject matter of IS in general terms (section 3.1) before presenting the main notions and distinctions that have been discussed in the theoretical literature (section 3.2)[1]. We identify the IS dimensions most relevant for our work before giving an overview of how these dimensions have been annotated in the literature in section 3.3. We then compare two dimensions, focus/background and given/new for the purpose of SAA in section 3.4) and finally discuss why the focus/background dimension is the most promising for our research goals.

3.1. What Is Information Structure?

While the concept of IS is much older, the term *information structure* was coined by Halliday (1967), who envisaged IS to be a separate layer of communication where so-called 'information units' are organized by the speaker:

> "Any text in spoken English is organized into what may be called 'information units'. The distribution of the discourse into information units is obligatory in the sense that the text must consist of a sequence of such units. But it is optional in the sense that the speaker is free to decide where each information unit begins and ends, and how it is organized internally; this is not determined for him by the constituent structure. Rather could it be said that the

[1] The sub-section on Givenness is partly based on Ziai et al. (2016)

> distribution of information specifies a distinct constituent structure on a different plane; this 'information structure' is then mapped on to the constituent structure as specified in terms of sentences, clauses and so forth, neither determining the other." (Halliday, 1967, p. 200)

In other words, IS relates and maps to other linguistic layers, such as syntax, but it is primarily a functional layer which allows the speaker the freedom to present information in the way she sees fit, while conforming to the linguistic system of the respective language. Note that Halliday explicitly mentions English here, but at an abstract level, IS is not language-specific. Chafe (1976) took the idea of how information is presented one step further and introduced the term 'information packaging' to clearer distinguish the information itself and its organization in the discourse.

To make things more concrete, let us consider (8), which is an adapted example from the corpus we present in chapter 4.

(8) Isabel geht joggen, das macht ihr Spaß.

'Isabel goes jogging, that's fun for her.'

(8) illustrates two important aspects: first, the speaker needs to introduce *Isabel* and *joggen* ('jogging') before referring to them as *ihr* ('her') and *das* ('this'), while the reverse order would be unnatural. Second, *Isabel* is established as the main topic of the utterance, and the clause *das macht ihr Spaß* ('this is fun for her') refers to that topic and adds new information about it.

Moreover, the utterance in (8) would not be uttered in isolation, but would likely attempt to satisfy some information requirement. Consider the modified version in (9).

(9) Welchen Sport macht Isabel?
which sport does Isabel

A: Isabel geht joggen, [das macht ihr Spaß].

'Isabel goes jogging, that's fun for her.'

Here, the information requirement is formulated through the question *Welchen Sport macht Isabel?* ('Which Sport does Isabel do?') and the relevant information in the answer is thus found in the first part, namely *joggen* ('jogging'), whereas the second clause can be considered to satisfy a different information requirement, such as *Warum geht Isabel joggen?* ('Why does Isabel go jogging?').

Examples (8) and (9) showcase three main independent distinctions that are commonly made in IS research. In (8), the expressions *das* ('this') and *ihr* ('her') in the second clause refers to a **given** concept since their referents have already been introduced, whereas the rest of the clause is **new**. *Isabel* is also marked as the **topic** of the sentence, whereas the rest of the sentence can be seen as the **comment** on this topic. Finally, the answer in example (9) provides the word *joggen* as the **focus** with regard to the question *Welchen Sport macht Isabel?*, whereas the rest of the utterance is referred to as the **background**.

In the next section, we will explain each of these distinctions in more detail.

3.2. Overview of IS Notions

3.2.1. Topic/Comment

Let us begin by discussing the notion of **topic vs. comment**. As the name suggests, the distinction separates an entity (the topic) from something that is said about it (the comment). The terms 'topic' and 'comment' were coined by Hockett (1958, p. 201), who built on a number of other works going back all the way to Aristotle. Reinhart (1981) subsequently developed a theory of communication that includes topic based on the notion of 'common ground', i.e., the information that can be assumed to be known by participants in a communicative scenario. In this theory, Reinhart assumes an entry-based storage of information, where the topic is the entry under which an assertion in the comment is stored. Based on this idea, Krifka & Musan (2012) define topic as follows:

Definition 1. *The topic constituent identifies the entity or set of entities under which the information expressed in the comment constituent should be stored in the common ground content.* (Krifka & Musan, 2012, p. 28)

To make this view of topic more concrete, we will consider the examples in (10). In (10a), assuming a context such as *Who does John live with?*, the noun phrase *John* is the topic under which the fact that he lives together with Mary will be stored. In (10b), given the respective context *Who does Mary live with?* the situation is reversed, with Mary being the topic and the fact that John lives together with her being stored under her entry. Of course, the end result of both statements in the world is the same, but the way in which information is integrated in the common ground differs.

(10) a. $[\![\text{John}]\!]_T$ lives together with Mary.
b. $[\![\text{Mary}]\!]_T$ lives together with John.

As one can see, the topic/comment distinction in these cases captures what a sentence is about, which has also led to the term 'aboutness topic'[2]. It is a useful distinction if one is interested in the entities that a discourse discusses, and what is said about them. It does, however, not capture whether the information asserted about a topic is new or whether it provides requested information.

3.2.2. Givenness

The **given vs. new** distinction separates information that is available in the common ground (given) from information that is not available (new) and hence must be integrated first. Givenness was discussed both by Halliday (1967) and Chafe (1976), but the most well-known and influential definition comes from Schwarzschild (1999), which we give below:

Definition 2. *An utterance U counts as* GIVEN *iff it has a salient antecedent A and either i) A and U co-refer or ii) A entails the Existential F-Closure of U* (Schwarzschild, 1999, p. 151)

This definition builds on the formal semantic concept of existential f-closure of *U*, which Schwarzschild defines as "the result of replacing F-marked phrases in *U* with variables and existentially closing the result, modulo existential type shifting" (Schwarzschild, 1999, p. 150).

Intuitively, Definition 2 counts an utterance as Given if it can be recovered from the common ground. This can mean that either the utterance has been explicitly mentioned before, or its meaning is entailed by some contextually available knowledge. Schwarzschild (1999) uses Givenness to predict which

[2] For other topic notions, see Krifka & Musan (2012, sec. 4 and 5).

parts of an utterance will be prosodically prominent, i.e., accented. The rationale is that Given expressions do not bear accents. To make this more concrete, consider example (11), which is example (12) from Schwarzschild (1999), where the relevant expression *convertible* has been mentioned literally before, and is thus given.

(11) John drove Mary's red convertible. What did he drive before that?
A: He drove her BLUE convertible.

The default stress assignment of English (cf., e.g., Culicover & Rochemont 1983) would put the main accent of the answer on *convertible*, but since it is Given, it is deaccented and the accent is placed on *blue* instead. Note also that the pronoun *her* is deaccented because it co-refers with *Mary*, who has been mentioned before. A semantically more interesting case of Givenness involves semantically similar words such as synonyms and hypernyms, as exemplified by *violin* and *string instrument* in (12), mentioned as example (7) by Büring (2007).

(12) (I'd like to learn the violin,) because I LIKE string instruments.

The existence of a violin entails the existence of a string instrument, so *string instrument* is *given* and deaccented under Schwarzschild's approach. To complete the empirical overview of the landscape of cases that the Givenness notion is expected to handle, let us briefly discuss the phenomenon known as 'bridging'. It can be exemplified using (13), which is example (29) of Schwarzschild (1999).

(13) a. John got the job.
b. I KNOW. They WANTed a New Yorker.

Here, the phrase *New Yorker* is Given and deaccented on account of available background knowledge establishing that the individual *John* is from New York. This is only captured vaguely by Definition 2 through the notion of salience, and it remains to be worked out exactly how the salience is established (Schwarzschild, 1999, p. 153–154).

Since these types of Givenness are quite different, some authors have characterized these differences in taxonomies. For example, Prince (1981) distinguishes the types 'New', 'Inferrable' and 'Evoked', each with various subtypes, on the basis of how exactly the information is accessible from the previous discourse or common ground. We discuss several annotation approaches building on this type of taxonomy in section 3.3.2.

Summing up, Givenness subsumes several ways in which something can be previously mentioned or accessible, and separates this content from new content. It is therefore a useful notion if one wants to determine whether an utterance contains information not already present in the common ground.

3.2.3. Focus

The last distinction we want to discuss is the one most central for this thesis, namely **focus vs. background**. Intuitively, focus marks the relevant part of an utterance, with regard to what is currently being discussed. This is however a rather vague definition, as it does not make explicit exactly what relevance means. A more useful basis for defining focus was put forward by Rooth (1985, 1992), who established the notion of **alternatives** in natural language semantics: roughly speaking, an alternative is one of a set of semantically compatible expressions that a speaker can use in a given scenario. Based on the idea of alternatives, let us define focus according to Krifka & Musan (2012):

Definition 3. *Focus indicates the presence of alternatives that are relevant for the interpretation of linguistic expressions.* (Krifka & Musan, 2012, p. 7)

This is a rather general definition which simply states that wherever focus occurs, it signals that there are semantic alternatives to the focused expression available in the context. Let us illustrate this with example (14), which is a simplified version of (8). Here and throughout this thesis, focus is indicated by double braces.

(14) Isabel geht ⟦joggen⟧$_F$.
Isabel goes jogging

Here, *joggen* ('jogging') is focused which indicates that there are other members from the set of sports potentially relevant here, and *joggen* is explicitly chosen from that set. To make the notion of alternative sets more concrete, consider the modified version in (15), where an explicit question has been added, to which the aforementioned sentence is a possible answer.

(15) Welchen Sport macht Isabel?
which sport does Isabel

A: Isabel geht ⟦joggen⟧$_F$.
Isabel goes jogging

The question *Welchen Sport macht Isabel?* makes explicit what is being asked for here, namely a sport. Focus here is the part of the answer that selects from the set of alternatives (possibly including swimming, running, etc.).

In order to incorporate questions into our definition of focus, we must take a step back. So far, we have seen focus as indicating the presence of alternatives, as per Definition 3. This definition however (deliberately) makes no statements about the nature and location of said alternatives. As example (15) shows, questions appear to be one way of making alternatives more concrete. In fact, formal semantics has tied the notion of alternatives to an explicit relationship between questions and answers called Question-Answer Congruence (QAC, see Stechow 1991). The central idea of QAC is that an answer is congruent to a question if both evoke the same set of alternatives. This is evidently the case in (15), where both the question and the answer evoke the set of sports.

In a slightly different strand of research, formal pragmatics has established the notion of Question Under Discussion (QUD) as a way of modeling discourse (Roberts, 1996, 2012). The general idea here is that discourse is structured through implicit or explicit questions representing the current topic under discussion. Discourse participants negotiate these questions and move between them, e.g. becoming more specific or general, or abandoning a question altogether. For example, a more specific QUD to the question in (15) would be *Geht Isabel joggen oder schwimmen?* ('Does Isabel go jogging or swimming?').

To tie these notions together, a definition of focus in terms of QUDs is needed. While such a view is assumed by several researchers, only few follow the example by Roberts (2012) in stating it explicitly. One such case is the definition by Riester & Baumann (2013), which we will assume for the remainder of our discussion and this thesis in general:

Definition 4. *"A focus is either an answer to the immediate (explicit or implicit) Question under Discussion (at-issue focus) or to a supplemental question (not-at-issue focus)."* (Riester & Baumann, 2013, p. 221)

This definition is not contrary to Definition 3, but assumes a **top-down** approach to focus, from context (here: the question) to utterance (here: the answer). In contrast, Definition 3 is **bottom-up** in nature, starting out from the focus marking and moving to the alternatives indicated by it. This distinction is also made by Riester & Baumann (2013).

The connection between questions, alternative sets and focus is crucial for the notion of focus we pursue in this thesis. In narrative discourse, the QUD is typically implicit, but in other types of data, such as interviews or dialogues, one finds more overt types of information requests. We return to this issue at the end of section 3.3.2 when we discuss focus annotation approaches.

The semantic type of alternatives, and hence of focus, can take different values. (16) illustrates this by presenting a different question to the answer we saw in (15).

(16) Was macht Isabel?
what does Isabel

A: Isabel ⟦geht joggen⟧$_F$.
Isabel goes jogging

In (16), the set of alternatives does not contain sports, but all things Isabel could plausibly do, such as eating, watching a movie, or going to bed. Consequently, the focus in the answer is not just *joggen*, but the whole verb phrase *geht joggen*. In the literature, this variation in the scope of focus is known as 'narrow' vs. 'wide' focus (cf., e.g., Selkirk 1984).

In an even more complex case, the alternative set can contain propositions, which are typically expressed as whole clauses. Consider example (17), where the question asks for a reason:

(17) Warum braucht Isabel neue Sportschuhe?
why needs Isabel new sports shoes

A: ⟦Isabel geht joggen⟧$_F$.
Isabel goes jogging

As we can see in (17), the whole answer is in focus because focus here selects from an alternative set of reasons, not of actions as in (16) or of individuals as in (15). The notion of alternatives thus flexibly captures different kinds of questions and corresponding foci.

In conclusion, the notion of **topic** deals with common ground management and how information is organized there, whereas **givenness** and **focus** deal with how languages mark new and relevant information with respect to the previous discourse. In the next section, we will therefore take a close look at givenness and focus from the perspective of evaluating short answers.

3.3. Information Structure Annotation

Having given an overview of the main IS notions, we will now review previous research in annotating these notions in corpus data. Before launching into this review proper, let us first define what annotation is and what purposes it serves, both generally and in the context of this thesis.

3.3.1. What is Annotation?

According to Bird & Liberman (2000), the term **linguistic annotation** "covers any descriptive or analytic notations applied to raw language data" (Bird & Liberman, 2000, p. 23). While the simplest and most low-tech form of linguistic annotation would be to take a text on paper and mark it up with the relevant notations using a pen, application to language data here typically means that the notations are coded electronically, as is the source text.

A useful view of annotation is that of enriching language data by making properties of it explicit. In linguistic annotation, this usually means marking instances of a language phenomenon in corpora. To give a popular example, consider the example of part-of-speech annotation: a sentence such as *The man snores* can be annotated to *The/DT man/NN snores/VVFIN*, where the uppercase codes are part-of-speech tags in the Penn Treebank tagset (Marcus et al., 1993a) meaning 'determiner', 'noun' and 'finite verb', respectively. On the syntactic layer, the same example can then be further enriched to the following Penn Treebank notation:

(18) `(S (NP (The DT) (man NN)) (VP (VVFIN snores))).`

The basic ingredients for an annotation approach are the following: first, one needs the actual language data, preferably of a genre which contains enough instances of the property one wants to annotate. Second, one needs to create systematic guidelines for the annotation process, in order to make it as deterministic as possible. Finally, annotations should not be produced by the researcher who devised the guidelines, but by at least two independent annotators, in order to a) prevent idiosyncratic solutions and b) to be able to assess the success of the approach by measuring annotation consistency between raters. Artstein & Poesio (2008) discuss various measures of inter-annotator agreement for this purpose, but the most widespread ones are the various versions of κ (Cohen, 1960; Fleiss, 1971) and Krippendorff's α (Krippendorff, 1980). Since most of the annotation approaches we discuss use Cohen's κ, its definition is given in equation 3.1, where p_o is the observed percentage agreement and p_e is the expected percentage agreement given the label distribution for each annotator.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3.1}$$

Interpreting κ-values has been the subject of some debate (cf.,e.g., Artstein & Poesio 2008, sec. 4.1.3), but values above .6 are generally interpreted to indicate at least substantial agreement. After agreement evaluation, the different annotation versions are merged into a single version by yet another annotator acting as judge in cases of conflict. The result is a definite version of the annotated resource, known as the **gold standard**.

Once performed, an annotation functions like an index into language data: annotated properties can be searched for and used as filtering criteria for linguistic analysis. Given a corpus of part-of-speech annotated texts analogous to the example above, it is straightforward to formulate queries such as "give me all nouns" or "give me all nouns preceded by determiners". Meurers (2005), for instance, discusses the possible benefits of annotated corpora for theoretical linguistic analysis, showing that what may look like a sound theory from an introspective point of view can often be falsified by searching authentic annotated data.

For computational linguistics, annotation has been an invaluable method of providing training data for supervised machine learning approaches. During the 1990s, this led to a breakthrough in parsing (cf., e.g. Collins 1996; Charniak 1996) and continued on to other more semantic NLP tasks such as semantic role labeling (cf., e.g., Carreras & Marquez 2005) and machine translation (cf., e.g., Och 2003), and even reaching discourse-level tasks such as co-reference resolution (cf., e.g., Soon et al. 2001).

The discourse level is where annotation work becomes relevant for the specific goals of this thesis. In the next section, we will therefore review some important IS annotation work that has been done on topic, givenness and focus, with an emphasis on the latter.

3.3.2. Overview of IS Annotation Approaches

Givenness, or information status, as it is often called, has traditionally been the more researched area in terms of annotation approaches, while topic and focus have not been investigated as much in corpus data. As we will see, this is partly due to the abstractness of topic and focus, making them difficult to operationalize in an actual annotation approach. We will briefly review information status approaches first, before turning to the more challenging other two notions.

This overview is not meant to be exhaustive, but to provide a view of the landscape in IS annotation. Others have provided a broader perspective of corpus annotation in connection with IS (cf., e.g., Lüdeling et al. 2016).

Information Status

MULI The MULI project (Baumann et al., 2004a,b; Kruijff-Korbayová & Kruijff, 2004) aimed at annotating the relevant factors of information structure in German and English. The rationale for this is that a theory introduces bias in the annotated resource, which in turn lowers its general value. The annotation was done in three layers, syntax, prosody and discourse, representing the linguistic dimensions which the authors deemed relevant for IS. On the syntax layer, the authors build on the annotation already present in the Tiger Treebank (Brants et al., 2002) and the Penn Treebank (Marcus et al., 1994), adding topological field information following (Becker & Frank, 2002) and an encoding of non-canonical word order structures. On the prosody layer, the authors created recordings of the treebank texts to support the annotation of pitch accents, boundary tones and the position and strength of phrase breask following the ToBI (Tones and Break Indices, Silverman et al. 1992) and GToBI (Grice & Baumann, 2002) schemes. Finally, the discourse layer includes a variety of semantic information about discourse entities, including their semantic type, denotation characteristics, specificity and quantification. The most relevant of these for our discussion here is the annotation of familiarity status following Prince (1981), distinguishing the categories 'new', 'unused', 'inferrable', 'textually evoked' and 'situationally evoked'.

While the objectives and the sheer scope of the project is laudable, no evaluation has ever been reported, so the success of the effort is unclear.

Hempelmann et al. (2005) Coming from a more cognitive perspective, Hempelmann et al. (2005) annotated information status in written language, with the objective of predicting it in a computational approach using LSA (Landauer et al., 1998). They again based their effort on the taxonomy by Prince (1981), but collapsed it into the three categories 'given', 'new', and 'inferrable' due to data sparseness. Subsequently, the scheme was applied to all NPs in four texts from 4th grade textbooks, using two annotators. Across the 195 sentences containing 478 NPs, the authors achieved an inter-annotator agreement of $\kappa = .74$ (88% percentage agreement), which is substantial. The class distribution obtained was 317 'given' NPs, 116 'inferrable', and 45 'new'. All NPs were then hand-annotated with respect to whether they are pronominal, definite and whether they overlap with previous NPs in terms of content words. Together with two LSA-based measures, these variables were used to train two logistic regression models to predict information status. The best model achieved an accuracy of 80% (with a majority baseline of 66%).

Switchboard Nissim et al. (2004) describe annotation work on information status in the Switchboard corpus, which consists of telephone conversations on pre-defined topics. The portion of Switchboard used here is part of the Penn Treebank and thus is syntactically annotated. Nissim et al. (2004) build their annotation scheme on the taxonomy by Prince (1992) and also draw on Eckert & Strube (2000), distinguishing the high-level categories 'old', 'mediated' and 'new'. At the same time, the scheme includes finer-grained distinctions for 'old' and mediated', resulting in a total of 16 categories. In an annotation study using two annotators on three Switchboard dialogues, 1,738 NPs were annotated, excluding locative, directional and adverbial NPs. One dialogue was used to train the second annotator, while the first annotator was in fact the first author. The agreement obtained for the high-level categories was $\kappa = .845$, and $\kappa = .788$ for the finer-grained distinctions. The authors observed that is was easier for annotators to assign subtypes of the 'old' class than those of the 'mediated' class, and that the subtypes for which syntactic clues are relevant were annotated more reliably.

LISA Concerning information status annotation in other languages, Ritz et al. (2008) present an annotation study across three different text types: question/answer pairs elicited through a questionnaire (Skopeteas et al., 2006), map task dialogues where interlocutors collaborate on reaching a destination, and newspaper commentaries from the Potsdam Commentary corpus (Stede, 2004). Two undergraduate students of linguistics performed the annotation, after a three day testing phase. Besides information status, the study also encompassed topic and focus[3] annotation, all based on the LISA guidelines (Götze et al., 2007) developed at the SFB 632 in Potsdam for cross-language annotation of information structure (cf., e.g., Chiarcos et al. 2009). Before annotating information structure categories, the annotators performed syntactic annotation, which was subsequently corrected and merged into a gold standard. The category system used for information status builds on the one by Nissim et al. (2004) mentioned above, also featuring a coarse-grained level with the classes 'given', 'accessible' and 'new', and a fine-grained level with more subclasses for 'given' and 'accessible'. In total, 7 classes are distinguished on the fine-grained level. In the evaluation, κ-values for the coarse-grained tagset were highest for the NPs in the 42 question/answer pairs ($\kappa = .80$), followed by the 2 map task dialogues ($\kappa = .66$) and finally the newspaper commentaries ($\kappa = .60$). For the fine-grained tagset, the results are generally lower, but follow the same pattern, with $\kappa = .73$ for the question/answer pairs, $\kappa = .61$ for the dialogues, and $\kappa = .55$ for the commentaries. The authors do not report how many instances were annotated in each case. Concerning the sources of disagreement, the authors identify the referentiality status of elements as a source of confusion among the annotators, this was especially the case with relative and reflexive

[3] We describe this effort in more detail below.

pronouns, and with expressions in metaphors and collocations.

RefLex The RefLex scheme (Riester et al., 2010; Riester & Baumann, 2011; Baumann & Riester, 2010, 2012; Riester & Baumann, 2013) was developed as a finer-grained way of characterizing information status than what had previously been done. The scheme was created in a German context. However, unlike all previously discussed schemes, it distinguishes two dimensions of givenness, a lexical and a referential dimension. The lexical dimension is used for lexically mentioned or entailed concepts, such as lemmas, synonyms and hyperonyms of the expression in question. The referential dimension is used for all referring expressions, such as pronominals and most definite expressions. In a way, the referential dimension thus corresponds to clause *i)* of Schwarzschild's definition of Givenness (see Definition 2), while the lexical dimension corresponds to clause *ii)*. The separation also translates to the syntactic layer upon which RefLex is applied: where lexical givenness typically applies to individual content words and non-referential phrases, the referential level applies at the level of referring determiner phrases (DPs) or prepositional phrases (PPs). This makes it possible to label the noun *man* in the DP *The man* as lexically new, while at the same time labelling the whole DP as given, since it refers to a previously introduced individual. Riester & Baumann (2013) list the following categories as the most important ones for the R(eferential) level: R-GIVEN, R-BRIDGING, R-UNUSED, R-NEW, R-GENERIC and OTHER. For the L(exical) level, they mention L-GIVEN, L-ACCESSIBLE and L-NEW as the main labels. In a validation effort of RefLex on written news text from the DIRNDL corpus (Eckart et al., 2012), two trained student annotators applied the R-level to 3,445 referring DPs and PPs, and the L-level to 5,045 content words. The agreement obtained was $\kappa = .75$ on the R-level and $\kappa = .64$ on the L-level. Riester & Baumann (2013) suggest that the lower score for the L-level is due to the confusion among annotators on when to assign the label L-ACCESSIBLE.

Topic

Prague Topic-Focus Articulation The Czech Prague Dependency Treebank (Hajič et al., 2006) builds on the Prague School of functional and structural linguistics (Sgall et al., 1986), where IS annotation plays an important role due to the interaction with Czech's relatively free word order. The notion annotated in the treebank which is most relevant for our discussion here is called Topic-Focus Articulation (TFA). It is annotated on top of a dependency-based tree representation for sentences called 'tecto-grammatical structure' where each node receives a TFA value (one of 't'(opic), 'f'(focus) and 'c'(contrastive)). TFA in turn builds on another distinction, 'contextual boundedness', which essentially captures the aboutness relation: a declarative sentence asserts contextually unbound information about its topic, which is contextually bound. Contextually bound nodes can be contrastive, resulting in a 'c' value for the TFA attribute. It is important to note that what is called "Focus" here is actually more akin to what we have introduced as 'comment' earlier in this chapter.

In an annotation experiment using this distinction, Vesela et al. (2004) used three annotators across four annotation phases on a data set of 441 trees and 6,402 tecto-grammatical nodes in total. The annotators were not the same for all four phases, which is noted as one reason why the agreement varies between 76% and 90% for nodes, and 26% and 36% for trees. Some of the main problem areas reported by Vesela et al. (2004) are the distinction between contrastive topic and focus, between contrastive and non-contrastive topic and the TFA status of nodes in free modifiers, such as adjuncts specifying place and time.

Cook & Bildhauer (2013) Another approach, targeting topic in German, is presented by Cook & Bildhauer (2013). They report on two experiments where aboutness topics have been annotated, and the insights from the first experiment were used to improve the guidelines for the second one. The original guidelines were the ones described by Götze et al. (2007), where aboutness topics are determined by tests such as "An NP X is the aboutness topic of a sentence S containing X if S would be a good answer to the question 'What about X?".

In the first experiment, two expert annotators selected aboutness topics from a small number of pre-defined syntactic constituents in a data set of 588 sentences from the DeReKo corpus (Kupietz et al., 2010). Fleiss' κ (Fleiss, 1971) computed to values between .19 and .57, depending on the sentence category. In the second experiment using the revised guidelines and a data set of 56 sentences from the TüBa-D/Z (Telljohann et al., 2004), four student annotators were trained, who were asked to label each noun phrase as either an aboutness topic or not. In addition, they were asked to state whether they are dealing with a sentence that has a topic/comment structure at all. For the aboutness topic distinction, Fleiss' κ was at .45, while for the topic vs. non-topic structure distinction, it only reached .23. Cook & Bildhauer (2013) state that selecting the aboutness topic among various candidates is non-trivial, as the tests in the annotation scheme can be positive for several candidates in a sentence. Also, they suggest that the binary distinction between topic and topicless sentences may not be reflected in authentic data.

Stede & Mamprin (2016) In a very recent approach, Stede & Mamprin (2016) present the topic-annotated version of the Potsdam Commentary Corpus (Stede, 2004). The annotation comprises several kinds of topics, but in the evaluation Stede & Mamprin (2016) concentrate on aboutness in order to compare their results to the ones of Cook & Bildhauer (2013) described above. Stede & Mamprin (2016) state that while their original annotation guidelines were based on the ones by Götze et al. (2007), they were revised to eventually differ significantly. The improvements draw on insights from Jacobs (2001) and include the criteria of informational separation (topics are presented separately in the linguistic form from the rest of the clause) and addressation (information presented in the comment is stored under the entry of the topic, equivalent to Definition 1).

The agreement study uses 10 texts, while the complete corpus contains 175 annotated texts, making it one of the largest topic-annotated resources. In the agreement evaluation, Stede & Mamprin (2016) separated discourse segmentation from topic annotation, where the results of segmentation were merged into a gold standard before topics were annotated. Two expert annotators were used. In the topic annotation experiment, they followed Cook & Bildhauer (2013) in requiring annotators to on the one hand decide whether they are dealing with a topic/comment sentence, and on the other hand determine for every referring noun phrase or prepositional phrase whether it is an aboutness topic. For the former distinction, Cohen's κ was .6, while for the latter it was .71. Both results are considerably higher than the ones obtained by Cook & Bildhauer (2013), but Stede & Mamprin (2016) note that the number of annotators (four vs. two) and discourse segments (58 vs. 139) differs between the studies.

Focus

Switchboard Calhoun et al. (2005) and Calhoun et al. (2010) report on focus annotation work in the aforementioned Switchboard corpus. They call the notion they annotate 'kontrast' instead of focus, but aim for both implicit and explicit constrasts, contrary to what has been called 'contrastive focus' in the theoretical linguistic literature (cf., e.g., Selkirk 2002). The idea is that an explicit contextual trigger, i.e., one or more relevant alternatives in the sense of Rooth (1992), must be identified first, which then gives rise to an instance of focus. An example of this is shown in (19), which is example (7) of Calhoun et al. (2010).

(19) they're talking about having it [the prison system] as a <u>business</u>... so... the $[\![$government$]\!]_F$ doesn't have to deal with it.

In (19), the topic is the prison system, and the explicit trigger *business* gives rise to an instance of 'kontrast' on the word *government*. Calhoun et al. (2010) specify six different types of 'kontrast' to be annotated in total: 'correction', 'contrastive' (see example above), 'subset', 'adverbial', 'answer', 'other' and 'background'. With the exception of 'background', all of these describe the relationship between the respective 'kontrast' instance and the trigger that gave rise to it. The authors restricted the annotation markables to the word classes noun, verb, adjective, adverb and demonstrative pronoun, since only these were expected to bear 'kontrast' (Calhoun et al., 2005). In an annotation study on 145 Switchboard conversations, two annotators identified 'kontrast' at either the word or the NP level. For the 124,440 words in the data set, Calhoun et al. (2010) found an agreement of $\kappa = .67$ for the binary distinction between 'kontrast' and 'background', and also $\kappa = .67$ for the distinction between different 'kontrast' types. As a main problem area, they identified uncertainty about the scope of 'kontrast', since annotators could choose between the word or the NP level. In addition to that, there was confusion between different types of 'kontrast', and between 'kontrast' and the less well-defined 'other' class.

LISA The aforementioned LISA guidelines developed at the SFB 632 in Potsdam also include a scheme for focus annotation. The working focus definition is that focus is "that part of an expression which provides the most relevant information in a particular context as opposed to the (not so relevant) rest of information making up the background of the utterance" (Götze et al., 2007, p. 179). The scheme distinguishes between 'new information focus' (*nf*) and 'contrastive focus' (*cf*) at the top level. For 'new information focus', the authors distinguish between 'solicited' (*nf-sol*), e.g. licensed by a question, and 'unsolicited' (EMPH). An example of 'solicited' focus is shown in (20), which is example (12) of Ritz et al. (2008). 'Unsolicited' focus is exemplified in (21), which similarly corresponds to (13). For 'contrastive focus', they distinguish the sub-categories 'replacement' (*cf-repl*), 'selection' (*cf-sel*), 'partiality' (*cf-part*), 'implication' (*cf-impl*) and 'truth value' (*cf-ver*). The latter describe the relation of the focus to the element it contrasts with, similar to the Switchboard approach described above. (22) demonstrates basic 'contrastive' focus, with co-indices indicating which focus belongs to which contrastive element.

(20) [Who]$_{nf}$ is reading a book?
[Mary]$_{nf-sol}$ is reading a book.

(21) [Once upon a time, there was a wizard]$_{nf-unsol}$. He [lived in a beautiful castle]$_{nf-unsol}$.

(22) My [older]$_{cf_1}$ sister [works as a secretary]$_{cf_2}$, but my [younger]$_{cf_1}$ sister [is still going to school]$_{cf_2}$.

Parallel to the information status annotation described above, Ritz et al. (2008) employed two undergraduate students to annotate focus in three text types: question/answer pairs, map task dialogues and newspaper commentaries. The agreement results generally are significantly lower than those for information status: token-level κ for the core categories (*nf*, *cf*) is at .51 on the question/answer pairs, at .44 on the map task dialogues and finally at only .19 on the newspaper commentaries. On the NP level, results are somewhat better, but represent only a partial evaluation (due to the fact that there are foci outside of NPs), with the κ-values .62, .48 and .41, respectively. As the main source of disagreements, the authors identify the extent of the focus: one annotator defined focus extensions to phrasal heads rather than whole phrases. This is reminiscent of the problems with the Switchboard approach above, and thus appears to be a central issue in focus annotation which we will return to at the end of this chapter.

DanPASS Paggio (2006) describes focus annotation work in a corpus of spoken Danish. The notion of focus pursued here builds on the work of Lambrecht (1994), who defines focus in terms of presupposition: presupposed content is the background, whereas non-presupposed content forms the focus. Moreover the former is assumed to be optional whereas the latter is obligatory. Additionally, annotators are required to annotate a sentence topic if one can be identified. Example (23), which is (1) from Paggio (2006), demonstrates the identification of both categories.

(23) Så ⟦tager du en lille firkant⟧$_F$ [. . .] Du lægger ⟦den⟧$_T$ ⟦midt på trekanten⟧$_F$ der fungerer som tag

'Then ⟦you take a small square⟧$_F$ [. . .] You put ⟦it⟧$_T$ ⟦in the middle of the triangle⟧$_F$ that functions as a roof.'

A set of general annotation principles are defined that include rules such as 'all sentences must have a focus', and 'focus needs not coincide with a syntactic phrase'. Most importantly, there is a rule about accentuation, 'there must be at least one main accent in a focus domain, but there may be several'. This means that prosody is used as a defining feature for focus. Concerning annotation in general, the author states: "The annotation work relies largely on the coders' intuition, for example to decide what is presupposed information, [. . .]" (Paggio, 2006, p. 1606). The κ values obtained are between .7 and .8, which is rather high. This could be due to the aforementioned prosodic evidence which helps in identifying focus, but also due to the specialized corpus type, a collection of monologues and dialogues, where alternatives and contrasts are typically easier to pin down. As the main sources of disagreement, Paggio (2006) mentions the identification of the left focus boundary, i.e., the extent of the focus.

3.4. Zooming In: Givenness vs. Focus

Having introduced the main notions and having seen how they have been annotated in data by others, the question arises which notion is most relevant for the purpose of assessing the content of answers to questions.

Let us first briefly discuss why the notion of topic is not part of this section's title. As we have stated in our introduction of the notion, it captures what entity or entities a sentence is about and what information is stored under that entity. Consequently, a topic/comment partitioning of the answer in SAA would tell us what the topic of the answer is, and what is said about it. Also, by extending the approach to the question, we could possibly determine whether the topic of the answer is the one introduced by the question.

However, this is only useful in two cases: either the response does not address the question at all, which is easy to detect for an SAA system because the language material in the response is very different from the reference. Or the response does provide relevant information with respect to the question, but associates it with the wrong topic, resulting in a wrong answer. This is however relatively rare, since the topic is introduced by the question and hence reasonably easy to reproduce.

We therefore decided not to pursue the notion of topic any further in the context of the research questions in this thesis, but rather to dive more deeply into givenness and focus in the context of SAA, examining what each notion offers for the task. We will proceed mainly by way of examples, drawn and adapted from the empirical basis we will introduce in chapter 4.

Before launching into the discussion proper, we must state an underlying assumption that we make concerning focus analysis in the SAA setting. We have outlined the relationship between QUDs, QAC and focus in the previous section. In SAA, questions are explicit by definition and it is reasonable to assume that the answer tries to address the question, and if that questions is the current QUD, QAC is established, providing us with the necessary prerequisites for focus analysis in this type of data.

3.4.1. Given/New

Let us first assume the perspective of the given vs. new distinction. We have seen that givenness is a useful notion in predicting prosodic prominence as demonstrated by Schwarzschild (1999), but its role in evaluating answers to questions has yet to be determined.

In (24), we present an example where the question asks for an individual, which is provided by the answer.

(24) Which cultural figure is most important to the people of Salzburg?

A: Mozart [is most important to the people of Salzburg]$_{given}$.

In this example, the answer reprises a fair amount of question material, which can thus be regarded as given in the answer. The only piece of new information is the name *Mozart*, which happens to be what answers the question. No other new information is provided. We can thus state that when the new information is identical to the requested information, the given vs. new distinction works as intended for separating relevant from irrelevant content with respect to the question.

However, this is not the case for all answers. In (25), the answer presented contains several pieces of new information, not all of which is necessary for answering the question.

(25) What do many people think of when they hear Belarus?

A: [Many people]$_{given}$ do not associate vacation with [Belarus]$_{given}$, but rather [think of]$_{given}$ the Cernobyl catastrophe of 1986.

While the answer in (25) contains some given material, like the one in (24), most of the answer material is in fact new information. This does however not automatically make it requested information: the question *What do many people think of...* is actually answered by the phrase *the Cernobyl catastrophe of 1986*. The remainder of new information is unrequested and extraneous, and givenness is not very helpful here in distinguishing relevant and irrelevant information.

An even more extreme case are alternative questions, where all possible alternatives are already present in the question, as in example (26).

(26) Is the apartment in a new building or in an old building?

A: [The apartment is in a new building]$_{given}$.

Here, the question explicitly requires the answer to select either *in a new building* or *in an old building*, which means that requested content must by definition be given. The answer contains no new information at all and is still perfectly acceptable, which shows that the criterion of given vs. new is not helpful here.

We can conclude that the given vs. new distinction is helpful in evaluating answers in cases where new information is identical to requested information. However, it fails to capture the relevant content in cases where extraneous information is present, and where given information is part of the requested content.

3.4.2. Focus/Background

We will now look at the same cases under the perspective of the focus vs. background distinction. Consider example (27), which corresponds to example (24):

(27) Which cultural figure is most important to the people of Salzburg?
A: ⟦Mozart⟧$_F$ is most important to the people of Salzburg.

Since the answer successfully addresses the alternative set of individuals encoded through the question by mentioning *Mozart*, focus selects the relevant information here. However, since *Mozart* is both requested and new, this is not an advantage yet over the given vs. new distinction. Let us look at the second case, shown in example (28), which corresponds to (25):

(28) What do many people think of when they hear Belarus?
A: Many people do not associate vacation with Belarus, but rather think of ⟦the Cernobyl catastrophe of 1986⟧$_F$.

Focus offers the clear advantage here of selecting exactly the information requested, and nothing else. Finally, example (29) shows that focus can also deal with alternative questions, as opposed to the problem exemplified in (26):

(29) Is the apartment in a new building or in an old building?
A: ⟦The apartment is in a new building⟧$_F$.

Since the definition of focus we employ does not build on newness at all, it is not a problem that the alternatives are given in the question here. Focus still selects the relevant information in the answer (in this case the entire answer), regardless of where in the context the alternatives are introduced.

We have illustrated based on corpus examples that focus is more suited than givenness when it comes to evaluating answers to questions, where it is necessary to detect relevant instead of new information. As a result, in this thesis we concentrate on the focus vs. background distinction, though givenness will still turn out to be useful due to the fact that new information often coincides with focused information.

3.5. Issues in Focus Annotation

After reviewing the most prominent work in IS annotation and determining that we are mainly interested in focus for the purposes of this thesis, we will now summarize the issues that arise in focus annotation approaches, and that we will attempt to address in our focus annotation effort described in chapter 6. We identify two main problem areas, which we will describe in more detail below.

3.5.1. Determining Relevant Alternatives

The first major issue concerns determining the relevant alternatives as a prerequisite to identifying focus in the first place. If such alternatives are not explicitly determined in the context, as appears to be the case with the efforts based on the LISA guidelines, focus identification has to rely much more on surface indicators, such as prosodic prominence, and can become relatively arbitrary. In the Switchboard annotation effort, this problem is tackled by marking explicit triggers, such as elements that focus contrasts with. These can be seen as a form of explicit alternatives in the context, and it is therefore not surprising that Calhoun et al. (2010) report higher agreement values than Ritz et al. (2008). We also observe that the availability of explicit alternatives and their nature is closely linked to the type of data used in the annotation project: both Switchboard and DanPASS are dialogue corpora, whose communicative nature and restricted domain presumably make the identification of relevant alternatives easier than this would be the case in e.g. newspaper text. In fact, Ritz et al. (2008) report consistently higher agreement results for question/answer pairs and map task dialogues than for newspaper commentaries across different IS categories. This leads us to the conclusion that, while in principle focus annotation is applicable to all language data, it benefits from data sources in which the nature of the alternatives can be clearly identified. Fortunately, with the reading comprehension corpus we introduce in chapter 4, we have an empirical basis including explicit questions, providing a suitable development basis for a reliable focus annotation approach.

3.5.2. Determining the Extent of the Focus

The second major issue centers around determining the extent of a focus instance: once the approximate location (or nucleus) of a focus has been identified, the question arises what the boundaries of the focus are. If, for example, a noun has been identified as being part of the focus, and it is preceded by a determiner, is the determiner also part of the focus? Almost all approaches we discussed mention this issue as a source of disagreement (Paggio, 2006; Ritz et al., 2008; Calhoun et al., 2010), either because the guidelines leave the decision to the annotator, or because annotators do not properly follow the guidelines. Constraining focus instances *a priori* to certain syntactic elements is however also not a solution: as Ritz et al. (2008) suggest[4] and as we have shown by example earlier in this chapter, the syntactic unit corresponding to a focus instance can vary according to the nature of the alternatives. It thus seems what is needed is a way of testing whether individual words are part of a focus instance or not. We return to this issue in chapter 6 when we discuss our focus annotation scheme.

Summary

In this chapter, we gave an introduction into the general idea of information structure, which is to organize utterances into units of information (or meaning) so that they fit into the discourse. We gave an overview of the three most important distinctions within IS: topic vs. comment, which deals with the entities information is organized around, given vs. new, which categorizes information according to how accessible it is in the discourse, and focus vs. background, which separates an utterance into a part that answers a current (implicit) question and one that does not.

[4] They state that NP-based κ values only represent a partial evaluation for focus, meaning that foci exist outside of NPs.

In the next section of the chapter, we gave an overview of existing annotation approaches for topic, information status and focus, with an emphasis on the latter. Having ruled out topic vs. comment for the purposes of this thesis, we then zoomed in on given vs. new and focus vs. background in the context of evaluating answers to questions, and showed by example how focus represents more accurately the part of an answer that we are interested in.

Finally, we discussed the main problem areas that focus annotation efforts are faced with. We identified two main issues: (1) determining relevant alternatives in the context, so focus can be pinpointed, and (2) determining the extent of the focus, i.e., its borders. Both problems will be addressed in later chapters of this thesis.

Part II.

The Empirical Basis and Our Experimental Sandbox

4. Empirical Basis

In this chapter, we describe the data that serves as a basis for all following experiments, both in SAA and focus annotation. In section 4.1, we first review some desirable characteristics that an empirical basis for meaning comparison and IS analysis should have, before discussing concrete options of naturally occurring data that fit these desiderata. In section 4.2, we then zoom in on the case of reading comprehension corpora and describe the particular corpus we collected, paying close attention to its contents and structure and the corpus creation process. The presentation of the corpus is partly based on Meurers, Ott & Ziai (2010) and Ott, Ziai & Meurers (2012). Finally, we characterize subsets of CREG we use for annotation and evaluation purposes later in this thesis.

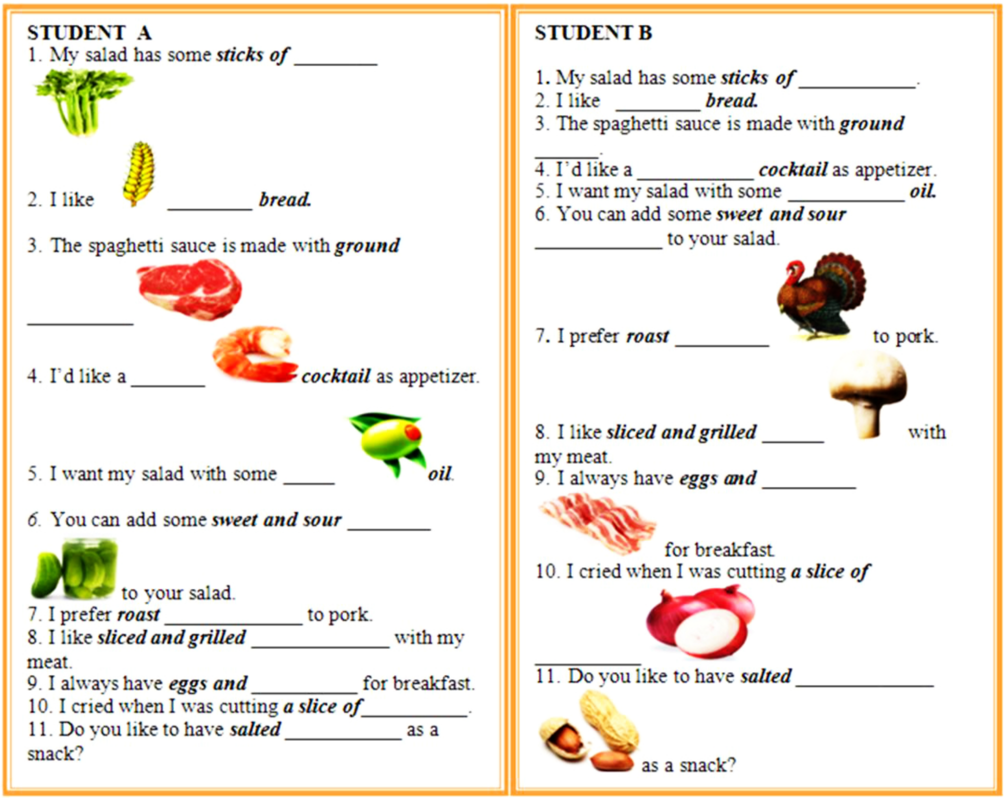
4.1. An Empirical Basis Including an Explicit Task Context

When we evaluate the meaning of a natural language expression, we need some form of reference that the meaning of the expression has to be compatible with. For example, when reading a newspaper article, we constantly try to integrate the statements made in the article with what we already know. For a human with sufficient background knowledge, it is possible to assess the content of a statement.

However, a serious problem arises when the necessary background knowledge is not available: we cannot determine whether a statement's content is adequate. This is especially problematic in automatic approaches, because the knowledge needs to be accessible in machine-readable form and the machine must be able to do some form of reasoning based on the facts it has access to. These issues, while interesting, belong to the field of AI and are outside the domain of language processing proper.

So how does one avoid having to tackle AI problems while still being able to study form variation and information structuring in meaning comparison? In order to answer that question, let us take a look at other language production settings where more explicit context is available. Our assumption will be that the more constrained a language production task is, the easier it is to evaluate the meaning of the language produced.

We have already mentioned that genres such as newspaper text are on the loosely constrained end of the spectrum. While a good candidate for syntactic analysis, as demonstrated by various well-known treebank projects (cf., e.g., Marcus et al. 1993b; Brants et al. 2002), newspaper text does not lend itself to tasks meaning where meaning has to be evaluated. Looking at the other end of the spectrum, we find very tightly constrained language production settings as they occur in learning scenarios, such as fill-in-the-blanks exercises and information gap activities. An example of the latter is shown in Figure 4.1.



STUDENT A

1. My salad has some ***sticks of*** ________

2. I like ________ ***bread.***

3. The spaghetti sauce is made with ***ground***

____

4. I'd like a _______ ***cocktail*** as appetizer.

5. I want my salad with some _____ ***oil.***

6. You can add some ***sweet and sour*** ________

to your salad.

7. I prefer ***roast*** _____________ to pork.

8. I like ***sliced and grilled*** _____________ with my meat.

9. I always have ***eggs and*** _________ for breakfast.

10. I cried when I was cutting ***a slice of***_________.

11. Do you like to have ***salted*** ___________ as a snack?

STUDENT B

1. My salad has some ***sticks of*** ___________.

2. I like ________ ***bread.***

3. The spaghetti sauce is made with ***ground*** ______.

4. I'd like a ___________ ***cocktail*** as appetizer.

5. I want my salad with some ___________ ***oil.***

6. You can add some ***sweet and sour*** ____________ to your salad.

7. I prefer ***roast*** _________ to pork.

8. I like ***sliced and grilled*** ______ with my meat.

9. I always have ***eggs and*** _________ for breakfast.

10. I cried when I was cutting ***a slice of***

____

11. Do you like to have ***salted*** _____________ as a snack?

Figure 4.1.: Information gap activity (Uriarte, 2013)

While one can very clearly define what the intended solutions for the gaps are, the task offers very little room for variation: because of the sentence context and the pictures, the possible solutions are extremely limited and thus can be hard-coded, rendering automatic meaning evaluation unnecessary. One could argue that information gap activities can be pushed to the phrasal and also the clausal level, but they still lack the possibility of unrestricted input, since some sentence context is always given.

Staying within (language) learning scenarios but moving towards more unrestricted tasks, we find formats such as picture description. Given a prompt, the task here is to address the prompt using the information in the pictures. An example is shown in Figure 4.2, where the learner is prompted to describe in their own words what the girl Michelle is doing at a given time.

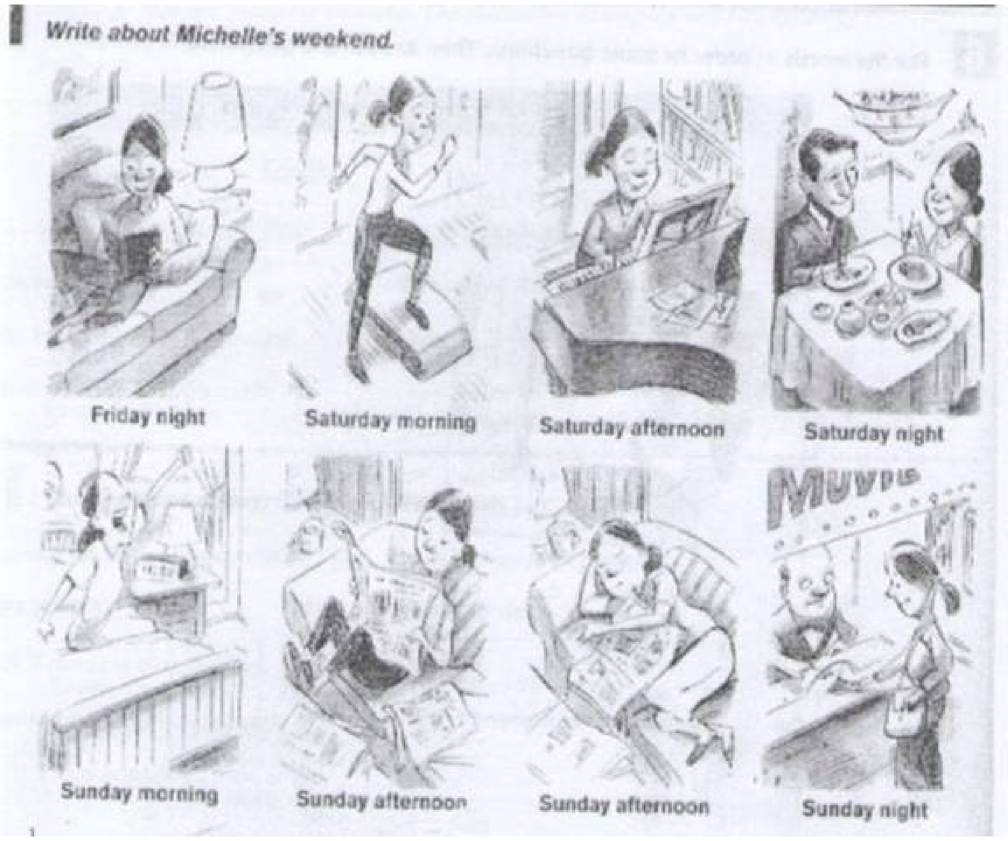


Figure 4.2.: Picture description activity (Razagifard & Rahimpour, 2010)

This type of language production setting clearly yields more form variation, since no explicit sentence context is given, providing learners with the opportunity of completely free input. It is also in principle possible to formulate target answers for each picture. However, it is very hard to control exactly what aspects of the pictures learners will describe, how much information is to be expected in the input, and where the distinction between a correct and an incorrect answer should be. Moreover, the task is harder to automatically evaluate because the information source is not textually encoded, again requiring the encoding of knowledge outside the linguistic system.

It seems clear that we need a setting with linguistically encoded context on the one hand, and enough potential for significant form variation on the other hand. Furtunately, the educational sector does provide such a setting: reading comprehension tasks. They are a meaning-focused activity where learners are supposed to demonstrate their understanding of a reading text by answering specific questions. Since the input is in principle unrestricted, there is ample opportunity for form variation in learner answers, however it is clear what a correct answer should look like due to the explicit reading text and the comprehension questions. Moreover, it is common practice for teachers to formulate target answers based on the text in order to facilitate consistent grading.

In the next section, we therefore present the reading comprehension corpus that served as a basis for the research carried out in this thesis.

4.2. The Corpus of Reading Comprehension Exercises in German

The Corpus of Reading Comprehension Exercises in German (CREG) is a German corpus of answers to reading comprehension questions, collected as part of project A4 of the SFB 833. While most work in SAA targets English, German offers the challenge of richer morphology and freer word order, making it an interesting language for meaning comparison research. The corpus was collected in collaboration with Nina Vyatkina at Kansas University (KU) and Kathryn Corl at The Ohio State University (OSU). Both institutions are large midwestern universities with German programs of substantial size. The reason why collection took place in the US rather than in Germany was homogeneity: American learners of German have a far more homogeneous language background than foreign learners of German living in Germany, where their different native languages and the everyday interactions in German heavily influence their language production. At both locations, teaching assistants were hired for data collection and meaning assessment of the student answers. In order to enable the creation of a highly structured corpus through data entry and annotation by non-technical staff, a special tool was developed: the Web-based Learner Corpus Machine (WELCOME). The tool is presented in more detail in section 4.2.1.

Besides the answers themselves, CREG contains the reading texts, the comprehension questions and the target answers specified by teachers. In addition, student metadata was collected in order to enable research on Second Language Acquisition (SLA) aspects. The different types of data and the relationships between them result in a richly structured corpus, whose layout we present in greater detail in section 4.2.2.

4.2.1. Collection Process

Collection of CREG took place during the four years of the first phase of project A4. As mentioned in the previous section, the data was collected at two different sites in the US, Kansas University (KU) and The Ohio State University (OSU). At both sites, students were observed in their normal classroom behavior, no extra work was required of them beyond consenting to offer their answer data for research purposes. This meant that they did their exercises in the traditional paper-based fashion, without any electronic data immediately available. Moreover, no exercises were altered for the sake of research purposes.

The written exercises then had to be digitized and rated, for which two teaching assistants were hired at each site. The objective was to also observe and capture the rating process in its natural form, so teaching assistants were told to assess the student answers just like any normal assignment. It thus became clear that for a successful corpus creation effort, a software was needed that supports

1. decentralized data entry and annotation,

2. incremental addition of data over a longer period of time,
3. an intuitive interface for non-technical users and
4. a complex interaction of different data types (student answers, target answers, questions, reading texts, and metadata).

A search of the available software for corpus creation yielded that there was no existing tool which met these requirements. We therefore decided to create our own specialized tool, the Web-based Learner Corpus Machine (WELCOME).

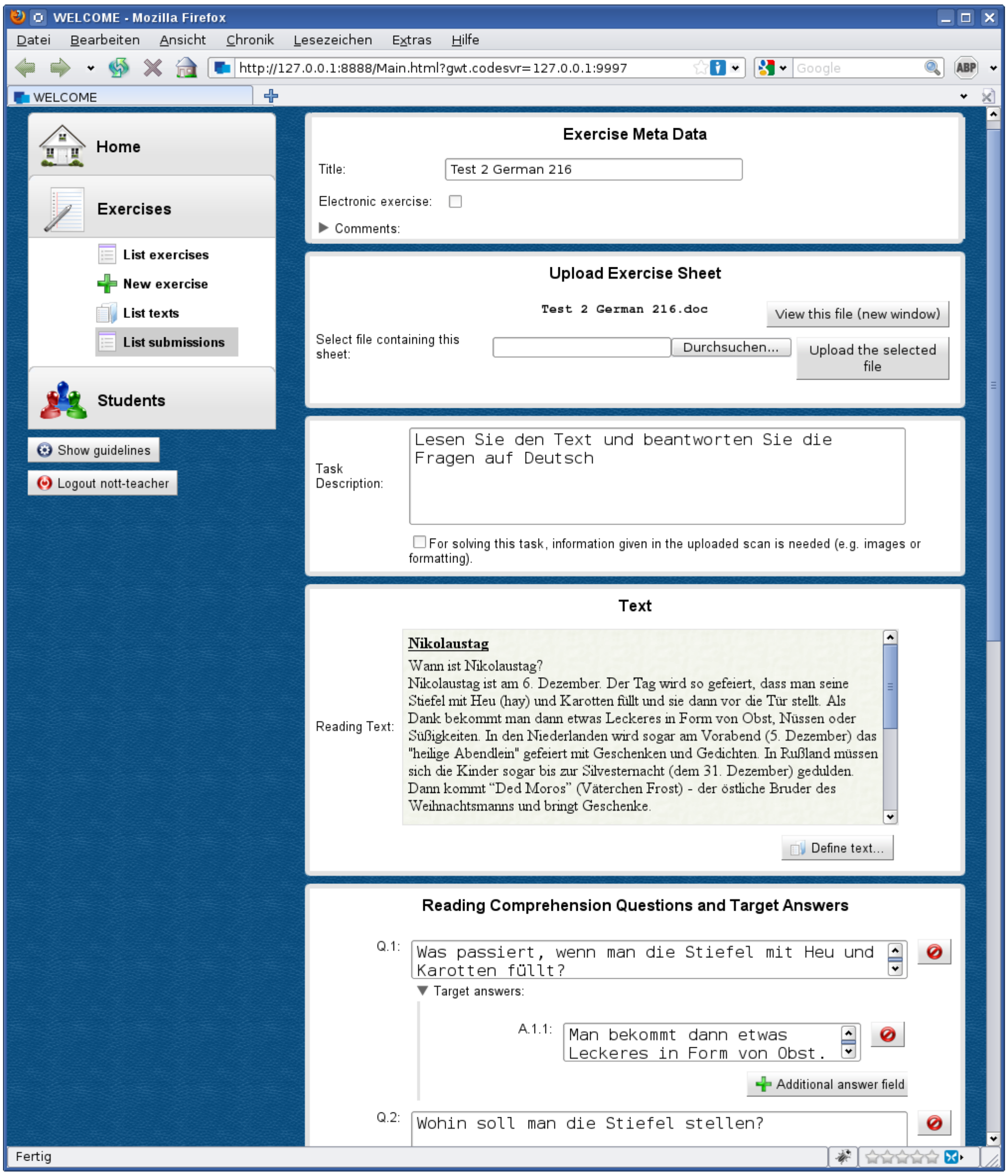
WELCOME is a collection software for reading comprehension corpora, developed by Niels Ott, Georgi Boychev and the author of this thesis, under the supervision of Detmar Meurers in project A4. It addresses the requirements we listed above in the following way: being an online tool, it supports decentralization by being accessible with a regular browser while storing the entered data in one place on a server. The format it is stored in is a relational database system (PostgreSQL[1]), which readily supports the incremental nature of the collection effort by always enforcing a consistent state of the corpus at any time. The user interface was implemented in the Google Web Toolkit[2], which allowed for an intuitive, desktop-like appearance and behavior that the teaching assistants could easily work with. Finally, communication between the user interface and the database system was realized using the industry-grade Hibernate object-relational mapping[3], which enables the transparent storage and retrieval of complex interrelated data structures directly from the application.

The corpus collection workflow was as follows: Starting from a paper version of the respective reading comprehension exercise, the teaching assistants first had to scan and upload that exercise before producing and entering an electronic version of the exercise in WELCOME. In the electronic version, they had to enter the exercise instructions separately from the reading text, since the same reading text can be (and was) reused in another exercise. In the last step of exercise creation, they had to specify the reading comprehension questions and the corresponding target answers. Figure 4.3 shows a screenshot of the exercise creation process.

[1] `http://www.postgresql.org/`

[2] `http://www.gwtproject.org/`

[3] `http://hibernate.org/orm/`



Exercise Meta Data

Title: Test 2 German 216

Electronic exercise: □

▶ Comments:

Upload Exercise Sheet

Test 2 German 216.doc

Select file containing this sheet:

Task Description: Lesen Sie den Text und beantworten Sie die Fragen auf Deutsch

□ For solving this task, information given in the uploaded scan is needed (e.g. images or formatting).

Text

Reading Text:

Nikolaustag

Wann ist Nikolaustag?
Nikolaustag ist am 6. Dezember. Der Tag wird so gefeiert, dass man seine Stiefel mit Heu (hay) und Karotten füllt und sie dann vor die Tür stellt. Als Dank bekommt man dann etwas Leckeres in Form von Obst, Nüssen oder Süßigkeiten. In den Niederlanden wird sogar am Vorabend (5. Dezember) das "heilige Abendlein" gefeiert mit Geschenken und Gedichten. In Rußland müssen sich die Kinder sogar bis zur Silvesternacht (dem 31. Dezember) gedulden. Dann kommt “Ded Moros” (Väterchen Frost) - der östliche Bruder des Weihnachtsmanns und bringt Geschenke.

Reading Comprehension Questions and Target Answers

Q.1: Was passiert, wenn man die Stiefel mit Heu und Karotten füllt?

▼ Target answers:

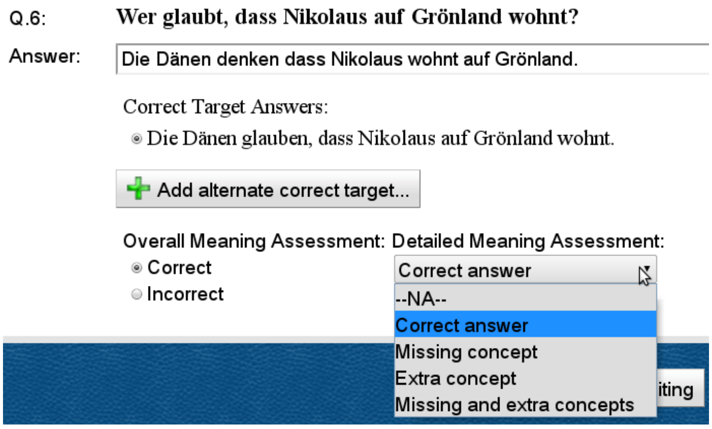
A.1.1: Man bekommt dann etwas Leckeres in Form von Obst.

Q.2: Wohin soll man die Stiefel stellen?

Figure 4.3.: Reading comprehension exercise in WELCOME

Once the exercises were created in the system, teaching assistants could begin to transcribe and assess the corresponding student answers. We required both teaching assistants to provide a transcription of student handwriting, because transcription is already an interpretation process, which may differ between individuals. Each teaching assistant then proceeded with assessing their respective transcription in relation to a specific target answer which they had to choose. If the student answer demonstrated a substantially different but correct way to answer the question, teaching assistants had the option to dynamically add a new target answer.

Assessment was done in two different schemes, which we called **binary** and **detailed**. In the **binary** scheme, there are only two categories: 'correct', which means the answer is appropriate, and 'incorrect', which means it is not. In the **detailed** scheme, the idea was to encode the nature of divergence in meaning with respect to the target answer, where the possibilities are 'correct answer' (no divergence), 'missing concept', 'extra concept', 'missing and extra concepts' and 'non-answer' (for off-topic responses). Figure 4.4 shows an example of the assessment process.



Q.6: **Wer glaubt, dass Nikolaus auf Grönland wohnt?**

Answer: Die Dänen denken dass Nikolaus wohnt auf Grönland.

Correct Target Answers:

● Die Dänen glauben, dass Nikolaus auf Grönland wohnt.

Overall Meaning Assessment: ● Correct ○ Incorrect

Detailed Meaning Assessment: Correct answer

Figure 4.4.: Answer assessment in WELCOME

As mentioned previously, we also included student metadata in the corpus. In order to enable longitudinal studies in the future where the same learner can be tracked over time, the teaching assistants collected metadata from the students via a questionnaire each semester (see section 4.2.2 for details on the metadata). The results of the questionnaire were accumulated in a spreadsheet and imported into the WELCOME database.

4.2.2. Corpus Layout and Characteristics

We now take a closer look at the data elements in the corpus and the relations between them. CREG is a highly structured corpus, which is due to the multitude of different types of information involved: teachers create exercises consisting of reading texts, instructions, and questions. Students in different courses with different background and proficiency levels complete these exercises by producing answers to the questions. Teachers then use constructed target answers to assess the answers that the students produced.

Figure 4.5 shows a diagram of data objects in the corpus and the mappings between them. A 1:n-mapping means that for every one element on the source side of a relation, there can be *n* elements on the destination side. n:1-mappings work in the opposite fashion. For example, for every Student Submission there is exactly one Reading Exercise that the student worked on. Likewise, every Reading Exercise has n Comprehension Questions, and so on.

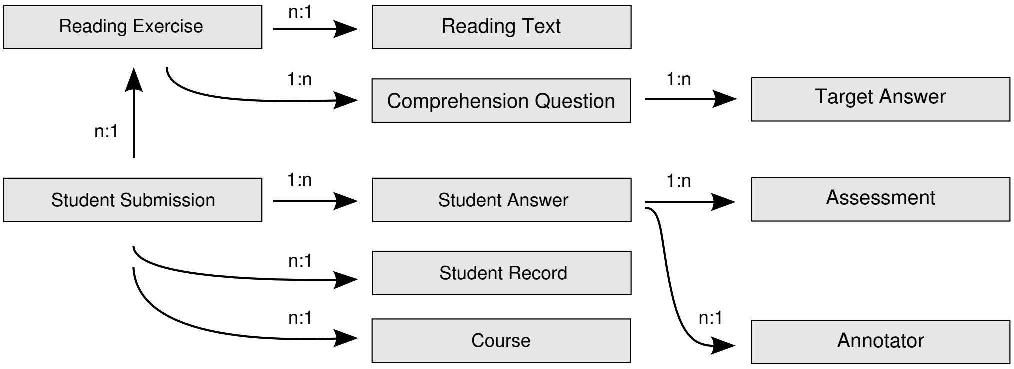


Figure 4.5.: CREG corpus structure

All of these objects and mappings are stored in a relational database, the source format of CREG. Because of this relational backbone, the database can in principle answer any research-related query that can be expressed in relational terms. Examples would be "Give me all students that worked on a specific exercise" or "Give me all exercises that student X worked on" or "Give me all courses where exercise X was administered".

For the purposes of SAA in the A4 project and for sharing CREG with other researchers working in the same field, we implemented an export function from the relational source format to a hierarchical XML structure containing only texts, questions, and corresponding target and student answers with assessments, but excluding any other information present in the database, such as student metadata, because it is not immediately relevant to the task of SAA.

However, student metadata may very well be of interest to other research fields, such as SLA, where it is often necessary to track the development of individual students. The metadata we collected includes background information such as age, gender, previous exposure to German, other foreign languages learned, and time spent in a German-speaking country. Tracking students' progress through multiple courses would in principle be possible with a single metadata record collected at the start of the first course. However, some of the metadata, such as time spent in Germany, may be subject to change. For this reason, we collected metadata at the start of each semester, which yields a series of metadata records for each student, allowing longitudinal studies to take into account how the students develop over time.

4.2.3. CREG Subsets

In total, CREG contains 148 reading texts, 1,517 reading comprehension questions, 1,642 target answers provided by the teachers, and 35,013 learner answers written by American learners of German, making it the largest German reading comprehension corpus that is currently available. However, the full corpus includes material that may not be useful to the research goal at hand. For example, very short answers (less than 5 tokens) tend to be uninteresting in terms of variation. Also, in cases where the two annotators disagreed on the meaning evaluation, it is unclear what the gold standard should be, so one might want to exclude these cases.

For these and other reasons, several subsets of CREG were compiled over the duration of the project. Since these subsets will be used and referenced later in this thesis, we will give an overview of each one's characteristics here.

CREG-1032

CREG-1032 is a balanced subset containing 1032 student answers, 223 target answers, 177 reading comprehension questions and 31 reading texts. It was compiled in 2011, roughly at the halfway point of the corpus collection process. Its main purpose was to provide a first testbed for Short Answer Assessment in the German language. For this purpose, it was important to get an even class distribution that would set the random baseline for the distinction between appropriate and inappropriate answers to 50%. Another consequence of the mainly computational purpose is that CREG-1032 only contains answers for which the two annotators agreed on the meaning assessment, in order to ensure a consistent gold standard that automatic approaches can be compared against. Furthermore, to ensure sufficient form variation, the answers had to be at least five tokens in length, which excludes very short elliptic answers, as in "Who ate the cake?" "Peter.". The resulting average token length was 11.87 tokens. Finally, in the same vein of obtaining more interesting form variation, we only included answers from intermediate courses and above.

CREG-5K

CREG-5K is essentially an updated version of CREG-1032 with the same characteristics, but compiled after the corpus collection effort was completed. It contains 5,138 student answers, 966 target answers, 877 reading comprehension questions, and 96 reading texts. In terms of student answers, it is thus roughly five times as large as CREG-1032, and is also a balanced set. The average token length is 11.58. Because CREG had grown substantially, some filtering of the texts was now also necessary in addition to the answer-related criteria described earlier: duplicate and near-duplicate reading texts were excluded, and three cases where the reading text contained fill-in-the-blank elements were also removed. Also, we found that some of the answers had been given in English, so these were semi-automatically detected and removed from the final answer subset.

CREG-2155

CREG-2155 is a random sample of CREG-5K that has no overlap with CREG-1032 in terms of questions and answers. The intention was to create a data set for the second round of manual focus annotation (see section 6.3) which has approximately the characteristics of CREG-5K but is smaller in size so that annotation can be finished within a reasonable time frame. Answers were sampled proportionately for each question, aiming for a corpus size of roughly 2000 student answers. The resulting corpus contains 2,155 student answers and 767 target answers to 728 questions on 83 texts.

CREG-17K

CREG-17K is a snapshot of the full corpus from the time of the study reported in Ott, Ziai & Meurers (2012), roughly at the halfway point of data collection. As the name implies, it contains about 17,000 student answers, although only 10,083 of them are annotated by two annotators. The subset is not balanced. Its main purpose was to make the agreement study in Ott, Ziai & Meurers (2012) possible, where we compared the annotators both for the binary and the detailed assessment scheme. We found a binary percentage agreement of 88.5% ($\kappa = .71$) for the KU part, and an agreement of 85.7% ($\kappa = .57$) for the OSU part. For the detailed categories, the percentage agreement was 86% ($\kappa = .77$) for the KU part and 70.6% ($\kappa = 0.47$) for the OSU part. The noticeably much lower values for the OSU corpus were traced both to a very skewed distribution in binary assessment (the vast majority of answers is correct) and a difference in understanding the categories for the detailed assessment.

CREG-23K

CREG-23K is the biggest subset available with two annotations for each student answer. It contains 23,147 student answers to 1059 reading comprehension questions on 113 reading texts. The number of target answers is 1302 for annotator 1, and 1336 for annotator 2. Date from entry-level courses was omitted again, and the minimum token length was 4, with an average token length of 12.43. Also, similar to CREG-5K unsuitable texts and non-German answers were removed. The overall binary agreement for this set is 86.1% with

$\kappa = 0.6$.

CREG-ALL

CREG-ALL is the full corpus as completed at the end of the four-year collection phase, without any filtering criteria whatsoever. It contains 35,013 student answers to 1,517 questions on 148 reading texts. There are 1,642 target answers corresponding to the questions.

CREG-TUE

For the purpose of comparison with learners, a corpus with answers by native speakers was also compiled, called CREG-TUE. The speakers were a sample of 100 students recruited from the population in Tübingen. They provided 3,546 student answers to 143 reading comprehension questions on 21 texts. The student answers were then rated by the same annotators that rated the OSU learner answers, using 180 target answers. The texts and questions were chosen in such a way that they represent the overlap in material between OSU and KU, in order to maximize comparison possibilities with learners answering the same questions.

4.2.4. Quantitative Overview

Table 4.1 presents an overview in terms of numbers of what each CREG subset contains. For reading texts, questions, target answers and student answers, we list the respective count, total number of tokens, and average number of tokens per item. We also list the average number of questions per text and student answers per question, since especially the latter is important when choosing training and testing data partitions.

One important characteristic needs to be highlighted above others here. The average token length of reading texts is far lower for early CREG subsets, with 318.33 for CREG-1032, 376.89 for CREG-17K, and 348.00 for CREG-TUE. This is in stark contrast to later CREG subsets[4], which can have up to 1083.08 tokens per reading text (CREG-2155), suggesting that the complexity of the texts, and hence of the whole reading comprehension task, rose significantly.

[4] See appendix A for example texts from CREG-1032 and CREG-5K

This assumption is backed up by recent research in reading comprehension (cf., e.g., Eason et al. 2012) where the impact of question types and reading texts on comprehension was investigated. As we will see in later chapters, this likely also affects classification performance.

Summary

We presented the empirical basis for our research efforts in this thesis, the Corpus of Reading Comprehension Exercises in German (CREG). The collection of CREG was part of the research of this thesis, and we contributed significantly by co-developing the Web-based Learner Corpus Machine (WELCOME). CREG is the foundation for both our Content Assessment experiments (see next chapter and chapter 8) and for our analysis of focus in answers (see chapters 6 and 7).

We first described some desirable characteristics for an authentic data source for Content Assessment: an explicit, linguistically encoded task context and free-text answers with sufficient form variation. Having decided that reading comprehension exercises fit our needs, we proceeded to describing CREG itself. We started with the collection process which included the development of the WELCOME tool. We then outlined the corpus layout and the relationships between its ingredients, before discussing several CREG subsets of interest for research described in later chapters. Finally, we concluded the chapter with a quantitative overview of CREG subsets.

	CREG-1032	CREG-5K	CREG-2155	CREG-17K	CREG-23K	CREG-ALL	CREG-TUE
Reading Texts	31	96	83	75	112	148	21
total Tokens	9868	93569	89896	28267	108437	118541	7308
avg. Token #	318.33	974.68	1083.08	376.89	968.19	800.95	348.00
Questions	177	877	728	624	1059	1517	143
total Tokens	1915	10366	8751	6334	12274	16839	1346
avg. Token #	10.82	11.82	12.02	10.15	11.59	11.10	9.41
Q's per text	5.71	9.14	8.77	8.32	9.46	10.25	6.81
Target Answers	223	966	767	716	1334	1642	178
total Tokens	2953	15350	12748	9344	21157	24384	2264
avg. Token #	13.24	15.89	16.62	13.05	15.86	14.85	12.72
Student Answers	**1032**	**5138**	**2155**	**16947**	**23147**	**35013**	**3545**
total Tokens	12291	60372	26356	151337	292578	360984	29778
avg. Token #	11.91	11.75	12.23	8.93	12.64	10.31	8.40
SA's per question	5.83	5.86	2.96	27.16	21.86	23.08	24.79

Table 4.1.: Statistics of different CREG subsets

5. Experimental Sandbox: the CoMiC System

In this chapter, we present the Comparing Meaning in Context (CoMiC) system, which is the basis for the integration of focus into SAA in chapter 8. We first review the key aspects of the Content Assessment Module (CAM) which formed the foundation for CoMiC in section 5.1. We then discuss the design and implementation of CoMiC in the UIMA framework (Ferrucci & Lally, 2004) in section 5.3 (partly based on Meurers, Ziai, Ott & Bailey 2011a), and the adaptation of the system to German in section 5.4 (partly based on Meurers, Ziai, Ott & Kopp 2011b). Finally, we evaluate CoMiC on different subsets of CREG in section 5.5.

5.1. The Content Assessment Module as the basis for CoMiC

CAM is a system developed by Stacey Bailey in the context of her PhD thesis (Bailey, 2008) in collaboration with Detmar Meurers (Bailey & Meurers, 2008). It was built to diagnose meaning errors in answers by learners of English to reading comprehension questions at The Ohio State University. There are three stages to the system:

1. **Annotation** of the student and target answer with linguistic information.
2. **Alignment** of answer parts using the annotated linguistic information.
3. **Classification** of student answers based on the number and the kind of alignments found.

The linguistic **annotation** performed in stage 1 ranges from tokenization and part-of-speech tagging to pronoun resolution. The idea is that every linguistic annotation step performed enables another type of comparison between linguistic units in answers by providing an abstraction: lemmatization abstracts over inflection, spell checking abstracts over typos, and so on. Table 5.1 lists the NLP components used in CAM.

Annotation Task	**NLP component**
Sentence Detection, Tokenization, Lemmatization	MontyLingua (Liu, 2004)
Lemmatization	PC-KIMMO (Antworth, 1993)
Spell Checking	Edit distance (Levenshtein, 1966), SCOWL word list (Atkinson, 2004)
Part-of-speech Tagging	TreeTagger (Schmid, 1994)
Noun Phrase Chunking	CASS (Abney, 1997)
Lexical Relations	WordNet (Miller, 1995)
Similarity Scores	PMI-IR (Turney, 2001; Mihalcea et al., 2006)
Dependency Relations	Stanford Parser (Klein & Manning, 2003)

Table 5.1.: NLP components used in CAM, (Bailey & Meurers, 2008, p. 110)

In the **alignment** stage, CAM then uses the abstractions to compute alignments between equivalent tokens, chunks and dependency triples. The problem arises that for a given token, chunk or dependency triple, there can be multiple alignment possibilities. Thus, CAM needs to choose a globally optimal alignment between student and target answer given local alignment possibilities. The system does so by employing the Traditional Marriage Algorithm (TMA, Gale & Shapley 1962) which optimizes alignments with respect to one side, in this case the student answer. Figure 5.1 shows an example with alignments between a target and a student answer to a question. Here, the blue dotted line represents a pronominal alignment between *He* and *Bob Hope*, the green dot-dashed one a surface token alignment between *was* and *was*, and the red dashed one a similarity alignment between *in his house* and *at home*.

Given one definite alignment configuration with respect to a student-target answer pair, the system proceeds to **classifying** student answers based on the number and kind of alignments found. Classification is based on a set of 13

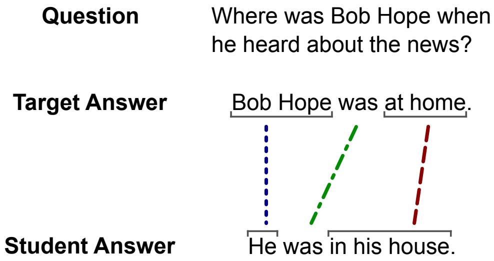


Figure 5.1.: CAM alignment example from Bailey & Meurers (2008, p. 110)

features which express the number and type of alignments found between student and target answer. Table 5.2 lists all the features used in detail. Features 1–7 relate to the number of respective linguistic units (token, chunk, dependency triple) aligned, whereas features 8–13 are concerned with the level of abstraction on which the alignments were made.

Classification itself is done with TiMBL (Daelemans et al., 2007), a memory-based learner which provides an implementation of the *k*-nearest-neighbor algorithm.

The system was trained and evaluated on a corpus of 566 answers to reading comprehension questions, written by foreign learners of English at The Ohio State University as part of their regular homework assignments. The 566 learner answers were divided into a training set (311 answers to 47 questions, called 'development set' by Bailey & Meurers 2008) and a test set (255 answers to 28 questions). Each of the answers had been rated by two annotators with respect to appropriateness of meaning, and answers where the two annotators did not agree were excluded from training and testing. We hereafter refer to this corpus as the Corpus of Reading Comprehension Exercises in English (CREE).

Performance was evaluated in terms of classification accuracy on both the training set, using leave-one-out testing, and on the test set. In the former case, the accuracy was 87% for the binary classification task, and in the latter case, accuracy even rose to 88%. Bailey & Meurers (2008) however noted that neither the training nor the test set was balanced, instead showing a strong skewedness towards correct answers (71% in the training set and 84% in the test set), which introduces a bias for machine learning approaches towards the majority class and makes results seem better than they actually are. This is readily admitted

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

Table 5.2.: Features used for classification in CAM, Bailey & Meurers (2008, p. 112)

to by Bailey & Meurers (2008): in an experiment using a reduced but balanced version of the training and test sets, they report an accuracy of 78% for the training set and 67% for the test set.

5.2. Shortcomings of the Content Assessment Module

The CAM approach sketched in the previous section provides a good starting point as far as the empirical and conceptual basis is concerned. But given its nature as a pilot study into content assessment, the authors did not focus on the NLP architecture and data structure choices. When pursuing this strand of research further, on the practical side questions arise on how such an approach is best realized in a general NLP architecture. On the one hand, it should support modular experimentation and development of content assessment approaches for other languages and research questions, such as the research carried out in this thesis. It should also facilitate integration into current architectures motivated for ICALL systems such as TAGARELA (Amaral, Meurers & Ziai, 2011) or the new FeedBook project[1], where the idea is to provide immediate, automated feedback to students based on NLP analysis. On the theoretical side, a number of research issues present themselves, of which the investigation of the role of the context and information structure on content assessment is most central to us here. Besides this main goal, a more dynamic integration of different levels of linguistic representation, which would also benefit from a general and flexible NLP architecture and explicit data structures considerations is also a desirable characteristic. For these practical and theoretical reasons, we pursue an architecture satisfying the following requirements:

- **Representations and alignment:** CAM only aligns tokens to tokens, chunks to chunks, etc. However, in general the same meaning can be expressed by linguistic units of different complexity and type, e.g., the token *initially* could be aligned to chunk *in the beginning*. Thus, alignments between different representations should be supported.

- **Marking contextual relevance of material:** Some parts of the student and target answer, such as material already given in the question (a point to which we return from time to time in this thesis) or punctuation, should not be taken into account when doing a semantic comparison. The original CAM simply deleted such material from the answers, destroying syntactic structures and leaving the source text incoherent. A mechanism is needed which excludes the relevant units from alignment but otherwise leaves the answers intact.

- **Explicitness of data structures and modularity of analyses:** As it is not clear from the start which NLP tool will perform best for a given task, we need a way to make explicit the data structures we want to work with regardless of which particular tool will provide them. Moreover, new

[1] https://www.uni-tuebingen.de/en/research/core-research/collaborative-research-centers/sfb-833/knowledge-transfer.html

analysis components should be straightforward to add without interfering with the ones already present in the system.

Besides these architectural issues, CAM has also not been scaled up and tested on data sets in the area of 1,000 or more answers. As Bailey & Meurers (2008) state themselves, "more extensive testing with a larger corpus is needed", because a good result on one relatively small corpus does not necessarily mean the method is successful on bigger and more diverse corpora, which possibly come from other sources. And finally, the system only demonstrates its good performance for English, with no indication on whether a similar approach would be viable for other languages. We will return to both these issues in section 5.4.

In the next section, we will therefore present CAM's re-design and re-implementation in the A4 project, the CoMiC system. As we stated in the introduction to this chapter, it forms the basis for all meaning assessment experiments described in this thesis.

5.3. Building CoMiC Using the UIMA Framework

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

In CoMiC, each NLP tool we use (see Table 5.3) is encapsulated as a UIMA Annotator that contributes a specific analysis result to the CAS. Figure 5.2 shows the overall CoMiC architecture. A UIMA Collection Reader takes care of reading in the corpus data and setting up the initial CAS before it is enriched with annotations. While such a variety of parallel analysis results would pose problems for most file-based annotation formats, they are not problematic for UIMA, because annotations are typed and stored in a stand-off manner in a common index, hence they integrate well and do not interfere with each other.

Annotation	**CoMiC-EN**
Sentence Detection, Tokenization	OpenNLP `https://opennlp.apache.org`
Lemmatization	morpha (Minnen et al., 2001)
Spell Checking	Edit distance (Levenshtein, 1966), SCOWL word list (Atkinson, 2004)
Part-of-speech Tagging	TreeTagger (Schmid, 1994)
Noun Phrase Chunking	OpenNLP
Lexical Relations	WordNet (Miller, 1995)
Similarity Scores	PMI-IR (Turney, 2001; Mihalcea et al., 2006)
Dependency Relations	MaltParser (Nivre et al., 2007)

Table 5.3.: NLP tools used in the English CoMiC system

Before alignment takes place, two components take care of marking material that is not to be included in alignment: the givenness filter marks every word whose lemma appears in the question, and the punctuation filter marks all punctuation tokens. Thanks to the explicit data structures, marking can simply be done by setting a Boolean feature on the type *Token* to a certain value. Alignment modules can then check this value and exclude unwanted material.

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

When all alignments have been determined and the TMA has selected the

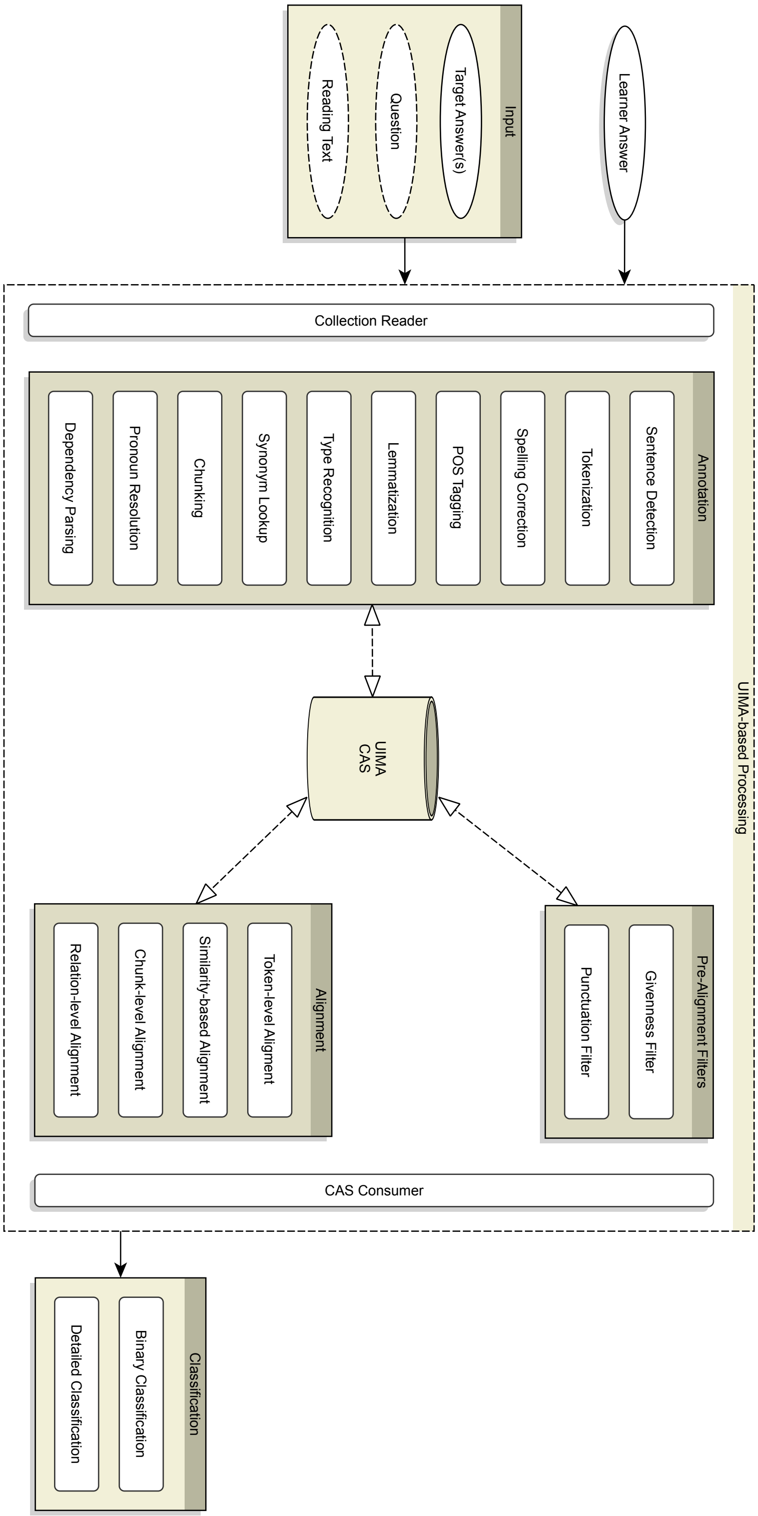


Figure 5.2.: CoMiC architecture overview

optimal configuration, a UIMA CAS Consumer uses the alignment information in the CAS to extract features for training or calling the classifier, for which we use either TiMBL (Daelemans et al., 2007) (as in the original CAM), or a memory-based classifier from the WEKA package (Hall et al., 2009). At this point, UIMA-based processing ends and the feature configurations are written to a simple text file that the classifier can read.

For the purpose of comparing CoMiC-EN (CoMiC for English) to the original CAM approach, we evaluated it against the same data set, the CREE corpus, which is described in section 5.1 in more detail. The memory-based learner TiMBL was trained on the 311 student and target answers from the training set and evaluated via leave-one-out testing, and against the 255 student and target answers from the test set. We used the same seven distance measures with TiMBL as in the original implementation: overlap, Levenshtein, numeric overlap, modified value difference, Jeffrey divergence, dot product and cosine distance. Different distance measures reflect different ways of comparing features in memory-based learning, and instead of relying on any single one of them, the best choice was automatically selected according to a majority voting of the distance measures for each instance.

The results obtained are summarized in Table 5.4.

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

Table 5.4.: Evaluation results of the original CAM and CoMiC-EN

We report two numbers for both the development set and the test set: *Binary Classification* refers to the accuracy achieved in the task of deciding whether a student answer was correct or incorrect. *Detailed Classification* refers to the accuracy in predicting the correct detailed assessment: correct, missing concept, extra concept, blend, or non-answer. Both classification tasks were carried out using the 13 features in Table 5.2.

As aimed for, the performance of CoMiC-EN using the new architecture reaches the same high level as the original CAM implementation. There are slight differences, which are to be expected given that, as we saw in Table 5.3, different NLP tools were used for five of the nine annotators. But in an architecture making use of such a wide range of parallel representations for the alignments, the specific choice of NLP tools does not seem to be crucial to the performance of the overall approach.

5.4. Extension to the German Language

With the English CoMiC system (CoMiC-EN) in place and working, the next step was to implement a system for German based on the same architecture. Because of the architecture's modularity, implementing the German system was essentially a matter of exchanging any English-specific components or models for German-specific ones. The general procedure, including how the alignments are selected and the answers are classified, remained the same. Table 5.5 lists the NLP components used in the German system.

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

Table 5.5.: NLP tools used in the German system

Figure 5.3 shows a German example of the type handled by CoMiC, with a question (Q), a target answer (TA), a student answer (SA), and different alignments between TA and SA. The token alignments shown occur either directly on the `Token` level, as with *und* and *Schiffsdiesel*, on the `Spelling` level, as with *zon–von* and *Fish–Fisch*, or on the `SemType` level (semantic type derived from GermaNet), as with *Geruch–Gestank*.

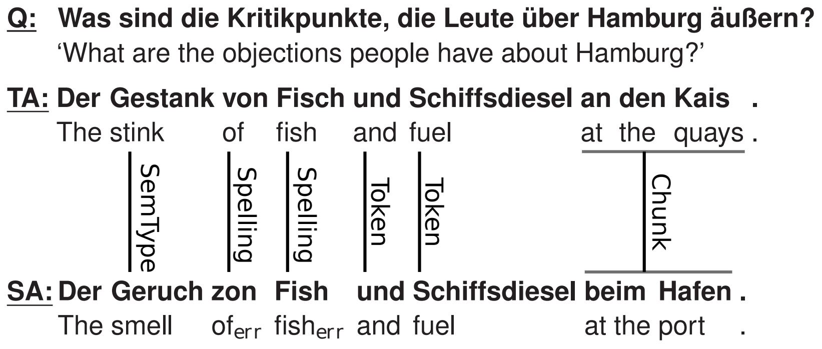

Q: Was sind die Kritikpunkte, die Leute über Hamburg äußern?
'What are the objections people have about Hamburg?'
TA: Der Gestank von Fisch und Schiffsdiesel an den Kais .
The stink of fish and fuel at the quays .
SemType
Spelling
Spelling
Token
Token
Chunk
SA: Der Geruch zon Fish und Schiffsdiesel beim Hafen .
The smell of$_{err}$ fish$_{err}$ and fuel at the port .

Figure 5.3.: CoMiC alignment example on German data

We evaluated CoMiC-DE using the CREG-1032 data set introduced in section 4.2.3, where one part comes from Kansas University and the other from The Ohio State University. Both of these contain only records where both annotators agreed on the binary assessment (appropriate/inappropriate meaning). Each set is balanced, i.e., it contains the same number of appropriate and inappropriate student answers.

In training and testing the TiMBL-based classifier, we followed the methodology previously described, where seven classifiers are trained using the different available distance metrics. Training and testing was performed using the *leave-one-out* scheme (Weiss & Kulikowski, 1991) and for each item the output of the seven classifiers was combined via majority voting.

The classification accuracy for both subsets of CREG-1032 is summarized in Table 5.6. We report accuracy and the total number of answers for each data set.

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

Table 5.6.: Classification accuracy for the two data sets

The 84.6% accuracy figure obtained for both data sets shows that CoMiC-DE is quite successful in performing content assessment for the German data collected so far, a result which is competitive with the one for English obtained by Bailey & Meurers (2008), who report an accuracy of 78% for the binary assessment task on a balanced English data set, as described in more detail in section 5.1.

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

5.5. Performance on Different CREG Subsets

In previous sections, CoMiC was evaluated on relatively small sets of data, since they were what was available at the time. However, as more of CREG was collected and hence more quality gold standard data became available, we were able to test it also on bigger CREG subsets.

We obtained results for the bigger CREG-5K subset (see section 4.2.3) using two different train/test setups, 10-fold cross-validation and a randomly chosen split with roughly 80% training data and 20% test data. For the latter, we created two versions, one based on held-out answers and one based on held-out questions, i.e., all answers to a sampled question are held out. This creates a more challenging test case since the system has never seen similar answers to the ones in the test set before.

The results are summarized in Table 5.7. Note that newer versions of CoMiC use the *k*-nearest-neighbor implementation of WEKA (Hall et al., 2009) instead of TiMBL, since the performance is comparable and WEKA is more convenient to use and integrate. Moreover, in order to further strengthen CoMiC's flexibility and ease of use, the system was ported to uimaFIT (Ogren & Bethard, 2009), a front-end library to UIMA which includes powerful mechanisms to configure and test components. Finally, CoMiC was adapted to use the DKPro Core component repository and type system (de Castilho & Gurevych, 2014), which enables access to a host of new analysis components that can be integrated with little additional work[2]. As a result of this significant change in the system, the numbers reported from here on through the remainder of the thesis can vary slightly when compared to results on the same data sets with the early CoMiC.

	10-fold CV	Unseen answers	Unseen questions
# of answers	5138	1001	1121
Maj. baseline	50.0%	51.3%	51.1%
Accuracy	83.1%	81.5%	78.8%

Table 5.7.: Results for the CREG-5K test sets

One can see that the cross-validation test case is the easiest among the ones presented here, resulting in the highest accuracy of 83.1%. The held-out 'unseen answers' test set presents a harder challenge, coming out at 81.5%. As expected, the hardest test case for CoMiC is the transfer to new questions, demonstrated by the lowest accuracy of 78.8%. Note that the two held-out test sets are not exactly balanced: 'unseen answers' has a slightly higher proportion of correct answers (51.3%) and 'unseen questions' leans slightly towards incorrect answers (51.1%). However, all results are on a similarly high level and far above the baseline.

Besides investigating results on CREG-5K in more detail, it is also interesting to look at CoMiC results on different CREG subsets, as they differ in size and other characteristics (see section 4.2.3). Table 5.8 summarizes the results for 10-fold cross-validation across different CREG subsets.

Looking at Table 5.8, we notice that the early CREG-1032 yields the best results. However, the difference to CREG-5K is not dramatic (83.7% vs. 83.1%). CREG-2155, a subset of CREG-5K which shares no questions or answers with CREG-1032, seems to be a little harder (82.3%), suggesting that later material is more challenging to assess automatically. This could be due to the fact that the reading texts became far longer and hence more complex (see Table 4.1 in

[2] We are greatly indebted to Björn Rudzewitz for his work on porting CoMiC to DKPro.

Data set	# of answers	Majority baseline	10-fold CV
CREG-1032	1032	50.0%	83.7%
CREG-2155	2155	51.4%	82.3%
CREG-5K	5138	50.0%	83.1%
CREG-23K A1	23146	74.0%	81.4%
CREG-23K A2	23146	80.2%	84.5%

Table 5.8.: Accuracy for different CREG subsets

chapter 4 for statistics, and appendix A for examples).

Finally, the much larger CREG-23K proves to be a problematic testbed for two reasons: first, the two annotators only agreed on the meaning assessment in 86.1% of the cases ($\kappa = .6$), and the inconsistency leads to noise in classification. And second, it is very skewed towards correct answers, with a 74% majority baseline for annotator 1 (A1) and an even higher 80.2% majority baseline for annotator 2 (A2). For these reasons, the classification performance obtained is to be taken with a grain of salt and is less reliable than the other results we report.

5.6. Application to Other Tasks

CoMiC has also been applied successfully to other tasks than SAA. In Rudzewitz & Ziai (2015), we adapted CoMiC to the problem of answer selection in community question answering in Task 3 of SemEval 2015 (Màrquez et al., 2015): given a forum question and a sequence of answers, the task is to classify each of the answers as 'definitely relevant', 'potentially useful' or 'bad or irrelevant'. Since no target answers are available in this setting, we instead aligned the question with each answer, and developed several question–answer features which try to approximate the relationship between what the question asks for and what the answer supplies. The results were moderate, but the approach was noted for being the only one that tackles the problem using a sequence classification approach in an attempt to consider individual answers in the context of the thread.

Rudzewitz (2016b) extended this approach to an investigation of which information sources are most useful in which task. To that end, he extracted "over

250 different features from five information sources: question features, answer features, question–answer features, answer–answer features, and user features". He compared the impact of these features on both the abovementioned community question answering problem, and on traditional Question Answering (QA). Results show that domain adaptation and modeling the relationship between question and answer is most effective for traditional QA, but that analyzing the properties of an answer and how it integrates into the thread context is more beneficial for community QA.

Finally, Rudzewitz (2016a) connected SAA to plagiarism detection and showed that CoMiC performs highly competitive in this task when augmented with traditional authorship attribution features. In addition, he carried out the reverse experiment and applied the augmented CoMiC to SAA again, demonstrating that plagiarism features also have a positive impact on SAA.

Summary

We presented the CoMiC SAA system, which forms the basis for all Content Assessment experiments later reported in this thesis. CoMiC is based on the conceptual approach of CAM (Bailey & Meurers, 2008), and is an alignment-based system which classifies answers with regard to content by comparing them to pre-specified target answers. We showed that CoMiC's architecture is superior to CAM's in several regards, such as representation of annotations and exchangeability of processing components. We then demonstrated that CoMiC's performance is on a par with CAM on the English CREE corpus, before describing how CoMiC's architecture enabled a transfer of the approach to German. Equipped with this system, we evaluated CoMiC on various subsets of the German CREG and demonstrated the robustness of the approach through consistently high accuracy figures in the $> 80\%$ range. Finally, we briefly described three works which have successfully applied CoMiC to other computational linguistic tasks, such as QA and plagiarism detection.

Part III.

Focus: Internal and External Relevance

6. Focus Annotation of Reading Comprehension Data

In this chapter, we discuss the manual annotation of focus. We first discuss where in the data empirical evidence on focus might be obtained (section 6.1), and highlight the main problems with relying on surface features in the operational definition of focus. Then we launch into a description of our own focus annotation scheme (section 6.2), pointing out interesting problems to solve before detailing how our annotation guidelines tackle them and what the technical considerations for the actual annotation procedure were. This section draws on Ziai & Meurers (2014).

Section 6.3 describes the results of the annotation process by our trained expert annotators. We discuss what the options for measuring inter-annotator agreement for focus are in the first place before reporting the agreement between our annotators and pointing out sources of disagreement. This section is based on Ziai & Meurers (2014) and De Kuthy, Ziai & Meurers (2016a).

In section 6.4, we turn to an implementation of focus annotation for non-experts, through crowd-sourcing of judgments on a major online platform. We first state the problem of formulating focus annotation for non-linguists, then outline our solution and how different crowd judgments on the same sentence are combined into one annotation. Finally, we discuss results in terms of agreement between the crowd and experts and also show how the quality of crowd annotation may be predicted independently. This section draws on work described in De Kuthy, Ziai & Meurers (2016b).

6.1. Sources of Evidence for Focus Annotation

In this section, we will discuss how focus may be determined in utterances and what reliable sources of evidence in authentic data are. The goal will be to find an operationalizable definition of focus. After discussing several surface criteria, we point out a major problem of defining focus in terms of surface features. Finally, we sketch our solution to this problem, which is then fleshed out in section 6.2.

6.1.1. Surface Criteria and Why They Are Not Sufficient

As we outlined in chapter 3, focus is signalled on various linguistic levels across different languages. One prominent area where this happens in Western languages is prosody, especially the **intonation and stress** assigned to parts of the utterance. Consider the question-answer pair in example (30) for illustration.

(30) Who came to the party?
⟦JOHN⟧$_F$came to the party.

Here, in the answer to *Who came to the party?*, *John* is both the focus and the bearer of the pitch accent. It would therefore be possible to identify focus by identifying pitch accents, without the need to resort to other features, if one has access to reliable prosody information.

However, this is not the case for all languages,and certainly not for written language, where focus is expressed through other means. Consider the example in (31).

(31) Who came to the party?
It was ⟦John⟧$_F$who came to the party.

In (31), a cleft construction is used to signal that the focus is *John*. If this was spoken language, *John* would additionally bear a pitch accent, but even without it there is clear syntactic evidence for focus placement. If all foci were so clearly marked, it would be possible to define focus in terms of **syntactic markedness**.

Another type of evidence exists in the lexical domain in the form of so-called **focus-sensitive particles**, such as *only*. They are claimed to signal that the focus of the sentence follows immediately to their right, as in (32):

(32) Who came to the party?
Only ⟦John⟧$_F$came to the party.

All of the above sources of evidence have in common that they are possible indicators of focus, but not necessary indicators, as focus exists without their presence too. This is demonstrated by the "vanilla version" of the example we have been using in this section, shown in (33).

(33) Who came to the party?
$[\![\text{John}]\!]_F$came to the party.

We conclude that defining focus in terms of any single surface feature is not possible. In fact, this seems intuitive because there would be no need for the term focus if it always corresponded to a surface phenomenon. Thus, if one were to base focus annotation guidelines on a definition of focus in terms of surface criteria, good agreement in annotation would only mean that the surface criteria can be robustly identified, not that focus can in fact be annotated reliably. This problem for the annotation of theoretical notions in linguistics has been pointed out more generally by Riezler (2014).

In the next section, we are going to show how focus can be defined through meaning-based criteria that will then be operationalized in a concrete annotation scheme.

6.1.2. Using Meaning-based Criteria

In the previous section, we have seen that annotating focus in terms of surface criteria will be problematic. This is due to the fact that while focus may be reliably signalled through means such as pitch accents and syntactic markedness, such indicators need not be present for a sentence to contain focused material.

We therefore adopt the view that the focus-background distinction is ultimately a **partitioning of the meaning of an expression** (cf., e.g. Krifka 2001), not a feature of other linguistic layers. This partitioning is imposed by contextual requirements, in particular the current **Question Under Discussion (QUD)** (Roberts, 2012). In other words, focused material addresses an information requirement currently present in the discourse. For illustration, consider (34) which is example (9) from Krifka (2001), where the information requirement *PERSON* of the question is made explicit in the semantics:

(34) Who did Mary see? $\exists x[saw(x)(M)], PERSON$
Mary saw John. $saw(J)(M)$

For an annotation approach, it thus follows that in order to take this view seriously, the annotation scheme needs to explicitly take questions and their information requirements into account. Applied to our example from the previous section, this means that the scheme needs to capture that a question such as *Who came to the party?* is looking for a semantic object of type *PERSON* in the answer.

However, developing such an annotation scheme imposes the practical requirement that questions need to be explicitly present in the data to be annotated. Fortunately, this is the case with reading comprehension data where explicit questions are asked about a text, making this corpus type an ideal testbed for the development of a meaning-based focus annotation scheme.

6.2. Developing a Focus Annotation Scheme

In this section, we introduce the annotation scheme we developed. We first discuss what representations focus can be annotated on before going into a description of the actual annotation scheme, and finally a summary in the form of the general workflow that annotators were required to follow.

6.2.1. What to Annotate?

For a meaning-based notion such as focus, the first question that naturally arises in annotation is on what linguistic level it can be annotated: directly on the word level, based on some syntactic annotation, or even based on a semantic representation?

Given that we established focus as a partitioning of meanings, the logical step would be to annotate semantic representations of utterances. In doing so, we would be able to abstract over surface variations in meaning that result from lexical and syntactic choices, while cleanly separating the focused meaning from the background meaning of the utterance.

However, semantic representations are still hard to identify automatically in a robust manner. Semantic parsing is an active field of research in its own right (cf., e.g., Pradhan et al. 2004), and cannot be considered a solved problem, as e.g. the recent ACL 2014 workshop on this topic demonstrates (Artzi et al., 2014).

Manual annotation of semantic representations, on the other hand, would be extremely time-consuming and well beyond the scope of this thesis.

A similar argument can be made for automatic syntactic annotation, although it is less prone to errors. But the more serious problem here is that syntactic annotation does not provide a direct correspondence to meaning: what is a unit in syntax does not necessarily constitute a unit of meaning that should be annotated as one.

For these reasons, we decided to annotate focus directly on the word level. This approach has the disadvantage that annotators need to handle lexical and syntactic issues as well, but does not presume any preprocessing and is fine-grained enough to support distinctions in meaning.

6.2.2. The Annotation Scheme

An important characteristic of our annotation scheme is that it is applied incrementally: annotators first look at the surface question form, then determine the set of alternatives (Krifka, 2007), and finally they mark instances of the alternative set in answers. The rich task context of reading comprehension data with its explicit questions allows us to circumvent the problem of guessing an implicit QUD, except in the cases where students answer a different question (which we account for separately, see below). In the following, we present the three types of categories our scheme is built on.

Question Form

Question Form is used to mark the surface form of a question, where we distinguish *wh*-questions, polarity questions, alternative questions, imperatives and noun phrase questions. In themselves, question forms do not encode any semantics, but merely act as an explicit marker of the surface question form. Table 6.1 gives an overview and examples of this dimension.

Focus

Focus is used to mark the focused words or phrases in an answer. We do not distinguish between contrastive and new information focus, as it is not relevant for assessing an answer whether the alternatives are explicitly mentioned before

Category	Example	Translation
`WhPhrase`	'Warum hatte Schorlemmer zu Beginn Angst?'	'Why was Schorlemmer afraid in the beginning?'
`YesNo`	'Muss man deutscher Staatsbürger sein?'	'Does one have to be a German citizen?'
`Alternative`	'Ist er für oder gegen das EU-Gesetz?'	'Is he for or against the EU law?'
`Imperative`	'Begründen Sie diesen anderen Spitznamen.'	'Give reasons for this other nickname.'
`NounPhrase`	'Wohnort?'	'Place of residence?'

Table 6.1.: Question Forms in the annotation scheme

and contrasted against, or whether they are available in the context through other means (in our case this will usually be the reading text). We also allow for the encoding of multiple foci and they in fact do occur in the data.

The Semantic Contribution Test The starting point of our focus annotation is Krifka (2007)'s understanding of focus as the part of an utterance that indicates the presence of alternatives relevant to the interpretation (see Definition 3). We operationalize this by testing whether a particular part of the utterance is needed to distinguish between alternatives evoked by the QUD. Recall that in chapter 3 we also established a top-down view of focus, where focus is seen as the answer to a question (see Definition 4), a view that is supported by the semantic phenomenon of Question-Answer Congruence (QAC) (Stechow, 1991).

Concretely, we train annotators to perform substitution tests in which they compare two potential extents of the focus to identify whether the difference in the extent of the focus also selects a different valid alternative in the sense of discriminating between alternatives evoked by the QUD. The test is phrased for annotators as shown in Definition 5.

Definition 5. *If the focus status of some utterance part (e.g. a preposition) is unclear, test for its semantic contribution: Does the meaning change when it is left out or changed?* (Semantic Contribution Test)

For illustration, consider the example in (35), where the question asks for a location.

(35) Q: Where does Heike live?

A: She lives ⟦in Berlin⟧$_F$.

Here "in" needs to be part of the focus because exchanging it for another word with the same POS changes the meaning of the phrase in a way picking another alternative, as in "She lives *near* Berlin". Consider the same answer to a slightly different question in (36). Here the set of alternatives is more constrained (the set of cities vs. the set of locations) and hence "in" is not focused.

(36) Q: In what city does Heike live?
A: She lives in ⟦Berlin⟧$_F$.

To illustrate that these examples also work for German, consider example (37), where a similar meaning and corresponding distinction between alternatives is expressed.

(37) Q: In welche Stadt zieht Sandra?
into which city moves Sandra
'To what city is Sandra moving?'
A: Sie zieht nach ⟦München⟧$_F$.
she moves to Munich
'She is moving to Munich.'

The test in Definition 5 does not only work for questions that ask for locations, however. In (38) from the CREG corpus, the question asks for a reason why the houses are broken.

(38) Q: Warum waren die Häuser kaputt?
why were the houses broken
A: Die Häuser waren kaputt wegen ⟦des Krieges⟧$_F$.
the houses were broken due to the war

The relevant reason mentioned in the answer is *des Krieges* ('the war'). Reasons are most often expressed as either full sentences or as embedded clauses introduced with *weil* ('because'), so this example is an interesting atypical example where a reason is expressed in a prepositional phrase. Again, the annotator needs to determine whether the preposition *wegen* ('due to') is part of the focus: since *wegen* is only present for syntactic well-formedness and does not contribute to a distinction between alternatives, it is unfocused. *des* ('the') on the other hand is part of the focus because of the definiteness it introduces, distinguishing between 'some war' and 'a specific war'.

General Criteria Besides the test described above, there are also a number of more general criteria we defined to guide focus annotation. The most important one of these is the relation to givenness: even though we do not assume givenness to be a part of the definition of focus, we recognize its correlation with background material and instruct annotators to avoid marking given material, unless it is needed to distinguish between alternatives.

Another issue that frequently comes up in our data is focus in coordination structures, where the issue is to decide whether the whole coordination is one focus or whether multiple foci are coordinated. We opted again for a semantic solution: annotators should mark separate foci depending on whether they constitute separate valid alternatives. An example is shown in (39), where the foci *Reisen* ('traveling') and *Bücher* ('books') represent separate valid alternatives.

(39) Q: Was mag Tina?
what likes Tina
'What does Tina like?'
A: Sie mag ⟦Reisen⟧$_F$ und ⟦Bücher⟧$_F$.
she likes traveling and books

There is also a set of well-formedness constraints, which includes the fact that focus is a sentence-based notion and never crosses sentence boundaries, and that, contrary to e.g. pitch accents, it does not apply to sub-lexical units such as syllables. Also, we generally ignore punctuation at focus boundaries.

Relation to the QUD Each sentence is assumed to include at least one focus. If it does not answer the explicit question, it must be annotated with a different QUD. So in addition to marking focus, we annotate the relation between the explicitly given question and the QUD actually answered by a given response. In the most straightforward case, the QUD is identical to the explicit question given, which in the annotation scheme is encoded as **question answered**. In cases where the QUD differs from the explicitly given question, we distinguish three cases: In the cases related to the implicit moves discussed in Büring (2003,

p. 525) exemplified by (40), the QUD answered can be a sub-question of the explicit question (indicated below in slanted font after the explicit question), which we encode as **question narrowed down**. When it addresses a more general QUD, as in (41), the response is annotated as **question generalized**.

(40) Q: What did the pop stars wear?
QUD: *What did the female pop stars wear?*
A: The female pop stars wore caftans.

(41) Q: Would you like a Coke or a Sprite?
QUD: *What would you like to drink?*
A: I'd like a beer.

An authentic example from the CREG corpus for a question narrowed down is (42), where the answer addresses two sub-questions of the explicit question, likely inspired by the information available in the reading text.

(42) Q: War dieses Event legal oder illegal?
was this event legal or illegal
QUD$_1$: *War dieses Event legal oder illegal in West-Berlin?*
QUD$_2$: *War dieses Event legal oder illegal in Ost-Berlin?*
SA: ⟦Legal⟧$_F$ in West-Berlin, ⟦illegal⟧$_F$ in Ost-Berlin.
legal in west Berlin, illegal in east Berlin

Finally, we also mark complete failures of QAC with **question ignored**. In all cases where the QUD being answered differs from the question explicitly given, the annotator is required to specify the QUD apparently being answered. Naturally, in the latter case of ignored questions the assumed QUD is somewhat more arbitrary in nature, since no relation to the explicit question could be determined.

Answer Type

Answer Type expresses the semantic category of the focus in relation to the question form. It further describes the nature of the question-answer congruence by specifying the semantic class of the set of alternatives. The answer types discussed in the computational linguistic literature generally are specific to particular content domains, so that we developed our own taxonomy. Examples include `Time/Date`, `Location`, `Entity`, and `Reason`. In addition to semantically restricting the focus to a specific type, answer types can also provide syntactic cues restricting focus marking. For example, an `Entity` will typically be encoded as a nominal expression. For annotation, the advantage of answer types is that they force annotators to make an explicit commitment to the semantic nature of the focus they are annotating, potentially leading to higher consistency and reliability of annotation. On the conceptual side, the semantic restriction encoded in the answer type bears an interesting resemblance to what in a Structured Meaning approach to focus (Krifka, 1992) is referred to as *restriction of the question* (Krifka, 2001, p. 3).

6.2.3. Annotation Workflow

Summing up, the basic annotation procedure is as follows:

1. Examine the surface form of the question and label it with the appropriate **Question Form** tag.
2. Determine the set of alternatives opened up by the question. For example, *Wo wurde Mozart geboren?* opens up the set of places where Mozart could have been born, whereas *Was hat Herbert getan?* opens up the set of activities Herbert might have done.
3. For each answer, mark the part of it as **Focus** that either explicitly chooses one alternative or at least narrows down the set of alternatives. For example, a question such as *Wer kam gestern zur Party?* might be answered by an exhaustive list of individuals, but could also partly be answered by something like *Alle die eingeladen waren.*, thus describing a set of individuals through a common property.
4. Finally, connect each instance of **Focus** that addresses an explicit question back to the corresponding **Question Form** by using the appropriate **Answer Type**. The **Answer Type** needs to further describe the set of alternatives that the respective **Focus** is part of, so it must be possible to read the relation as "[Focus] is a [Answer Type] answering [Question Form]...". For example, "*Salzburg* is a `Location` answering `WhPhrase` *Wo?*".

Category	Description	Example (translated)
`Time_Date`	time/date expression, usually incl. preposition	The movie starts *at 5:50*
`Living_Being`	individual, animal or plant	*The father of the child* padded through the dark outskirts.
`Thing`	concrete object which is not alive	For the Spaniards *toilet and stove* are more important than the internet.
`Abstract_Entity`	entity that is not concrete	The applicant needs *a completed vocational training as a cook*.
`Report`	reported incident or statement	The speaker says *"We ask all youths to have their passports ready."*
`Reason`	reason or cause for a statement	The maintenance of a raised garden bed is easier *because one does not need to stoop*.
`Location`	place or relative location	She is from *Berlin*.
`Action`	activity or happening.	In the vegetable garden one needs to *hoe and water*.
`Property`	attribute of something	Reputation and money are *important* for Til.
`Yes_No`	polar answer, including whole statement if not elliptic	*The mermaid does not marry the prince*.
`Manner`	way in which something is done	The word is used *ironically* in this story.
`Quantity/Duration`	countable amount of something	The company seeks *75* employees.
`State`	state something is in, or result of some action	If he works hard now, *he won't have to work in the future*.

Table 6.2.: Answer Types with examples

6.2.4. Annotation Tool

Having described the annotation scheme and its theoretical motivation in detail, we now turn to how the annotation scheme was applied in practice using annotation software. Our desiderata for such a focus annotation tool were the following:

1. **Flexible marking of answer words in a discourse setting**: the tool needs to allow annotators to mark everything the annotation scheme allows, and ideally no more than that. It also needs to display all the context needed to perform the task.

2. **Customizable annotation scheme**: instead of hard-coded notions that we have to then divert from their intended use, the tool should provide an intuitive means of specifying annotation categories and constraints.

3. **Ease of use**: the tool should not require annotators to go through a complicated installation mechanism and it should be reasonably self-explanatory if the task is conceptually clear.

4. **Support for multiple annotators**: evaluation of the scheme in terms of inter-annotator agreement (see next section) requires multiple annotation versions of the same data, so the tool needs to readily support this.

5. **Straightforward use of annotation results**: in order to perform an evaluation or to use the resulting annotation in another task (e.g. SAA), the results need to be easy to read and process in other programs.

Several tools have been used for IS annotation, and they fulfill the above criteria to varying degrees. For example, Ritz et al. (2008) used EXMARaLDA (Schmidt, 2004) which is in essence a transcription tool meant for spoken language data. However, its support for multiple layers of annotation made it somewhat popular for other types of corpora. The main drawback of EXMARALDA is that there is no way to make the annotation scheme explicit, i.e., there is no support in guiding annotators through the annotation process. Also, the resulting annotation is stored in a somewhat cumbersome XML format, making further processing more difficult.

Another approach is the Nite XML Toolkit used for the Switchboard corpus (Calhoun et al., 2010). It features a powerful data model and query language supporting multiple layers of linguistic annotation, as needed by the dialog data used in the Switchboard project. However, the complexity of the toolkit in connection with the correspondingly complex file format makes it rather difficult to work with, so we did not choose this toolkit for our annotation effort.

In other discourse annotation tasks such as co-reference resolution, MMAX2 (Müller & Strube, 2006) is frequently used. It displays a whole document and allows for specification of custom annotation schemes via stylesheets, from which the user interface is then generated. While these features sound very promising, customizing MMAX2 in practice is non-trivial.

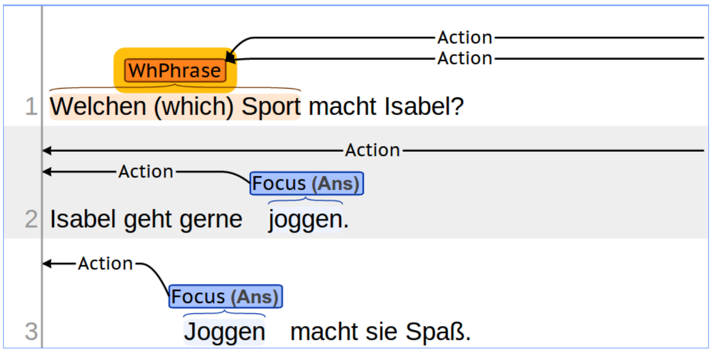
All of the tools discussed above also have in common that they need to be installed locally, and do not come as a web-based version. Besides making things more complicated for the annotator, this has the disadvantage that the data under annotation also resides on the annotator's machine instead of a central server where the researcher supervising the annotation has direct control over the data, and existing infrastructure can be used to prevent data loss.

In contrast, the brat annotation tool[1], is a more modern, web-based solution. brat offers an intuitive way of customizing the annotation scheme via configuration files, and comes with a straightforward user interface. Annotations are stored in a simple but functional text format, separate from the source data. Additionally, brat allows for relations between span-based annotations, which we used for implementing our Answer Type category: an Answer Type is a relation between a Focus and a Question Form. brat also readily supports multiple annotators through the creation of multiple user accounts. For these reasons, we chose it as the basis for our annotation efforts.

Figure 6.1 shows a *brat* screen shot with an example including a `WhPhrase` Question Form (line 1) and two answers, a target answer (TA, line 2) and a student answer (SA, line 3), containing a word selected as focus with Answer Type `Action`. We display target answers together with student answers in order to provide the annotator with some contextual knowledge that facilitates the interpretation of the often ungrammatical student answers. The screenshot also demonstrates how different colors can be used in brat for different annotation categories.

Equipped with a concrete annotation scheme and a tool to carry out the annotation experiment, we can now proceed to discussing annotation results in the next section.

[1] `http://brat.nlplab.org`



1 Welchen (which) Sport macht Isabel?

2 Isabel geht gerne joggen.

3 Joggen macht sie Spaß.

Q: ‘*Which sport* does Isabel do?’

TA: ‘She likes to go ⟦jogging⟧$_F$.’

SA: ‘⟦Jogging⟧$_F$ is fun for her.’

Figure 6.1.: Brat annotation example

6.3. Expert Annotation: Empirically Validating the Annotation Scheme

In this section, we describe the manual focus annotation experiments that were carried out as part of this thesis research. We first outline the procedure and the iterative nature of the annotation effort, before asking ourselves how focus annotation can be evaluated and describing the results. Finally, we discuss some important sources of disagreement between annotators through example cases.

6.3.1. Annotation Setup and Training

The annotation effort was carried out in two phases. Each phase used different CREG sub-corpora: in the first phase, the annotators worked on CREG-1032 and in the second phase, the bigger CREG-2155 was used (see section 4.2.3 for details on these subsets). In both phases, the annotation was performed by two graduate research assistants in linguistics using brat directly on the token level. Each annotator was given a separate directory containing identical source files to annotate.

In order to sharpen distinctions and refine the annotation scheme to its current state, the first phase included a piloting phase: we drew a random sample of 100 questions, target answers and student answers from each subcorpus (KU and OSU) of CREG and trained our two annotators on them. During this piloting process, we met with the annotators to discuss difficult cases and decide how the scheme would accommodate them.

For the second phase, we had to use different annotators, which meant they again needed to be trained. They underwent the same training that was required for the annotators of the first phase and we again met with them to discuss clarify difficult distinctions in case they were not clear from the category descriptions in the scheme.

After the second phase was completed, we used a third annotator as judge in order to merge the two annotation versions from both phases into one gold standard. In cases of conflict, whenever a focus annotation in line with the guidelines was provided by one of the annotators, the judge picked that annotation, resorting to a different annotation only when both versions were incorrect.

6.3.2. How to Measure Success?

Having performed the manual annotation, the question arises how to compare and calculate agreement of spans of tokens in focus annotation. In other span-based CL problems, such as Co-reference Resolution or Named Entity Recognition (NER), the set of markable linguistic units is usually compared using some combination of Precision and Recall. In NER, for example, both Precision and Recall are defined in terms of exact matches between annotated units, which means every missed or spurious token is an error (cf., e.g., Tjong Kim Sang & De Meulder 2003).

An important characteristic of such tasks is that the set of markables explicitly constrained beforehand, usually in a syntactic manner: both co-reference and NER are defined for **noun phrases** only, which means other syntactic units need not be part of the evaluation.

IS notions are different in that they structure the meaning of utterances according to contextual requirements. Applied to our task, this means that in principle **any word** can be focused and no word classes can be excluded a priori. The consequence for evaluation schemes is that each word must be treated as a markable for which the annotator needs to make a decision.

Given that we compare annotations on the word level, we decided to follow standard evaluation procedures in calculating percentage agreement and Cohen's Kappa (Cohen, 1960), which we already introduced in chapter 3 (equation 3.1).

In principle it would of course be possible to use other agreement measures. Artstein & Poesio (2008) discuss Fleiss' κ and multi-κ (Fleiss, 1971) along with Krippendorff's α coefficient (Krippendorff, 1980), which allows for different weights for disagreements between annotators. Multi-coder versions of κ are however not necessary in our case since we only use two annotators at any time. The possibility of weighting disagreements seems more interesting, but our basic annotation problem distinguishes only two classes (*focus* and *background*), so it seems unnecessary to complicate the agreement measure here.

6.3.3. Quantitative Results of Phase One

Table 6.3 summarizes the agreement results for the first phase of annotation. We based this phase of annotation on the CREG-1032 subset (see section 4.2.3), since it has already been used for evaluating our own (Meurers, Ziai, Ott & Kopp, 2011b) and several other SAA approaches (Hahn & Meurers, 2012; Horbach et al., 2013; Pado & Kiefer, 2015) and thus provides a good testbed for integrating focus into SAA later (see chapter 8).

Type of distinction	Type of answers	# tokens	%	κ
Binary (focus/background)	Student	10557	85.6	.69
	Target	2013	91.1	.82
	Both	12570	86.4	.71
Detailed (13 Answer Types + background)	Student	8748	77.5	.70
	Target	1978	82.4	.76
	Both	10726	78.4	.71

Table 6.3.: Agreement on student and target answers in CREG-1032

For both student and target answers, we report the granularity of the distinction being made (focus/background vs. all answer types), the number of tokens the distinction applies to, and finally percentage and Kappa agreement.

The results show that all numbers are in the area of substantial agreement ($\kappa > .6$). This is a noticeable improvement over the results obtained by Ritz et al. (2008), who report $\kappa = .51$ on tokens in questionnaire data, and it is on a par with the results reported by Calhoun et al. (2010), even though the latter is somewhat artificially constrained to certain word classes, as we have described in our annotation review in section 3.3.2.

Annotation was easier on the more well-formed target answers than on the often ungrammatical student answers. Moving from the binary focus/background distinction to the one involving all Answer Types, we still obtain relatively good agreement. This indicates that the semantic characterization of foci via Answer Types works quite well, with the gap between student and target answers being even more apparent here.

In order to assess the effect of answer length, we also computed macro-average versions of percentage agreement and κ for the binary focus distinction, following Ott et al. (2012, p. 55) but averaging over answers. We obtained 87.5% and $\kappa = .73$ for student answers, and 93.0% and $\kappa = .86$ for target answers. A few longer answers which are harder to annotate thus noticeably affected the agreement results of Table 6.3 negatively.

6.3.4. Incremental Changes to the Annotation Guidelines

After the first phase of annotation, we met with the annotators and discussed problematic cases in order to improve the guidelines. One of the issues that came up was the aforementioned rule about foci in coordination structures. While the general principle of marking separate foci in case of multiple alternatives was clear, there were certain borderline cases where the relation between possible foci was unclear. An example is shown in (43).

(43) Q: Was aßen sie zum Kaffee?
what ate the to the coffee
'What did they have with the coffee?

A: $[\![$Kekse und Knabbereien$]\!]_F$.
cookies and snacks

The problem here is that while *Kekse* ('cookies') and *Knabbereien* ('snacks') seem to be two separate valid alternatives, they are semantically not disjunctive:

Knabbereien is arguably a hyperonym of *Kekse*. We added a clause in the guidelines that in cases like this, where the semantic relationship between alternatives involves hierarchy, annotators should only mark one focus.

Another problem concerned the distinction of the Answer Types `State` and `Action`, which was sometimes hard for annotators to do because both have similar syntactic correlates, namely clauses. We resolved this by examining the lexical aspect of the verb in the answer: if it is dynamic (e.g. 'to run'), i.e., describing an event, the resulting type should be `Action`, whereas if it is static (e.g. 'to own'), the type should be `State`.

We also simplified the marking of determiners at focus boundaries: since we could not find an instance where the determiner does not contribute to the meaning of a noun phrase at all, we opted to always include determiners as part of the focus.

6.3.5. Quantitative Results of Phase Two

The second phase of annotation was performed for two reasons. First, we wanted to validate the annotation approach by applying it to a broader range of language material, and see whether it was general enough. Second, we needed to obtain more high-quality training data for the automatic focus detection approach we describe in chapter 7.

The generalization aspect is also supported by the fact that we used two new annotators for the second phase, who were trained the same way as the first. As far as the new language material is concerned, ideally we would have extended manual focus annotation to the whole CREG-5K corpus (see section 4.2.3), which has over 5,000 student answers. However, since we did not have the resources to annotate the entire CREG-5K, we sampled a subset of it in the following way: we first removed all questions and their answers that are already present in CREG-1032[2], since they were part of the first phase of annotation. From the remainder, we randomly sampled approximately 2,000 answers, balancing them over questions so that all questions are proportionally represented (stratified random sample). The resulting corpus is called CREG-2155, due to its 2,155 student answers.

[2] We did not remove questions based on entire reading texts, however, so there is some reading text overlap between CREG-1032 and CREG-5K.

Type of distinction	Type of answers	# tokens	%	κ
Binary (focus/background)	Student	22755	83.4	.64
	Target	10952	90.8	.79
	Both	33714	85.8	.69
Detailed (13 Answer Types + background)	Student	18468	76.8	.70
	Target	10811	75.6	.70
	Both	29283	76.4	.70

Table 6.4.: Agreement on student and target answers in CREG-2155

The agreement results of the second phase are summarized in Table 6.4. They are somewhat lower than the ones for CREG-1032 in Table 6.3, but still one the same high level. While individual annotator performance may play a role here, we strongly suspect the lower results to be a consequence of the higher task complexity in CREG-5K, and hence CREG-2155: as we briefly hinted at in section 4.2.3 at the end of chapter 4, the reading texts of CREG-2155 are on average approximately three times as long as the ones of CREG-1032, and thus the answers can be expected to be of a much more heterogeneous nature, which complicates the focus annotation task. This suspicion is also supported by the fact that CoMiC results on CREG-2155 are somewhat lower than on CREG-1032, showing that automatic content evaluation is harder here also.

Finally, in Table 6.5 we summarize the agreement results for both phases of annotation. In total, 4,177 answers with 46,284 tokens were annotated, of which 3,187 are student answers and 767 are target answers. The overall agreement resulting from treating both annotation phases as one data set is 86.0% with $\kappa = .7$. We call the overall expert-annotated corpus CREG-ExpertFocus, which in its adjudicated gold standard version forms the basis for focus detection in chapter 7.

	# answers	# tokens	% agreement	κ
CREG-1032	1255	12570	86.4%	0.71
CREG-2155	2922	33714	85.8%	0.69
CREG-ExpertFocus	4177	46284	86.0%	0.70

Table 6.5.: Overall inter-annotator agreement for focus/background

6.3.6. Sources of Disagreement

To explore the nature of the disagreements in both annotation phases in a qualitative manner, we showcase three characteristic issues here based on examples from the corpus. Consider the following case where the annotators disagreed on the annotation of a student answer:

(44) Q: Warum nennt der Autor Hamburg das "Tor zur Welt der
why calls the author Hamburg the gate to the world of the
Wissenschaft"?
science
'Why does the author call Hamburg the "gate to the world of science"?'
SA$_{A1}$: ⟦Hamburg hat viel renommierte Universitäten⟧$_F$
SA$_{A2}$: Hamburg hat ⟦viel renommierte Universitäten⟧$_F$
'Hamburg has many renowned universities'

Whereas annotator 1 (A1) marks the whole answer on the grounds that the focus is of Answer Type `Reason` and needs to include the whole proposition, annotator 2 (A2) excludes material given in the question. Both can in theory be justified, but annotator 1 is closer to our guidelines here, taking into account that *Hamburg* indeed discriminates between alternatives (one could give reasons that do not include *Hamburg*) and thus needs to be part of the focus.

The second example illustrates the issue of deciding where the boundary of a focus is:

(45) Q: Wofür ist der Aufsichtsrat verantwortlich?
for what is the supervisory board responsible
'What is the supervisory board responsible for?'
SA$_{A1}$: Der Aufsichtsrat ist für ⟦die Bestellung⟧$_F$ verantwortlich.
SA$_{A2}$: Der Aufsichtsrat ist ⟦für die Bestellung⟧$_F$ verantwortlich.
'The supervisory board is responsible for the appointment.'

Annotator 1 correctly excluded *für* ('for') from the focus, only marking *die Bestellung* ('the appointment') given that *für* is only needed for reasons of well-formedness. Annotator 2 apparently thought that *für* makes a semantic difference here, but it is hard to construct a grammatical example with a different preposition that changes the meaning of the focused expression.

Finally, in (46) we have an example that illustrates a combination of both problems:

(46) Q: Wer tappte durch die dunkle Vorstadt?
who padded through the dark outskirts
SA$_{A1}$: Der Mann, der ist ⟦der Vater von das Kind⟧$_F$.
SA$_{A2}$: ⟦Der Mann, der ist der Vater von das Kind⟧$_F$.
'The man who is the father of the child.'

The question asks for an individual. Annotator 1 marks the part of the relative clause that most closely describes that individual, which however disregards the fact that the relevant alternative is introduced by the main clause including the restrictive relative clause. Annotator 2 correctly marks the whole sentence, recognizing that *Der Mann* ('the man') already selects among alternatives, and the restrictive relative clause further narrows it down.

6.3.7. Open Problems

Despite the incremental improvements and the overall success of the annotation effort, several problem areas remain, some of which we will discuss briefly in this section.

Focus in Answers to *why*-questions

One problem concerns the Answer Type `Reason`, generally used for foci which are answers to *why*-questions. While other alternative sets such as individuals, places and times are quite clearly defined, reasons are not: whether something is a valid reason or not is determined by what we know about the situation and the world in general, and the inferences we can draw based on that knowledge. Knowledge is not part of a definition of focus (or any language phenomenon), so one could argue that focus can not be expected to provide a solution for this problem.

However, since focus is actually marked in answers to *why*-questions, as evidenced by pitch accents and other surface markers, the answers do apparently indicate some form of alternative. One possible approach would thus be to find a sub-question to the explicit *why*-question which more effectively constrains the alternative set and to which the given answer is congruent.

Focus and Syntactic Omission

Another issue emerges from the annotation setup we use, which is to annotate a partitioning of meaning (focus) on the basis of a surface representation (words). As long as there is a direct correspondence between surface forms and units of meaning, the surface forms can be used as proxies for annotation. However, certain syntactic phenomena, such as ellipsis, allow for omission of language material which we need for annotating distinct foci. This is especially apparent in coordination structures, which tend to occur frequently in our data.

We do not see a direct solution to this problem in the current setup, but rather suggest that this problem might be solved by annotating focus on a deeper linguistic level, perhaps akin to the tecto-grammatical structures found in the Prague Dependency Treebank (see section 3.3.2).

Non-wellformed Language

A related problem is the frequent occurrence of non-wellformed language in the learner data that we use. In most cases, the errors learners make do not preclude the interpretation of the answer, but they can still pose problems for focus annotation, as shown by the significantly lower agreement figures on student answers in comparison to the well-formed target answers.

If one sees non-wellformedness as a dimension of ill-formed variation (Meurers & Dickinson, 2017, sec. 2.2), the issue connects the one we described before, where the problem was that certain well-formed syntactic constructions make focus annotation on surface representations difficult. However, while well-formed variation has been extensively studied in syntactic frameworks, it is unclear how ill-formed variation may be normalized.

One possible solution would be to first construct a so-called minimal target hypothesis (Lüdeling et al., 2005), which includes the minimum number of edits necessary to make a sentence grammatical. Given such a target hypothesis, focus annotation could then proceed as usual.

6.4. Crowd-sourcing Annotation: External Grounding and Scaling Up

As we mentioned at the beginning of section 6.3.5, manual focus annotation by experts is a very time-consuming task, since it includes question analysis, determining the alternative set and testing for the extent of the focus.

We thus wanted to investigate whether a meaning-based notion such as focus can also be annotated with some success by untrained native speakers in a crowd-sourcing setup, as has been done successfully for other annotation tasks (cf., e.g., Snow et al. 2008). This is interesting for two reasons: first, it would be a means of obtaining more annotated data in a much faster way, since annotation on crowd-sourcing platforms can be done by hundreds of so-called 'workers' at a time. Second, it provides an external grounding of the notion of focus which is not dependent on a small number of linguistic experts' understanding of it, a problem in linguistic annotation which has been pointed out by Riezler (2014).

In this section, we thus present a crowd-sourcing experiment on focus annotation. We first explain the setup of the experiment before describing annotation results both quantitatively and qualitatively. Finally, we delve into the question of how the quality of crowd-sourced focus annotation may be predicted without comparing it to an external gold standard such as our expert annotation, and define a measure that accomplishes this to a certain degree.

6.4.1. Setup of the crowd-sourcing experiment

To study non-expert focus annotation, we implemented a crowd-sourcing task using the crowd-sourcing platform CrowdFlower[3] 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 like Amazon Mechanical Turk do not provide as transparently, and it has a built-in quality control mechanism ensuring that workers maintain a certain level of accuracy on interspersed test items throughout the entire job.

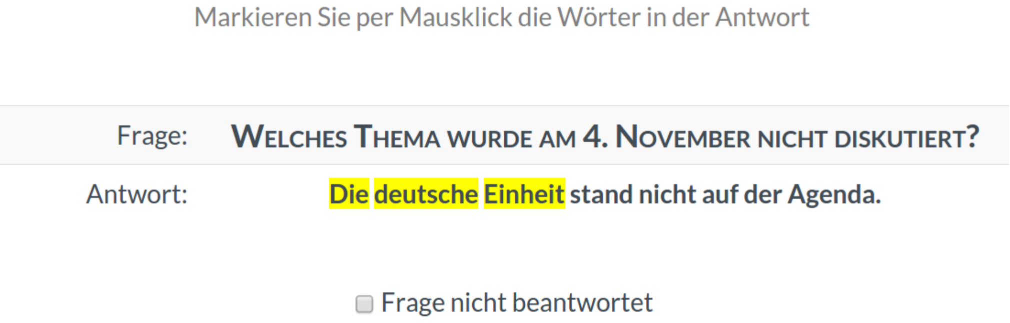
As data for our crowd-sourcing experiment, we used 5,597 question-answer pairs from the CREG-5K corpus (see section 4.2.3) and 100 manually constructed

[3] http://www.crowdflower.com/

test question-answer pairs. The task of the crowd workers was to mark those words in an answer sentence that “contain the information asked for in the question”. Workers were shown five question-answer pairs at a time. One of those five was from our set of hand-crafted test question-answer pairs. The workers were paid $0.02 per annotated sentence.

Since CREG-5K consists of reading comprehension questions and answers provided by learners of German, there are cases where a student response does not answer a given question at all, for example, when the learner misunderstood the question. In the gold standard annotation described in section 6.3, the annotators had the option to mark such cases as “question ignored”. Since we also wanted to provide the crowd workers with this option, we included a checkbox “Frage nicht beantwortet” (“*question not answered*”). When this option is selected, no word in the answer sentence can be marked as focus.

Figure 6.2 shows an example CrowdFlower task with the marked words in yellow. These marked words are the ones that we counted as focus. The English translation shown below was not part of the CrowdFlower task.



Markieren Sie per Mausklick die Wörter in der Antwort

Frage: WELCHES THEMA WURDE AM 4. NOVEMBER NICHT DISKUTIERT?

Antwort: Die deutsche Einheit stand nicht auf der Agenda.

□ Frage nicht beantwortet

Q: ‘Which topic was not discussed on November 4th?’
A: ‘⟦The German unification⟧$_F$ was not on the agenda.’

Figure 6.2.: Example CrowdFlower annotation task

We collected 11 focus annotations per answer sentence and crowd workers had to maintain an accuracy of 60% on the test question-answer pairs. Altogether we collected 62,247 annotated sentences.

6.4.2. Evaluation

To evaluate the quality of our crowd focus annotation, we wanted to find out how the annotations produced by the crowd workers compare to the gold standard expert annotation described in section 6.3. We therefore chose to calculate all possibilities of combining one through eleven workers into one "virtual" annotator using majority voting on individual word judgments. Ties in voting are resolved by random assignment. The procedure is similar to the approach described by Snow et al. (2008). We did not employ any bias correction or other types of weighting schemes, as discussed, e.g., by Qing et al. (2014), but plan to do so in future research.

In measuring agreement between crowd workers and the expert gold-standard on the word level, for the following reasons we opted for percentage agreement instead of Kappa or other measures that include a notion of expected agreement: *i)* Kappa assumes the annotators to be the same across all instances and this is systematically violated by the crowd-sourcing setup, and *ii)* calculating Kappa on a per-answer basis is not sensible in cases where only one class occurs, as in all-focus and no-focus answers.

Overall agreement of crowd with gold standard

We performed the evaluation on the CREG-5K data subset for which we obtained both expert and crowd annotations. Figure 6.3 shows the observed per-token percentage agreement reached by the crowd workers compared to the gold standard annotation.

As reference, the dotted lines show the percentage agreement between the two expert annotators. We see that the quality improves from 74.9% for one worker to 79.8% for eleven workers[4]. Given that this is below the agreement of 88.8% reached by the expert annotators for this data set, we next investigated which cases the crowd can handle, and which ones turn out to be difficult for the non-experts.

[4] Note that agreement does not improve when increasing from odd to even worker numbers, which is due to the fact that the probability of drawing a majority does not increase in these cases.

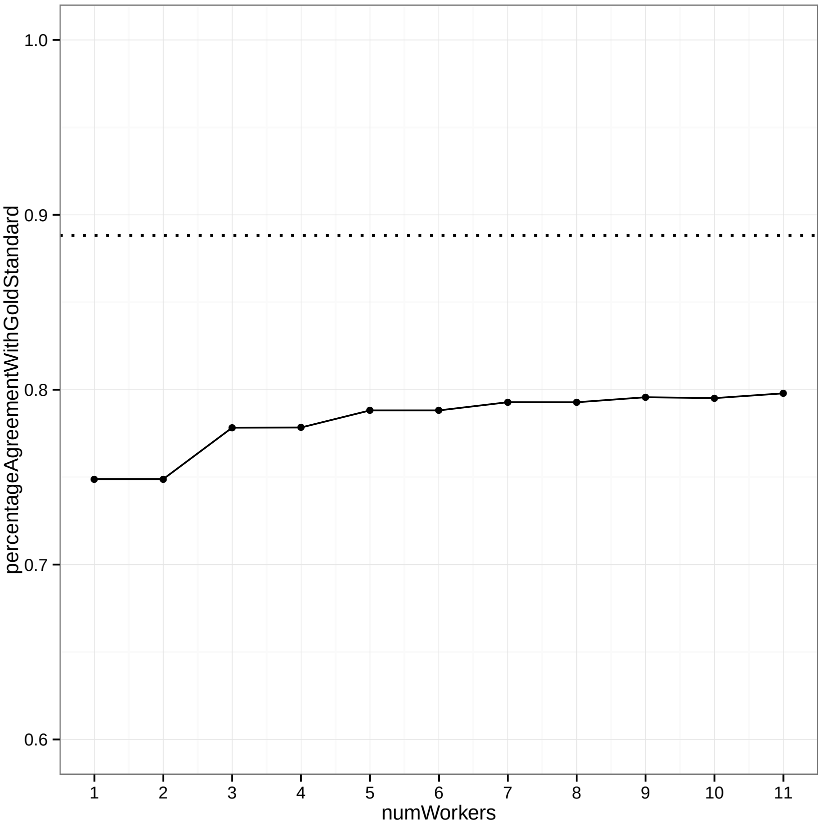


Figure 6.3.: Agreement of crowd with gold standard

Evaluation for different question forms

To identify patterns that show which types of data can be annotated with focus most consistently by crowd workers compared to the experts, we particularly want to look at properties of our data that take characteristics of the context into account – which in our case is the question context in which an answer annotated with focus occurs. We therefore investigated the impact of different types of questions on annotation agreement.

We carried out the comparison for the specific question form subtypes distinguishing surface forms of *wh*-questions as annotated in CREG (Meurers et al., 2011b). Figure 6.4 shows how the different question form subtypes impact the agreement between the crowd and the gold-standard focus annotation.

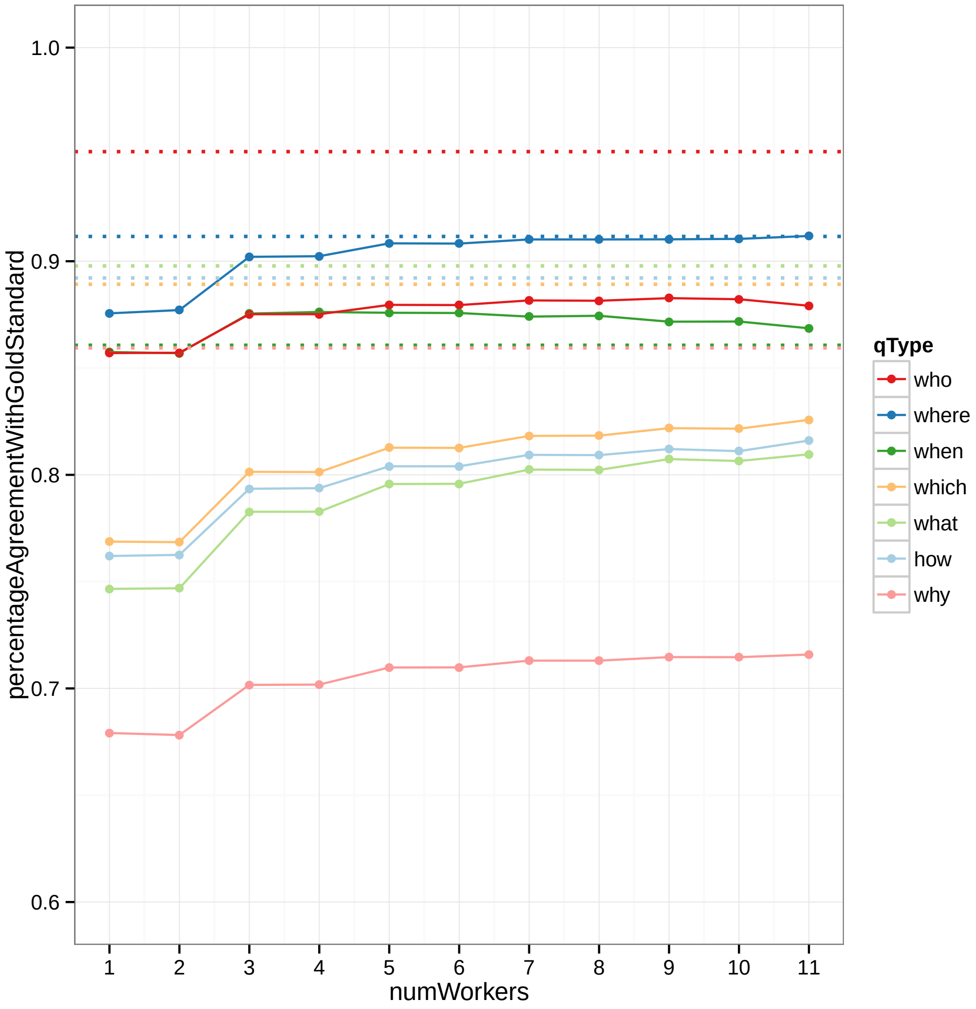


Figure 6.4.: Agreement by question form

As reference, the dotted lines again show the percentage agreements between the two expert annotators for the different question forms. The question forms make the answers fall into three broad categories in terms of worker-gold agreement: the most concrete ones (*who*, *when* and *where*) in terms of surface realization in answers come out on top with percentage agreements at 91% (*where*), 87% (*who*), and 86% (*when*).

The second group (*which*, *what* and *how*) are at 80–82% percentage agreement, which is likely due to their more ambiguous answer realization possibilities, e.g., a *what*-question can ask for an activity ('What did Peter do?') or an object ('What does Peter wear?').

The third group consists only of *why*-questions at an agreement level of 71%. For such questions asking for reasons, the range of possible answer realizations arguably is the greatest given that reasons are typically expressed by whole clauses. However, for the gold expert-annotation, the more explicit guidelines seem to have paid off in this case, as *why*-questions come out at a much higher agreement level of 86%.

To test whether more explicit guidelines could also help the crowd annotators to be more systematic in their focus annotation, we conducted a small additional crowd-sourcing annotation study with a smaller data set only containing answers to *why* and *what*-questions. While the general set up was the same as described in section 6.4.1, we provided the crowd workers with more examples illustrating focus in different kind of answers. The result was only a small improvement in agreement between crowd and gold standard annotation, with answers to *what*-questions 1% higher than before, and 2% higher for *why*-questions. Even more explicit guidelines thus do not seem to help the non-experts to handle answers occurring with *why*-questions when annotating focus.

Summing up the results so far, the crowd annotation study shows that i. the percentage agreement improves the more crowd workers are taken into account, and ii. majority voting on crowd worker judgments compared to the expert gold annotation can reach the expert level for specific cases (e.g., *where*-questions).

Qualitative discussion

To gain a better understanding of why the annotation agreement differs so widely with respect to question types for the crowd annotators, we take a closer look at the variation in the linguistic material that apparently impacts focus annotation. We discuss a typical example for a *who*-question (47) and a *why*-question (48) together with a sample of given answers from the CREG-5K data set as the two most extreme cases with respect to the observed annotation agreement.

In the case of the different answers to the *who*-question shown in (47), we can see that the variation both in meaning and form is very limited:

(47) Q: Wer war an der Tür?
who was at the door

A1: ⟦Drei Soldaten⟧$_F$ waren an der Tür.
three soldiers were at the door

A2: ⟦Drei Männer in alten Uniformen⟧$_F$ waren an der Tür.
three men in old uniforms were at the door

A3: ⟦Die drei Männer⟧$_F$ waren an der Tür.
the three men were at the door

A4: ⟦Drei alte Uniformen⟧$_F$ waren an der Tür.
three old uniforms were at the door

Syntactically, the focused part of the answers shown in ⟦. . . ⟧$_F$ is expressed as a nominal phrase. Contentwise, the same type of entity (a person) is expressed by semantically related words. The rest of the sentence shows no variation at all. The only inconsistency in annotation by the crowd occurred with NPs such as *Die drei Männer* in answer A3 in (47), where some of the crowd annotated the entire NP as the focus, while the rest of the crowd annotators only marked *drei Männer* as the focus, leaving out the definite article.

In the case of the various answers to the *why*-question shown in (48), multiple ways of answering the same questions can be observed, both syntactically and semantically.

(48) Q: Warum ist das Haus der Kameliendame so interessant?
why is the house of the lady of the camellias so interesting

A1: ⟦Ein Klimacomputer regelt Temperatur, Belüftung, Luftfeuchte
a air computer regulates temperature ventilation humidity
und Beschattung.⟧$_F$
and shading

A2: Das Haus der Kamelie ist so interessant, ⟦weil es 230 Jahre alt
the house of the camellia is so interesting because it 230 years old
und 8,90 m hohe ist.⟧$_F$
and 8.90 m high is

A3: ⟦In der warmen Jahreszeit wird das Haus neben die Kamelie
in the warm season is the house next to the camellia
gerollt.⟧$_F$
rolled

A4: Das Haus der Kamelie ist so interessant, ⟦weil es ist ein
the house of the camellia is so interesting because it is a
fahrbares Haus.⟧$_{\text{F}}$
mobile house

A5: Der Kamelie ist interessant ⟦wegen des Computers.⟧$_{\text{F}}$
the camellia is interesting because of the computer

Syntactically, the focused part of the answer is either expressed as the entire sentence as in A1 and A3 in (48), the subordinate clause starting with *weil* (because) as in A2 and A4 in (48), or as a PP introduced by *wegen* (because of) as in A5. Semantically, all four answers present a different propositional content. The relation between the question and potential answers thus is not particularly obvious or direct. Establishing the relation between question and answer – as needed to identify the focus of the answer – thus requires more effort by the annotator. This leads to less consistent results in the annotation for the crowd. For example, parts of the crowd annotators did not interpret the sentence A3 in (48) as an answer to the *why*-question in (48) at all and consequently did not mark any words in that sentence as focus, while the rest of the crowd annotators marked the entire clause as the focus.

For the expert annotators, the more explicit guidelines including a conceptual discussion of the key notions and explicit tests with minimal pairs, results in less pronounced differences in annotation quality for the different question types.

6.4.3. Predicting when the crowd is reliable

Apart from taking the question type into account, is it possible to predict when crowd focus annotation is particularly reliable based on characteristics of the crowd judgments?

Previous research on this issue has looked primarily at individual crowd worker characteristics, such as worker trustfulness (cf., e.g., Hantke et al. (2016). Hsueh et al. (2009) calculate sentiment ambiguity by considering the strength and the polarity of the sentiment's ratings. We here go into a similar direction for focus annotation, investigating the idea to take into account the diversity of the crowd performance, i.e., how diverse the focus annotations obtained from crowd workers for individual sentences are. Our hypothesis here is that sentences where the crowd agrees more on the annotation are annotated more reliably.

Calculating the cost of crowd consensus

We propose to measure the diversity of the focus annotation provided by the crowd workers in terms of the **Consensus Cost** in annotating a sentence of length n. The Consensus Cost (CC) is defined to be the sum of the minority annotation (i.e., focus or background) for all tokens in a sentence divided by the total number of tokens and the largest possible minority annotation for a token (in our case 5, since 6 would be a majority with 11 workers).

$$CC = \frac{\sum_{w=0}^{n} changeNeededForConsensus(w)}{largestPossibleMinority \times n} \tag{6.1}$$

The formula measures how many annotation changes would be needed to reach total consensus in annotating a given token. Sentences where the crowd workers mostly agreed on an annotation have a low consensus cost, because for every token only few annotation changes are needed to reach total agreement. Sentences where a larger number of workers diverge from the majority annotation have a higher consensus cost, since more changes would be needed in order to reach complete consensus on that annotation.

Figure 6.5 exemplifies the calculation of the Consensus Cost for the actual eleven crowd annotations from the crowd-sourcing experiment for the short example answer *Die/the drei/three Männer/men war/was an/at der/the Tür/door* from our CREG data.

For the first word *die*, only two of the 11 crowd workers marked the word as Focus, so the cost to reach total agreement (in this case that the token is (b)ackground, i.e., not focus) is 2. The next two words (*drei/three*) and *(Männer/men)* were marked as focus by 10 of the 11 of workers and thus each have a cost of one. The rest of the words in the sentence were unanimously not marked as focus by the crowd workers and thus have a cost of 0. The resulting Consensus Cost for the focus annotation for this sentence according to our formula is 0.11.

Since not all crowd workers perform equally well, it would in principle make

	Die	drei	Männer	war	an	der	Tür
1	**F**	F	F	b	b	b	b
2	**F**	F	F	b	b	b	b
3	b	F	F	b	b	b	b
4	b	F	F	b	b	b	b
5	b	F	F	b	b	b	b
6	b	F	F	b	b	b	b
7	b	F	F	b	b	b	b
8	b	F	F	b	b	b	b
9	b	F	F	b	b	b	b
10	b	F	F	b	b	b	b
11	b	**b**	**b**	b	b	b	b
Cost	2	1	1	0	0	0	0

$$\text{ConsensusCost} = \frac{4}{5\times 7} = 0.11$$

Figure 6.5.: Calculating the Consensus Cost

sense to incorporate their individual reliability. As a first step towards this idea, we are excluding all workers from annotation who fail to reach a particular accuracy threshold (0.6) on the test questions.

We can now investigate whether the Consensus Cost, i.e., the amount of agreement within the crowd, can serve as an indicator of the quality of the annotations provided by the crowd.

Consensus Cost and Annotation Quality

In order to determine whether Consensus Cost can function as a proxy for annotation quality, let us compare it to the agreement of the crowd workers with the gold standard expert annotation we discussed in section 6.4.2.

To explore the relation between Consensus Cost and quality of the annotation of an answer, we divided the possible values (0.0 to 1.0) of Consensus Cost into four ranges, using 0.25, 0.5 and 0.75 as boundaries. Figure 6.6 shows the boxplots for each of the four groups of answers by Consensus Cost, with the percentage agreement with the gold standard shown on the y-axis. The width of the box plots indicates the number of instances represented, whereas the height represents the distribution of agreement values.

For answers annotated with low Consensus Cost (< 0.5), the quality of annotation is generally high, with agreement with the gold standard between

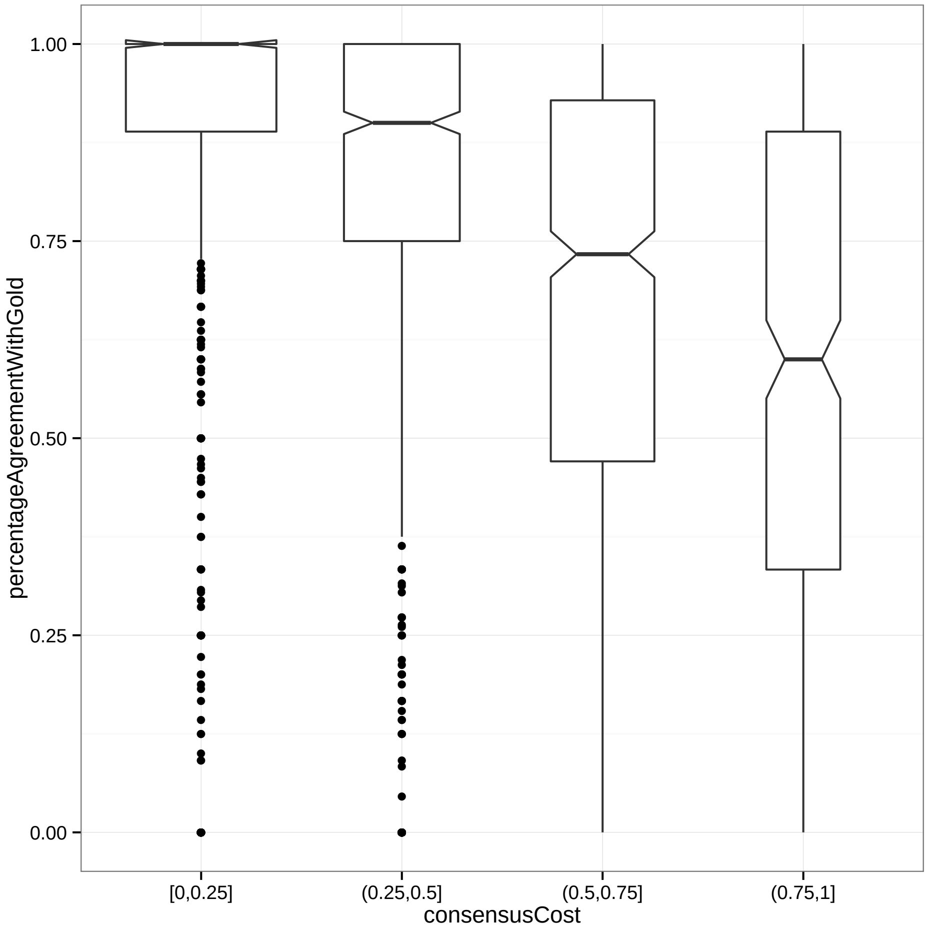


Figure 6.6.: Consensus Cost and Annotation Quality

0.7 and 1.0. The majority of data points fall into this interval. Interestingly, answers annotated with higher Consensus Cost values, in the intervals (0.5,0.75] and (0.75,1], show a more heterogeneous picture. While their median agreement is much lower, they also show a more varied distribution, including some high quality annotations.

In sum, we can conclude 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. The relationship is not a simple linear one, though, so that some annotations with high Consensus Cost may also be of high quality.

Consensus Costs by Question Type

When we evaluated the quality of the crowd focus annotation in relation to the gold-standard expert annotation in section 6.4.2, we found that the crowd annotations fall into three groups with respect to question types: Answers to the *who*, *when* and *where* questions showed a high percentage agreement with the expert annotation, answers to *which*, *what* and *how* questions had a much lower percentage agreement and answers to *why* questions were the most difficult ones for the crowd and had the lowest agreement numbers. The data by question type thus makes an interesting test case for Consensus Cost as a proxy for annotation quality. If sentences with a low consensus cost provide annotation of higher quality, we should be able to find a similar division of the annotation in terms of question types as in comparison with the expert annotation.

Figure 6.7 shows the consensus cost of our crowd annotation plotted according to question types. The figure shows clear differences by question type: The annotations of answers to *who*, *when*, and *where* questions have the lowest consensus costs, while answers to *why* questions have highest cost. And in addition, focus annotations of answers to *why* and *how* are most varied.

Consensus Cost by question type thus patterns parallel to the quality of the crowd annotation compared to the expert annotation. The analysis by question type thus confirms the overall analysis in the previous subsection establishing a low Consensus Cost in crowd annotation as a proxy for high quality annotation.

Summary

In this chapter, we presented our effort at manual focus annotation, both by experts and by crowd workers. We first discussed possible sources of evidence for focus annotation in corpus data, pointing out that surface criteria are not sufficient and settling on meaning-based criteria instead.

We then described our annotation scheme, including the annotation setup and tool used. The scheme operationalizes focus by making use of the explicit question to determine relevant alternatives. In order to pin down the extent of the focus, we use a substitution test with which the focus status of an individual word can be determined.

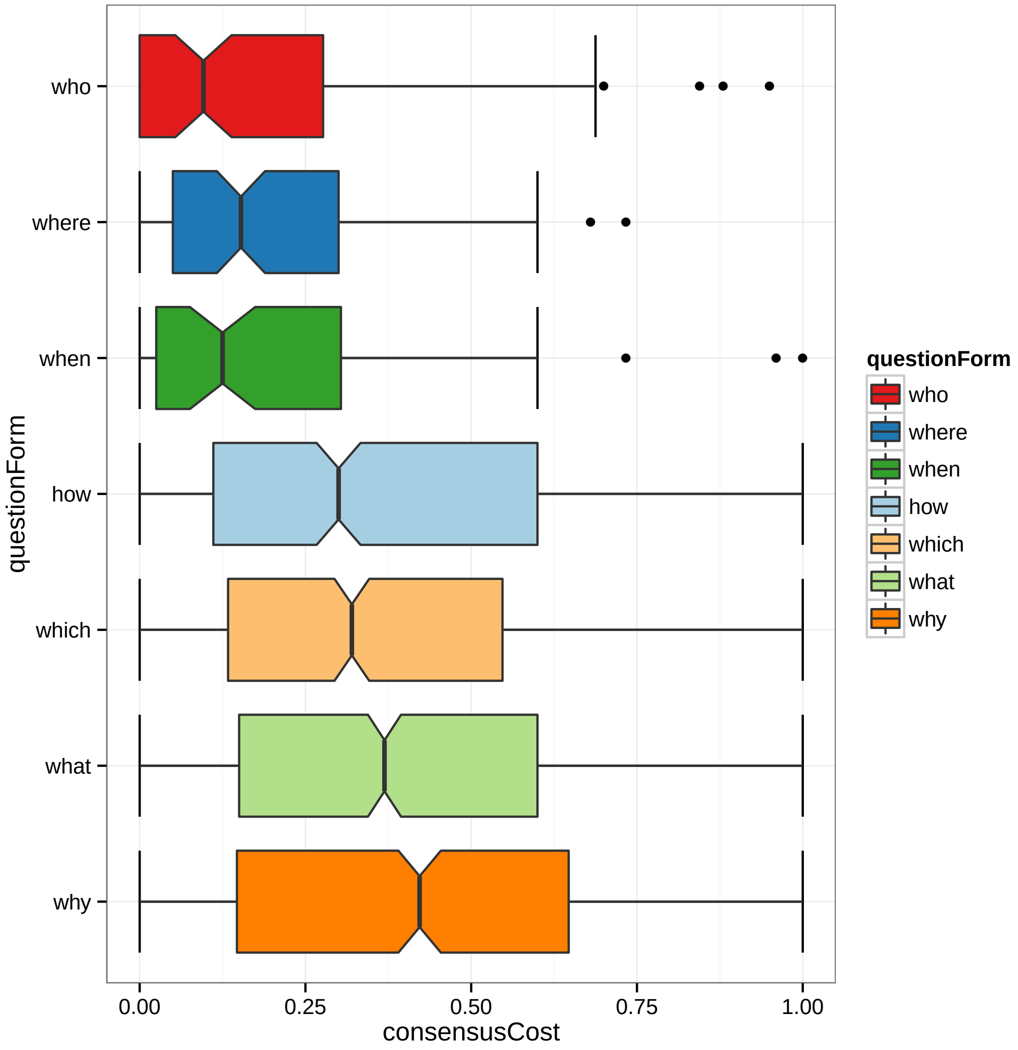


Figure 6.7.: Consensus Cost per Question Type

In the next section, we presented the results of our expert annotation study, which was conducted in two annotation phases. It produced 4,177 annotated answers with an agreement of $\kappa = .7$, which is very competitive with regard to the state of the art, where results reported were generally lower (cf., e.g., Ritz et al. 2008; Calhoun et al. 2010).

Finally, we described a crowd-sourcing experiment, showing that crowd workers do provide reliable focus annotation for some types of data, which is especially apparent in an analysis according to question types: crowd workers reach near-expert level for *who-*, *when-* and *where*-questions, produce acceptable results for *which-*, *what-* and *how*-questions and perform poorly on *why*-questions. We also showed how the quality of crowd annotation may be predicted independently of a comparison with experts, using Consensus Cost, an agreement-based measure we defined.

7. Automatic Focus Detection

In the previous chapter, we have discussed at length how focus can be annotated manually either by expert annotators, or through crowd-sourcing. In this chapter, we turn our attention to classifying focus in answers automatically. In doing so, we first review the few other works that have dealt with focus classification, and outline how our work differs from theirs (section 7.1). We then delve into a discussion of what observable linguistic properties could constitute good features for focus classification (section 7.2), before describing the resulting model (section 7.3) and classification approach (section 7.4). Next, we present results in terms of a comparison with the human-annotated gold standard (section 7.5) and showcase some key examples of success and failure that point out ways to improve the approach (section 7.6). Finally, we discuss three concrete extensions (section 7.7) for the initial model we implemented and show the impact they have on classification performance (section 7.8). Parts of this chapter are published in Ziai & Meurers (2018).

7.1. Previous Approaches

In this section, we briefly review relevant related work in the area of focus detection[1]. Overall, there is very little work done in the area of automatically determining a notion we would call focus as defined in chapter 3, and it almost exclusively centers on detecting the 'kontrast' notion in the English Switchboard corpus (see section 3.3.2). We therefore start with the Switchboard-based approaches before moving to the other work.

[1] For a broader perspective of computational approaches in connection with IS, see Stede (2012).

7.1.1. Switchboard-based Approaches

The availability of the annotated Switchboard corpus (Calhoun et al., 2005, 2010) sparked interest in information-structural categories and enabled several researchers to publish studies on detecting focus. This is especially true for the Speech Processing community, and indeed many approaches described below are intended to improve computational speech applications in some way, by detecting prominence through a combination of various linguistic factors. Moreover, with the exception of Badino & Clark (2008) and Zang et al. (2014), all approaches use prosodic or acoustic features.

All approaches listed below tackle the task of detecting 'kontrast' (as focus is called in the Switchboard annotation) automatically on various subsets of the corpus using different features and classification approaches. For each approach, we therefore report the features and classifier used, the data set size as reported by the authors, the (often very high) majority baseline for a binary distinction between 'kontrast' and background, and the best accuracy obtained. If available in the original description of the approach, we also report the accuracy obtained without acoustic and prosodic features.

Calhoun (2007)

Calhoun (2007) investigated how focus can be predicted through what she calls "prominence structure". The essential claim is that a "focus is more likely if a word is more prominent than expected given its syntactic, semantic and discourse properties". The classification experiment is based on 9,289 words with a 60% majority baseline for the 'background' class. Calhoun (2007) reports 77.7% for a combination of prosodic, syntactic and semantic features in a logistic regression model. Without the prosodic and acoustic features, the accuracy obtained is at 74.8%. There is no information on a separation between training and test set, likely due to the setup of the study being geared towards determining relevant factors in predicting focus, not building a focus prediction model for a real application case. Relatedly, the approach uses only gold-standard annotation already available in the corpus as the basis for features, not automatic annotation.

Nenkova & Jurafsky (2007)

Nenkova & Jurafsky (2007) use a combination of acoustic, prosodic and part of speech features in a logistic regression model. The main motivation for the work is to improve both automatic speech understanding and text-to-speech synthesis. The data set consists of 7,785 words, of which 72.38% belong to the 'background' majority. The model was tested using 10-fold cross-validation and obtained an accuracy of 76.88% using all features. Using only part-of-speech features, accuracy only decreases slightly to 76.42%.

Sridhar et al. (2008)

Sridhar et al. (2008) use lexical, acoustic and part-of-speech features in trying to detect pitch accent, givenness and focus. Concerning focus, the work attempts to extend Calhoun (2007)'s analysis to "understand what prosodic and acoustic differences exist between the focus classes and background items in conversational speech". 14,555 words of the Switchboard corpus are used in total, but filtered for evaluation later to balance the skewed distribution between 'kontrast' and 'background'. With the thus obtained random baseline of 50%, Sridhar et al. (2008) obtain 73% accuracy when using all features, which again drops only slightly to 72.95% when using only parts of speech. They use a decision tree classifier to combine the features in 10-fold cross-validation for training and testing.

Badino & Clark (2008)

Badino & Clark (2008) aim to model contrast both for its role in analyzing discourse and information structure, and for its potential in speech applications. They use a combination of lexical, syntactic and semantic features in an SVM classifier. No acoustic or prosodic features are employed in the model. In selecting the training and testing data, they filter out many 'kontrast' instances, such as those triggered across sentence boundaries, those above the word level, and those not sharing the same broad POS with the trigger word. The resulting data set has 8,602 instances, of which 96.8% are 'background'. The authors experiment with different kernel settings for the SVM and obtain the best result of 97.19% using a second-order polynomial kernel, and leave-one-out testing.

Zang et al. (2014)

Zang et al. (2014) build on the work by Badino & Clark (2008) and extend it by more explicitly modeling the relationship between the trigger and the instance of 'kontrast', and by using a CRF classifier. They base their experiment on 70,767 instances, where each instance is a word pair: for 'kontrast' instances, the 'kontrast' and the trigger form a pair, and for 'background', the first word in the same broad part-of-speech class forms a pair with the background word. All cases without any correspondence in broad part-of-speech class within the same sentence are discarded. The majority baseline is 88.86%, and the best model achieves an accuracy of 94.98%.

7.1.2. Other Approaches

Zhang et al. (2006) present work in detecting so-called 'focus kernels', which are to be understood as the novel part of an utterance. The work is based on an annotation of dialogues resulting from interactions of children with an Intelligent Tutoring System on basic math and physics concepts through illustration with Lego gears. The annotation reportedly target focus and contrast, however focus is defined in terms of newness and explicit contrast between words here, and not in terms of alternatives. The features for detecting the so-defined focus include prosodic prominence, knowledge- and corpus-based semantic relatedness measures and part-of-speech tags. The evaluation data set consists of 5,700 transcribed words and 48 minutes of corresponding speech waveforms, which were both fed into a time-delay recurrent neural network (Kim, 1998). The data set was partitioned into 90% for training and 10% for testing, the latter consisted of 536 words. Zhang et al. (2006) achieved an accuracy of 83.8% on the test set, where the majority ('nonfocus kernel') occurs in 69.2% of the instances.

7.1.3. Delineation of Our Approach

While the work discussed above is relevant to ours, there are several important differences that set us apart. First, in contrast to all approaches, we target the analysis of written texts, for which prosodic and acoustic information is not available, so we must rely on lexis, syntax and semantics exclusively.

Second, our annotation scheme and hence our classification approach is not based on a pre-selection of certain word classes, because our definition of focus is not based on syntactic constraints, but relies rather on a semantic notion of answerhood, where the notion of alternative sets in the sense of Rooth (1985, 1992) is operationalized in the annotation scheme.

Third, we tackle a different language, namely German, whereas all previous approaches are designed for English, where especially prosody follows very particular patterns unique to the language. We also have the added difficulty of dealing with language from foreign language learners due to the nature of our empirical basis.

Finally, the vast majority of the approaches discussed make direct use of the manually annotated information in the corpus they use in order to derive their features. While this is a somewhat viable approach when the aim is to determine the relevant factors for focus detection, it does not represent a real-life case where annotated data often unavailable. In our focus detection model, we only use automatically determined annotation as the basis for our features for predicting focus.

7.2. What Can Inform Focus Detection?

Let us now take a step back from related work and consider what may inform our own focus detection approach. We have seen what characteristics are relevant for manual annotation of focus in chapter 6, but these factors do not necessarily lend themselves directly to automatic focus detection. Therefore, we now turn to the issue of which automatically obtainable linguistic factors enable focus to be detected in answers to explicit questions. We will proceed by discussing relevant observable question and answer properties along with a computational approximation of Givenness.

7.2.1. Question Properties

When thinking about question-related features, it makes sense to first go back to our understanding of focus and the role the question plays in this understanding. Most importantly, the question defines the **alternative set**, as demonstrated in example (49):

(49) Who came to the party?
⟦John⟧$_F$ came to the party.

Here, the question *Who came to the party?* defines the alternative set of people through the *wh*-phrase *Who*. There is also the part of the question which is fully specified, namely the VP *came to the party*. This part can be seen as introducing the topic and setting the scene, but it also constrains the alternative set: the answer now needs to provide an element that is both a person and that came to the party. In order to satisfy both question requirements, the answer must select from the final alternative set, and it does so by providing one set member, *John*.

However, as already mentioned in chapter 6, we cannot robustly construct semantic representations of questions (or answers). For this reason, we must rather ask ourselves what reliable surface indicators of the relevant question properties discussed above are.

One property that quickly comes to mind is the question word itself. Question words such as *who* or *what* are a closed class and some of them have a strong connection to the alternative set: a *who*-question will likely ask for a person or other living entity, a *when*-question usually asks for a temporal expression, and so on.

Some question words, however, have multiple functions. For example, *what* can be the whole *wh*-phrase, as in *What did you eat today?*, or it can be part of the *wh*-phrase, as in *what*+NP questions such as *What city do you live in?*. It follows that in addition to the identity of the question word, one should also look at its context, such as the words next to it or its grammatical function in the question. The latter can be determined robustly by state-of-the-art dependency parsers, which can distinguish the first use of *what* as object from the second use as determiner. Such distinctions may even be possible using the question word's part-of-speech: the German STTS tag set (Schiller et al., 1995) in fact distinguishes between *wh*-determiners and *wh*-pronouns.

Concerning the non-*wh* part of questions, a robust approximation of constraints on the alternative set could be the content words present in that part. In example (49), these would be the words *came* and *party*. To model them as restrictions for possible foci, one could specify features which express their presence or absence in proximity to the focus in the answer.

Finally, not all questions are *wh*-questions: Polar and alternative questions also occur in our data, albeit with much lower frequency. In polar questions such as *Did John come to the party?*, the alternative set always contains two elements, p and $\neg p$, where p is the proposition expressed in the question. In alternative questions such as *Did John come to the party or did Frank come to the party?*, the alternative set can have more than two elements, but they are all explicitly given in the question. Both types of questions exhibit a different syntax compared to *wh*-questions: the auxiliary verb in the beginning is a strong indicator, and for alternative questions, there is usually a coordination structure present, involving the conjunction *or*.

7.2.2. Answer Properties

Looking at the answer in (49), we are faced with similar problems as when determining robust features for questions. As already mentioned, focus needs to select from the alternative set, so it must correspond to the type of that set. In (49), this condition is satisfied because *John* is a person who presumably came to the party. Precisely keeping track of such semantic relationships would mean building up a model of contextual and world knowledge where the relations can be looked up. This is of course not feasible to do manually, and would be fraught with errors if done automatically.

Starting with the most basic features that describe an answer word, there is the word's surface form itself or the lemma as an abstraction over its inflection. However, neither offers any insight into the semantics of the word by itself, and would occur too rarely for the learning algorithm to transfer to unseen data. The next more general attribute is the word's part-of-speech, which offers a more useful abstraction: semantic objects such as persons or places, for example, tend to be realized as nouns whereas actions tend to be realized as verbs. Sridhar et al. (2008) report that parts of speech are a competitive baseline for predicting focus[2]. The logical next step would be to characterize the words syntactic role beyond its part-of-speech, such as its grammatical function or the type of phrase it belongs to. Analogous to question words, such information can be obtained robustly by state-of-the-art parsers.

[2] However, this is somewhat unsurprising given that the focus annotation in the Switchboard corpus is restricted to certain word classes.

Another potentially relevant property is the position of a word in the answer. In our review of the Topic-Focus Articulation annotation approach of the Prague Dependency treebank in section 3.3.2, we have already seen that IS can interact with word order in a free word order language such as Czech, so it likely plays a role in other languages too. Besides encoding the word position numerically, there is also the possibility to characterize its immediate context or the topological field (Höhle, 1986) they occur in.

7.2.3. Givenness

As we outlined in chapter 3, the given vs. new distinction is quite separate from the focus vs. background one: focused material does not necessarily have to be new and background material does not have to be given. Nevertheless, the two often coincide, as shown in example (50), where *in Paris* is both new and focused.

(50) Where is the Eiffel tower?
The Eiffel tower is $[\![$in Paris$]\!]_F$.

Approaches such as Meurers, Ziai, Ott & Bailey (2011a) and Mohler et al. (2011) have successfully exploited this by excluding given material from answer evaluation, though only the former established the connection to IS. So while givenness represents a different IS dimension, it can be a useful factor in predicting focus.

The systems mentioned above have only used a very basic version of givenness which compares answer words to question words in terms of their surface forms. However, the phenomena involved in Givenness as defined by Schwarzschild (1999) include lexical entailment and co-reference, thus going well beyond simple string comparisons. It may therefore be worth investigating how true Givenness can be approximated computationally.

7.3. Evidence Used for Classification

We have motivated various types of linguistic information on different levels that can in principle be relevant for focus identification in the previous section. For our initial focus detection model (see section 7.7 for the extensions we made to it), we explored five different groups of features: lexical/syntactic answer properties, question properties, givenness, positional properties and conjunction features.

For some groups, we experimented with different concrete features manually selected them based on whether they improved overall accuracy or not. In follow-up work, we plan to replace this explorative methodology with an automatic feature selection approach able to determine good combinations of features. Nevertheless, in order to provide insight into which features did not improve classification, we discuss some of them at the end of this section.

The instance we characterize through the following features is a word within a sentence of the answer, as identified by the OpenNLP tokenizer and sentence segmenter[3], following the procedure we used to calculate inter-annotator agreement in chapter 6.

7.3.1. Syntactic Properties of the Answer (SynAns)

A word's part-of-speech and syntactic function are relevant general indicators with respect to focus: since we are dealing with meaning alternatives, the meaning of e.g. a noun is more likely to denote an alternative than a grammatical function word such as a complementizer or article.

Similarly, a word in an argument dependency relation is potentially a stronger indicator for a focused alternative in a sentence than a word in an adjunct relation. We therefore included two features: the word's **part-of-speech** tag in the STTS tag set (Schiller et al. 1995, see appendix B) determined using TreeTagger (Schmid, 1994), and the **dependency relation to the word's head** in the Hamburg dependency scheme (Foth 2006, see appendix B) determined using MaltParser (Nivre et al., 2007) as features in our model.

7.3.2. Question properties

The question constitutes the direct context for the answer and dictates its information structure and information requirements to fulfill. In particular, the type of *wh*-phrase (if present) of a question is a useful indicator of the type of required information: a *who*-question, such as 'Who rang the doorbell?',

[3] `http://opennlp.apache.org`

will typically be answered with a noun phrase, such as 'the milkman'. We identified **surface question forms** such as *who*, *what*, *how* etc. using a regular expression approach developed by Rudzewitz (2015) and included them as features. Related to question forms, we also extracted the question word's **dependency relation to its head**, analogous to the answer feature described above.

7.3.3. Surface givenness

As a rough and robust approximation to information status, we add a boolean feature indicating the **presence of the current word in the question**. We use the lemmatized form of the word as determined by TreeTagger (Schmid, 1994).

7.3.4. Positional properties

Where a word occurs in the answer or the question can be relevant for its information structural status. It has been observed since Halliday (1967) that given material tends to occur earlier in sentences (here: answers), while new or focused content tends to occur later. We encode this observation in three different features: the **position of the word in the answer** (normalized by sentence length), the **distance from the finite verb** (in words), and the **position of the word in the question** (if it is given).

7.3.5. Conjunction features

To explicitly tie answer properties to question properties, we explored different combinations of the features described above. Specifically, we encoded the **current word's POS depending on the question form**, and the **current word's POS depending on the *wh*-word's POS**. To constrain the feature space and get rid of unnecessary distinctions, we converted the answer word's POS to a coarse-grained version before computing these features, which collapses all variants of determiners, pronouns, adjectives/adverbs, prepositions, nouns and verbs into one label, respectively[4].

[4] For a list of the full tag set, see appendix B

Features Without Positive Impact

Several features were tried out and excluded again from the model due to their failure of improving the model. In a very early attempt, we used the surface form and lemma of the current word, which are likely too specific and do not generalize well. A similar result was obtained for features concerning specific syntactic constructions: we experimented with a feature indicating whether the current word is in a subordinate clause, determined using its position in the dependency syntax tree, but this also did not result in an improvement.

Regarding conjunctive features, we also tried out combinations of the current word's POS or dependency relation depending on the *wh*-word's dependency relation. Contrary to the conjunctive features described above, this did not improve the results, possibly due to data sparseness resulting from the many possible feature value combinations.

A different line of features were concerned with properties of neighboring words. We tried including the prediction outcome of the previous word as a feature for the current word, which was likely too noisy to contribute predictive power. The general validity of this approach was however demonstrated by using the gold version of this feature, i.e., the manually annotated focus label of the previous token, which improved results significantly but is not a realistic feature. We also experimented with using syntactic and givenness features of the previous and the next word as additional features for the current word, but this did not improve results either.

7.4. Classifier Training

Having motivated and developed our features in the previous sections, the question that arises next is what machine learning approach to use for training and testing an actual model. There are three largely distinct sub-problems to this issue, which we will address in separate sub-sections:

1. Choice of machine learning algorithm: what are relevant criteria for the choice of an algorithm for a task such as focus detection?
2. Training/testing setup: how can we partition the available data in such a way that sufficient training data is available while avoiding overfitting

and also setting data aside for extrinsic evaluation (see chapter 8)?

3. Expert vs. crowd-sourced training data: what is the optimal way of combining these? Which should be used for training, and which for testing?

7.4.1. Classification Algorithm

In choosing the classification approach, let us first recall that the basic unit we are going to classify is an individual word, following the decision we made for manual annotation in chapter 6. The task for the machine learning approach would thus be to decide for each word of an answer whether it is part of the focus or not. Any supervised classification approach is in principle applicable to this task, from Naive Bayes to Support Vector Machines (cf., e.g., Schölkopf & Smola 2003), or lazy learning approaches such as *k*-nearest-neighbor.

A first filtering criterion for classification approaches, or rather their implementations, is the **range of feature types** they can handle. Most of the features we described in the previous sections are string features, such as the part-of-speech of a word or the label of a dependency relation. Not all learning approaches can handle these, a problem which toolkits like WEKA (Hall et al., 2009) circumvent by converting string features to binary features which encode the presence or absence of the respective string feature. This has the side effect of blowing up the feature space, which can be a problem if either the training set is not large enough or the algorithm has problems with large numbers of features. Some algorithms are able to handle string features natively, either through the use of a string kernel function (Support Vector Machines and other kernel-based methods) or a string similarity function (*k*-nearest-neighbor and other lazy learning approaches).

Another important criterion is the **complexity** of an algorithm, both computationally and in terms of understandability for the researcher. Computational complexity translates to training time, which may seem as a practical non-issue with today's fast machines, but if one needs to try out many different feature combinations with many different data sets, training time can quickly become relevant. As far as understandability is concerned, the more complex an approach is, the harder it becomes to gain insight into the impact of particular features. If one is more interested in the factors of a linguistic phenomenon than in optimal classification accuracy, as we are here, then studying features is more important than bleeding edge machine learning.

Finally, one possible extension of the learning problem in connection with focus annotation is to classify **sequences of words** instead of words in isolation. Since the focus status of a word does to a certain extent depend on the focus status of the words around it, it makes sense to try an approach such as Conditional Random Fields (CRFs) (Lafferty et al., 2001) which is able to take the labels of previous instances into account.

We chose to run our experiments using **logistic regression** (Cox, 1958), as an learning approach that satisfies the criteria of feature types and complexity. It is a fast and well-understood algorithm for which efficient implementations are available that can handle very large numbers of features. We use the WEKA implementation which automatically converts nominal features to binary ones, and we use this functionality for all our non-numerical features, such as part-of-speech and dependency labels. Another advantage of logistic regression is that it is still being applied to current problems, including the ones we are interested in: Heilman & Madnani (2013) successfully employed it for SAA, and Calhoun (2007) as well as Nenkova & Jurafsky (2007) used logistic regression for focus identification.

Using early versions of our model, we also experimented with Support Vector Machine (SVM) implementations, but while taking significantly more time for training, their use did not result in better accuracy over the logistic regression approach. Likewise, we tried out an implementation of CRFs[5], which also did not result in better performance. We suspect that in order to leverage the power of sequence classification approaches, one needs to invest a significant amount of time into tuning feature templates, i.e., the specifications that determine which combinations of features from adjacent words are taken into account, and how large the context window should be. While such templates have been defined and used successfully for tasks such as chunking (cf., e.g., Tjong Kim Sang & Buchholz 2000), determining an optimal feature template for a less-studied task such as focus detection involves trying out a very large number of possible combinations. This is thus a topic we leave for future work.

[5] https://taku910.github.io/crfpp/

7.4.2. Training and Testing Setup

In order to obtain realistic and meaningful results, it is important to set up fair and generalizable training and testing setups. To accomplish this, it is standard practice to hold out a test set until all tuning and parameter estimation has been done. The tuning is either performed on a held-out development set, or via cross-validation on the training set.

The trade-off one has to deal with here is to keep the training set large enough to arrive at a robust model, but at the same time having enough testing data for results to be representative. In addition to that, one has to be careful to avoid overfitting: if one uses a setup without an explicit development set, it is easy to tune parameters too much to fit the training set. To counteract this effect while still using the maximum amount of training data available, one can make sure the folds in cross-validation are similar to the test set in size.

In our case, the situation becomes somewhat more complicated, because we want to also perform extrinsic evaluation of focus detection in SAA: to make sure that neither the focus classifier nor the SAA classifier have seen any test data before, the training data from focus detection also needs to be held out from SAA. This can quickly lead to insufficient training data in either classifying focus or classifying answers.

Our approach is the following: in order to make sure we do not train the focus classifier on data later used for testing SAA, we take CREG-ExpertFocus and remove all answers that are also in the CREG-5K test sets we use in chapter 8. The remainder is used to evaluate focus detection via 10-fold cross-validation. In order to avoid training the SAA classifier on data already used to train the focus classifier, we take the CREG-5K training set and remove all answers that are also in CREG-ExpertFocus. This is necessary because otherwise focus detection would likely produce unrealistically good results since the data was seen before, and the SAA classifier would trust these results too much, making it very difficult to generalize to unseen data.

The result is a focus detection training set with 3,064 answers (the 4,177 from CREG-ExpertFocus minus the ones in the CREG-5K test sets), and an SAA training set of 1,606 student answers (4,136 student answers from the CREG-5K training set minus the ones in CREG-ExpertFocus).

7.4.3. Expert vs. Crowd-sourced Data

Having carried out both expert (section 6.3) and crowd-sourced (section 6.4) focus annotation on various data sets, we are faced with the non-trivial question of what to use which type of annotated data for. Although the experts and the crowd agree to a large extent, their annotation behavior is nonetheless different, and it is possible a machine learning approach could be confused by these different versions of annotated focus.

Since the expert annotation is our reference for operationalized focus annotation in authentic data, we decided to start our focus detection experiments with expert annotation only, both for training and testing. In order to answer the question whether crowd data helps in addition to expert data, it can then be added gradually when a performance baseline has been established.

7.4.4. Summary of Setup for Testing Automatic Focus Detection

Summing up, our setup is as follows: we trained a logistic regression model using the WEKA toolkit (Hall et al., 2009). The data set used consists of the 4,177 answers of expert focus annotation available in CREG-ExpertFocus (see section 6.3), with the exception of the answers occurring in the extrinsic evaluation test sets we use in chapter 8, a total of 3,064 answers, of which 2,240 are student answers and 824 are target answers. We used 10-fold cross-validation on this data set to experiment and select the optimal model for focus detection.

On the technical side, we faced the problem of how to integrate the components used for annotation and feature extraction, so that the end result is both easily usable for intrinsic experimentation and testing, and can quickly be compiled into a focus detection model usable within the CoMiC system. Our solution was to build a UIMA pipeline using DKPro components (as we have done for CoMiC in chapter 5) for focus detection, whose output is a feature file readable by the WEKA toolkit. We use the ClearTK library (Bethard et al., 2014) for interfacing UIMA with machine learning components, as it allows to switch between training and testing mode very easily: once focus detection results have been deemed satisfactory, one can switch to testing mode and use the same ClearTK component with the trained model for automatic annotation within a larger UIMA pipeline.

7.5. Results

Table 7.1 lists the accuracies[6] obtained for our different feature groups, as well as three baselines: a POS baseline, following Sridhar et al. (2008), a baseline that only includes the simple givenness feature, and the majority baseline. The majority class is *focus*, occurring in 58.1% of the 26980 cases (individual words).

	Accuracy for		
Feature set	*focus*	*background*	both
Majority baseline	100%	0%	58.1%
Givenness baseline	81.5%	42.5%	65.1%
POS baseline	89.2%	39.6%	68.4%
SynAns	82.8%	50.3%	69.2%
SynAns + Question	83.8%	53.1%	70.9%
SynAns + Question + Given	84.8%	62.0%	74.8%
SynAns + Question + Given + Position	84.9%	66.5%	77.2%
All of the above + conjunction features	85.2%	66.7%	77.4%

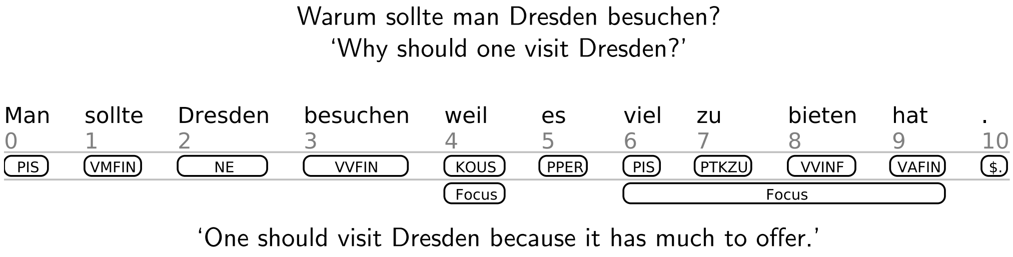
Table 7.1.: Focus detection performance using different feature sets

We can see that each feature group incrementally adds to the final model's performance, with particularly noticeable boosts coming from the givenness and positional features. Another clear observation is that the classifier is much better at detecting *focus* than *background*, possibly also due to the skewedness of the data set. Note that performance on *background* increases also with the addition of the 'Question' feature set, indicating the close relation between the set of alternatives introduced by the question and the focus selecting from that set, even though our approximation to computationally determining alternatives in questions is basic. It is also clear that the information intrinsic in the answers, as encoded in the 'SynAns' and 'Position' feature sets, already provides significant performance benefits, suggesting that a classifier trained only on these features could be trained and applied to settings where no explicit questions are available.

[6] We show per-class and overall accuracies, the former is also known as recall or true positive rate.

7.6. Qualitative Analysis

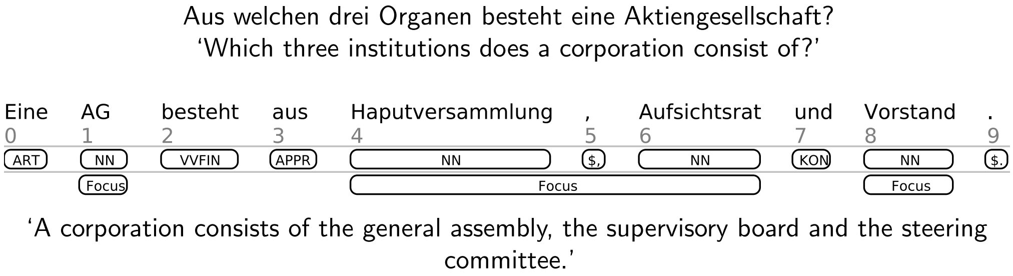
In order to complement our quantitative evaluation and shed light on typical focus detection errors, let us take a step back and qualitatively examine a few characteristic examples in more detail.


Warum sollte man Dresden besuchen?
'Why should one visit Dresden?'

'One should visit Dresden because it has much to offer.'

Figure 7.1.: Focus with a faulty gap in between

Figure 7.1 shows a case where a *why*-question is answered with an embedded *weil* ('because') clause. The classifier successfully marked *weil* and the end of the clause as *focus*, but left out the pronoun *es* ('it') in the middle, presumably because pronouns are given and often not focused in other answers. We did experiment with using a sequence classification approach in order to remedy such problems, but it performed worse overall than the logistic regression model we presented in the previous section. We therefore suggest that in such cases, a global constraint stating that *why*-questions are typically answered with a full clause would be a more promising approach, combining knowledge learned bottom-up from data with top-down linguistic insight.

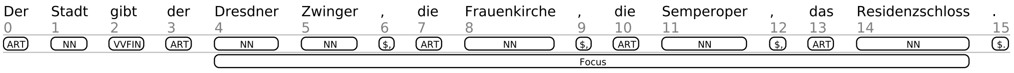

Aus welchen drei Organen besteht eine Aktiengesellschaft?
'Which three institutions does a corporation consist of?'

'A corporation consists of the general assembly, the supervisory board and the steering committee.'

Figure 7.2.: Focus with a faulty outlier (and a faulty gap)

In Figure 7.2, we can see two different problems. One is again a faulty gap, namely the omission of the conjunction *und* ('and'). The other is the focus marking of the word *AG* ('corporation') in the beginning of the sentence: since the question asks for an enumeration of the institutions that form a corporation, marking *AG* as focused is erroneous. This problem likely occurs often with nouns because the classifier has learned that content words are often focused. Moreover, the surface givenness feature does not encode that *AG* is in fact an abbreviation of *Aktiengesellschaft* and therefore given. It would thus be beneficial to extend our analysis of givenness beyond surface identity. We will get back to this issue in section 7.7.1.

Welche Sehenswürdigkeiten gibt es in der Stadt?
'Which places of interest are in the city?'



'The city exists the Dresden Zwinger, the Frauenkirche, the Semperoper, the Royal Palace.'

Figure 7.3.: Enumeration with correct focus

Finally, Figure 7.3 presents a case where an enumeration is marked correctly, including the conjunctive punctuation in between, showing that cases of longer foci are indeed within reach for a word-by-word focus classifier.

7.7. Extending the Approach

While our model has produced promising results already, there are naturally various ways in which the approach may be extended, some of which have become apparent through the qualitative evaluation in the previous section. In this section, we therefore discuss three extensions which we have implemented, and their impact on classification accuracy: a distributional approximation of givenness, a constituency-based approach to syntactic and topological features, and an exploration of using crowd-sourced data in focus detection.

7.7.1. Distributional Givenness

We have seen in section 7.5 that surface-based givenness is helpful in predicting focus. However, it clearly has limitations, as for example synonymy cannot be captured on the surface. We also exemplified one such limitation in Figure 7.2. In order to overcome these limitations, we implemented an approach based on distributional semantics. This avenue is motivated by the fact that in Ziai et al. (2016) we have shown that Givenness modeled as distributional similarity is helpful for SAA at least in some cases.

We first detail how the distributional model was built, before describing the features calculated using the model.

Creating a distributional model

To model Givenness as distributional similarity, we need an appropriate word vector model. While some others have developed models for German (cf., e.g., Dima 2015; Köper et al. 2015), we opted to train one ourselves in order to better tailor it to our needs.

As empirical basis, we used the DeWAC corpus (Baroni et al., 2009) since it is a large corpus that is freely available and already lemmatized, both of which have been argued to be desirable for word vector models. Further preprocessing consisted of excluding numbers and other undesired words such as foreign language material and words the POS tagger had labelled as non-words. The whole corpus was converted to lowercase to get rid of unwanted distinctions between multiple possible capitalizations.

To select an implementation for our purpose, we compared two of the major word vector toolkits currently available, word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). While word2vec is a prediction-based approach that optimizes the probability of a word occurring in a certain context, GloVe is a counting approach based on co-occurrences of words.

We compared the two on the lexical substitution task designed for GermEval 2015 (Miller et al., 2015). The task can be seen as related to recognizing Givenness: deciding what a good substitute for a word in context is requires similar mechanisms to deciding whether the meaning of a word is already present in previous utterances. For GloVe, we used the models trained by Dima (2015), which were also trained on a large German web corpus and were shown to perform well. However, results on the lexical substitution task put both of word2vec's training approaches, continuous bag-of-words (CBOW) and skip-gram, ahead of GloVe using the models previously mentioned, so we continued with word2vec.

Finally, to select the optimal training algorithm for word2vec for our purpose, we again used the GermEval task as a benchmark. We explored both CBOW and skip-gram with negative sampling and hierarchical softmax, yielding four combinations. Among these, CBOW with hierarchical softmax significantly outperformed all other combinations, so we chose it as our training algorithm.

The German model we obtained has a vocabulary of 1,825,306 words and uses 400 dimensions for each, the latter being inspired by Iacobacci et al. (2015).

Calculating Givenness

Having equipped ourselves with a word vector model, the question arises how to use it in focus detection in such a way that it complements the positive impact that surface-based givenness already demonstrates.

The fundamental representation offered by a distributional model is a vector for each word in its vocabulary which can be used as an approximation of its meaning, to be combined with or compared against vectors of other words. A common way of calculating similarity is to obtain the cosine of the angle between two vectors in multi-dimensional space (cf., e.g., Salton & McGill 1983). The resulting cosine similarity is a value between -1 and 1, where -1 indicates the exact opposite, 0 indicates an orthogonal vector and 1 indicates an identical vector.

In Ziai et al. (2016) we have used this measure in combination with an empirically determined threshold on answer vs. question words in order to decide whether an answer word is Given or not: if a cosine similarity with any of the question words exceeds the threshold, the word counts as Given. However, determining thresholds requires in-domain data, and we found that they are very dependent on the data set. Therefore, we here use raw cosine similarities[7] and calculate **maximum, minimum and average cosine between the answer word and the question words**. As a fourth feature, we calculate the **cosine between the answer word and the additive question word vector**, which is the sum of the individual question word vectors.

As a result of these four new features, accuracy on *focus* drops slightly from

[7] We normalize cosine similarity as cosine distance to obtain positive values between 0 and 2: $dist = 1 - sim$

85.2% to 84.7%, but rises noticeably on *background* from 66.7% to 68.0%, pushing the overall detection accuracy from 77.4% to 77.7%.

7.7.2. Constituency-based Features

Another source of evidence we wanted to exploit is constituency-based syntactic annotation. So far, we have worked with part-of-speech tags and dependency relations as far as syntactic representation is concerned. However, while discontinuous focus is possible, focus as operationalized in our expert annotation in chapter 6 most often marks an adjacent group of words. Such groups very often correspond to a syntactic phrase, so constituent membership is likely indicative in predicting the focus status of an individual word. Also, we have mentioned in our discussion of possible evidence for focus at the beginning of this chapter, the topological field a word appears in (Höhle, 1986) is potentially relevant for its focus status.

Cheung & Penn (2009) present a parsing model that demonstrates good performance in determining both topological fields and phrase structure for German. The model is trained on the TüBa-D/Z treebank (Telljohann et al., 2004), whose rich syntactic model encodes topological fields as nodes in the syntax tree itself. Following Cheung & Penn (2009), we trained an updated version of their model using the current version of the Berkeley Parser (Petrov & Klein, 2007) and release 10 of the TüBa-D/Z[8].

Before calculating any features, we wanted to investigate the distribution of constituent types across foci, in order to gain insight on the connection between question types and constituent types of focus spans. We parsed our training set with the model described above and, for each focus span, determined the constituent that most closely matches its boundaries. For the 7,642 focus spans in the data set, this produced 5,373 exact constituent matches, the rest are near-matches and a small number where no constituent could be found, likely due to parser errors. We then grouped the obtained constituents by surface question form of the corresponding question (see section 7.3.2). Table7.2 shows the result of this analysis, listing for each question form the three most frequent constituent types[9] of focus spans, and the number of focus spans in total.

[8] http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora/tueba-dz.html

[9] For a list of the constituent types in the TüBa-D/Z scheme, see appendix B.

Question form	3 most frequent constituent types	Instances
what	NX (54%), SIMPX (15%), MF (8%)	1901
several	NX (29%), SIMPX (17%), ADJX (14%)	1768
which	NX (72%), SIMPX (7%), ADJX (5%)	1536
why	SIMPX (48%), NX (19%), MF (17%)	1024
how	NX (38%), SIMPX (17%), FKONJ (11%)	885
where	NX (47%), PX (45%), ADJX (7%)	175
who	NX (86%), SIMPX (4%), ADJX (3%)	151
unknown	NX (66%), SIMPX (15%), MF (8%)	91
when	PX (62%), NX (21%), MF (10%)	52
alternative	NX (43%), ADJX (23%), VXINF (17%)	35
yes/no	SIMPX (67%), DM (17%), VROOT (13%)	24
All	NX (45%), SIMPX (18%), MF (8%)	7642

Table 7.2.: Constituent types of focus spans grouped by question form

One can see that some patterns emerge, for example the most frequent constituent type for *why*-questions is SIMPX (simplex clause) and the most frequent for *when*-questions is PX (prepositional phrase). This is in line with our intuition that answers to *why*-questions are most often full sentences or clauses, and answers to *when*-questions are time expressions frequently realized as prepositional phrases. Also, the second most frequent category for *where*-questions is also PX, demonstrating that locations also tend to be realized in a similar manner syntactically. In most cases, noun phrases (NX) are the predominant category, and they are also the most frequent category overall, as shown in the bottom line of Table 7.2.

Based on the new parsing model and the investigation above, we integrated two new features into our focus detection model: the **direct parent constituent node of a word** and **the nearest topological field node of a word**. The impact of these features raises accuracy on *focus* slightly from 84.7% to 84.8% and on *background* quite noticeably from 68.0% to 68.7%, resulting in an overall accuracy increase from 77.7% to 78.1%.

Beyond Question Forms

Looking beyond the present problem, we also wanted to find out how close the relationship between the manually annotated Answer Types (as part of the expert annotation scheme described in section 6.2) and automatically determined constituent types is. Recall that an Answer Type encodes the relation between the question and the focus instance by giving a label to the alternative set, such as `Location`. In the annotation scheme it is used as a semantic scaffold for reliable focus marking. There is thus a direct correspondence between foci and Answer Types. We did an analysis analoguous to the one for question forms above, only this time grouping the same constituent types of focus spans by the respective Answer Type of the span instead. Table 7.3 lists the result of this analysis.

Answer type	3 most frequent constituent types	Instances
Reason	SIMPX (50%), MF (19%), NX (16%)	1452
Abstract_Entity	NX (83%), MF (4%), PX (3%)	1343
Action	NX (28%), FKONJ (18%), SIMPX (17%)	1318
Property	ADJX (36%), NX (26%), FKONJ (13%)	819
Thing	NX (93%), ADJX (3%), MF (2%)	769
Location	NX (64%), PX (31%), ADJX (3%)	463
State	SIMPX (49%), MF (17%), NX (14%)	403
Living_Being	NX (87%), R (6%), PX (3%)	344
Yes_No	DM (55%), SIMPX (25%), N/A (10%)	200
Quantity_Duration	NX (62%), ADJX (18%), N/A (7%)	168
Time_Date	PX (61%), NX (23%), MF (6%)	145
Report	SIMPX (44%), FKONJ (14%), MF (12%)	140
Manner	PX (56%), ADJX (12%), SIMPX (10%)	78
All	NX (45%), SIMPX (18%), MF (8%)	7642

Table 7.3.: Constituent types of focus spans grouped by Answer Type

Table 7.3 reveals that Answer Types separate the same set of focus spans in a much more useful way. First, the distribution of the foci is more even across Answer Types, as evidenced by the fact that the median value of instances per type is only 175 for question forms, but 403 for Answer Types. Second, the most frequent constituent types differ much more between Answer Types than was the case for question forms, as becomes evident when comparing the top percentages: where in Table 7.2 most question forms have NX (noun phrase) in the first place and SIMPX (simplex clause) in the second, the picture is far more diverse in Table 7.3. For example, the most frequent constituent type for

Property is ADJX (adjectival phrase), which is exactly how one would expect properties to be realized syntactically. Third, the differences between the most frequent and the second most frequent constituent type are greater in Table 7.3 than they were in Table 7.2 (42.5% vs. 37.5% on average), indicating a clearer separation of data.

We conclude that there is a strong relationship between the semantic Answer Types and the syntactic constituent types. While we currently cannot exploit this relationship because we cannot detect Answer Types automatically, this is clearly an avenue that future research should pursue: automatically determining the Answer Type given the question should help significantly in defining the alternative set, and in turn narrow down the possibilities of syntactic realization for foci, facilitating the task of identifying focus ins answers.

7.7.3. Exploring the Use of Crowd-sourced Data

In this section, we explore the usefulness of crowd-sourced data in training a focus detection classifier. At the end of chapter 6, we presented an experiment on turning focus annotation into a crowd-sourcing task, motivated by the relatively low cost and annotation time requirements of crowd-sourcing. We concluded that while crowd-sourced focus annotation is generally of lower quality than expert annotation, the crowd does produce reliable annotation for some types of data. In order to predict the quality of crowd annotation, we defined a measure called Consensus Cost (see equation 6.1), and showed that it has a strong relationship with annotation quality.

We first used the crowd-annotated CREG-5K described in section 6.4. In an attempt to filter out cases where workers had trouble interpreting the answer or did not bother to carry out the annotation task properly, we selected only those answers from the annotated data for which the majority of crowd workers did not use the "question not answered" button. The resulting data set consists of 37,639 words, which we used in a first focus detection experiment, in order to compare the resulting performance to the classifier trained on the smaller, but expert-annotated training set we have used so far. Using the same feature set and evaluation method (logistic regression in 10-fold cross-validation), we obtain an accuracy of 74.6% (*focus*: 74.9%, *background*: 74.5%), which is a significantly lower result than the 78.1% we reported above for the expert-based classifier.

In a second attempt using the CREG-5K data, we filtered the data set using a Consensus Cost cutoff: all annotated answers for which the Consensus Cost was higher than 0.75 were discarded, following our analysis of Consensus Cost versus annotation quality in Figure 6.6, where a Consensus Cost above 0.75 resulted in a very heterogeneous annotation quality. Classification accuracy using the thus filtered data, which now consisted of 31,111 instances, improves to 76.7% (*focus*: 76.8%, *background*: 76.6%), a noticeable improvement.

To expand our training data base beyond what we have used so far, we selected all questions and corresponding answers from the bigger (and noisier) CREG-23K that do not occur in CREG-5K. The intention behind this step is to enable the classifier to better generalize towards answers to previously unseen questions. The resulting data set, hereafter called CREG-CrowdQuestions , was annotated by the crowd using exactly the setup described in section 6.4. We applied the same filtering steps as for CREG-5K above, considering only answers where the majority had not used the "question not answered" button and where Consensus Cost was not above 0.75. The resulting data set has 27,385 instances and leads to an accuracy of 77.0% (*focus*: 82.4%, *background*: 72.1%), the highest crowd-based focus detection result we obtained so far.

In a final step, we combined our expert-based training data with the CREG-CrowdQuestions data just described in an attempt to get the best out of both annotation worlds. This combined data set has 54,365 instances, and the accuracy we obtained is 76.5% (*focus*: 83.4%, *background*: 68.8%). The fact that this result is lower than using either data set on its own shows that the notion of focus targeted by experts and the crowd is still slightly different, resulting in a somewhat heterogeneous training base. Nevertheless, it is still possible that this model will perform well in extrinsic evaluation within SAA, a question to which we will return in chapter 8.

7.8. Final Results

Table 7.4 provides an overview of the incremental improvements we presented in the previous section for the expert-based training data. We list again the three baselines for reference, and the model from section 7.4 ("initial model").

	Accuracy for		
Feature set	*focus*	*background*	both
Majority baseline	100%	0%	58.1%
Givenness baseline	81.5%	42.5%	65.1%
POS baseline	89.2%	39.6%	68.4%
Initial model (section 7.5)	**85.2%**	66.7%	77.4%
Above + distrib. Givenness	84.7%	68.0%	77.7%
Above + constituency	84.8%	**68.7%**	**78.1%**

Table 7.4.: Final focus detection performance

While the improvements may seem modest quantitatively, they show that the added features are well-motivated and do make an impact. Overall, it is especially apparent that the key to better performance is reducing the number of false positives in this data set: whereas the accuracy for focus stays roughly the same, the one for background improves steadily with each feature set addition.

Summary

In this chapter, we presented our focus detection approach, building on the annotation work we presented in chapter 6.

We first gave an overview of existing work in focus detection, pointing out that it exclusively targets English and is mostly based on gold-standard annotation of spoken language including prosodic information, instead of automatic annotation on written language. We then discussed what are possible observable properties of questions and answers that one can automatically derive for focus detection in written language. Following that, we described the resulting feature set of our initial classification approach, which includes a combination of syntactic, positional and givenness features based on both the question and the answer. We also discussed classification-related problems such as algorithm and training data base in the context of focus detection.

We then presented quantitative and qualitative results. Quantitatively, our approach shows robust performance, beating all baselines significantly. We show the impact of each feature set and investigate some typical classifier errors using specific examples.

Complementing the success of the initial model, we launched into a description of three extensions of our approach: distributional givenness, constituency-based features and crowd-sourced training data. For the former two, we showed that they further increase the accuracy of our model, which is the basis for our extrinsic evaluation in chapter 8, where automatic focus detection is to be used within SAA.

8. Focus in Short Answer Assessment

In this chapter, we investigate the impact of each of the three sources of focus annotation that we have now established within the task of SAA: expert annotation, crowd annotation, and automatic annotation. In general, these also correspond to different levels of quality: crowd annotation is generally superior to automatic annotation, and expert annotation is generally superior to both crowd and automatic annotation.

We are interested in extrinsic evaluation for two reasons. First, it has been pointed out that evaluating annotation of a theoretical linguistic notion only intrinsically is problematic because there is no non-theoretical grounding involved (Riezler, 2014), so extrinsic evaluation is necessary to validate the annotation approach. And second, in this thesis we are ultimately interested in advancing SAA, which means that the final evaluation of our contribution must happen in the SAA context.

In order to perform the evaluation, we first recall what we want to accomplish by employing using focus annotation in section 8.1, illustrating the potential benefit of focus information in SAA with specific examples (partly based on Ziai & Meurers 2014). We then take stock of what our toolbox looks like in section 8.2, reviewing in detail how CoMiC aligns and classifies answers. In section 8.3, we discuss possible ways of incorporating focus into CoMiC and decide on using one of them, namely using focus as a filter for alignment. Finally, in section 8.4, we present the evaluation of each of the three sources of annotation separately[1].

[1] Sections 8.4.1 and 8.4.2 are based on De Kuthy, Ziai & Meurers (2016a) and De Kuthy, Ziai & Meurers (2016b), respectively. A version of Section 8.4.3 is published as part of Ziai & Meurers (2018).

8.1. The Goal: What We Want

The possible benefits of using focus to constrain alignment can take different forms: focus can lead us to exclude extra, irrelevant material, but it can also uncover the fact that the relevant piece of information has in fact not been included, as in the following corpus example:

(51) Q: Was machen sie, um die Brunnen im Winter zu schützen?

'What do they do to protect the wells in winter?'

TA: Zwölf der 47 Brunnen werden im Winter aus Schutz vor dem Frost und Witterungsschäden ⟦eingehaust⟧$_F$

'Twelve of the 47 wells are ⟦encased⟧$_F$ in winter for protection from freezing and damage from weather conditions'

SA: im Winter gibt es Frost und Witterungsschäden

'in winter there is freezing and damage from weather conditions'

The question asks what is being done to protect the wells in winter, for which the text states that twelve of wells are encased for protection (technically, this is an answer to a sub-question since nothing is asserted about the other wells). Additional new information such as *vor dem Frost und Witterungsschäden* does not distinguish between alternatives to the question *Was machen sie. . . ?*, which clearly asks for an `Action`. The target and student answer have high token overlap due to the presence of such extra information, but only the target answer contains the relevant focus "eingehaust". Without the focus filter, an SAA system would likely classify this answer as correct, but with the added focus information, it has the means to judge this answer adequately.

Another illustrative corpus example is the one we already saw in Figure 6.1, here repeated as (52). Recognizing the meaning equivalence between *Sie geht gerne* ('she likes to go') and *macht sie Spaß* ('is fun for her') is a non-trivial task for computational approaches, as semantic relatedness across part-of-speech classes is involved here. However, once *joggen* ('jogging') has been identified as the focus, the comparison would be simplified to an ordinary string match.

(52) Q: Welchen Sport macht Isabel?

'Which Sport does Isabel do?'

TA: Sie geht gerne ⟦joggen⟧$_F$.

'She likes to go ⟦jogging⟧$_F$'.

SA: ⟦Joggen⟧$_F$ macht sie Spaß.

'⟦Jogging⟧$_F$ is fun for her.'

In order to decide how to use focus to constrain alignment so that we get results such as the ones above, let us revisit how CoMiC aligns and classifies answers in the next section.

8.2. The Toolbox: The Alignment Process Revisited

In chapter 5, we established that CoMiC is an alignment-based system which operates in three stages:

1. Annotate linguistic units (words, chunks and dependencies) in student and target answer on various levels of abstraction.
2. Find alignments of linguistic units between student and target answer based on annotation.
3. Classify the student answer based on number and type of alignments, using a supervised machine learning setup.

Stage 1 is responsible for enriching the original input with linguistic information. Besides the components already present in the system (see section 5.4), this is where focus detection would need to be added in order for stages 2 and 3 to make use of it.

In stage 2, CoMiC computes alignments between parts of student and target answer. In stage 2, CoMiC integrates a surface approach to givenness, flagging all tokens thus identified as not to be aligned. This step happens before any alignments are computed, so it effectively constrains the number of possible alignments. For all remaining non-punctuation tokens, all possible alignments are then calculated on all levels, resulting in a graph. The Traditional Marriage Algorithm (TMA, Gale & Shapley 1962) is used to obtain one global alignment configuration where each student answer token is aligned to at most one target answer token.

In stage 3, CoMiC extracts features based on how many alignments of which type are found, relative to the number of all alignable tokens in either student or target answer. Givenness marking constrains the values of these features: all alignment counts now refer to the number of non-given words in student and target answer, not to the number of total words. In other words, this implicitly introduces a "givenness filter" into alignment and feature extraction.

8.3. The Solution: Focus/Background as Alignment Filter

The question we now need to address is how to make use of the focus information we obtained in the task of SAA in a way that is compatible with the alignment and feature extraction approach we reviewed in the previous section.

In principle, there are multiple ways focus could be encoded: for example, focus could be an **explicit additional feature** of answers, indicating the presence or absence of a focused expression and possibly its position in the answer or the words that are part of it. On the plus side, this variant is easy to implement, as it does not interfere with any existing features and integrates easily. However, we then have to rely on the machine learning algorithm to pick up interactions between existing features and the focus features, because the relationship between alignments and focus is not explicitly encoded.

Another way would be to encode focus **implicitly as filtering criterion** for other features, essentially making the system only look at focused parts of the answer to be classified. This approach is analogous to the one already implemented in the CoMiC system for surface givenness. This approach is still fairly straightforward to implement, as a blueprint for it exists in the form of the givenness filter. Moreover, while no explicit focus feature is present, focus is implicitly encoded here in that only focused tokens are considered for alignment, disregarding all background tokens. On the negative side, the commitment to focus here is very strong: what is not focused cannot possibly play a role in meaning comparison.

Finally, a third way would be to use focus as a sort of **weighting criterion** for other features or alignments. This would introduce a direct interaction of focus with existing features, allowing for fine-grained prioritization of linguistic units according to information-structural status. Also, it would allow to take a notion of confidence about the focus status into account, making it possible to lessen the commitment to focus when the quality of focus annotation is doubtful. This approach is however far more complex to implement, as it presents a number of open issues: on what basis would one turn focus into weights? Does such an approach involve some sort of empirically determined threshold and if yes, how can one avoid having to tune it to each data set?

For the present work, in the spirit of both givenness and focus as IS distinctions expressing different perspectives on an utterance (as outlined in chapter 3), we therefore decided to treat the focus/background distinction the same way as the given/new one: instead of looking only at non-given words, alignment and feature extraction can take only focused words into account. Consequently, we transferred the underlying method to the notion of focus and implemented a component that excludes all non-focused words from alignment, resulting in alignments between focused parts of answers only.

8.4. The Results: Externally Evaluating Focus Annotation

In this section, we quantitatively investigate the impact of focus annotation on SAA, using focus as a filter as described in the previous section. We do so separately for expert, crowd and automatic annotation, since in each case the research question and testing setup are slightly different. Because of these differences in setup, which also relate to training set sizes and available annotated data, the results between the three evaluations are only roughly comparable. However, we always report a baseline for individual results to be compared against.

8.4.1. Expert Annotation

For the evaluation of expert annotation, we experimented with three different settings involving the basic givenness filter and our focus annotations: i) using the givenness filter by itself as a baseline, ii) aligning only focused tokens as described above and iii) combining both by producing a givenness and a focus version of each classification feature.

Table 8.1 summarizes the quantitative results for the data sets annotated in the two phases of annotation we described in section 6.3. The figures were obtained using leave-one-out testing with TiMBL (Daelemans et al., 2007), using the *k*-nearest-neighbor algorithm with default settings and several distance measures, as in earlier CoMiC versions (see section 5.4).

In all cases, the results show that focus beats the basic givenness baseline on its own, pushing the classification accuracy substantially from 85.9% to 88.0% in the case of CREG-1032, and from 82.1% to 83.8% in the case of CREG-2155.

	Basic givenness	Focus	Combined
CREG-1032	85.9%	88.0%	88.6%
CREG-2155	82.1%	83.8%	85.1%
CREG-ExpertFocus	83.2%	84.6%	85.6%

Table 8.1.: Answer classification accuracy with CoMiC

While this is an encouraging result already, the combination of basic givenness and focus performs even better, improving slightly to 88.6% accuracy for CREG-1032, and more substantially to 85.1% in the case of CREG-2155.

In terms of the conceptual notions of formal pragmatics, this is an interesting result. While the notion of givenness implemented here is surface-based and mechanistic and thus could be improved, the results support the idea that both of the commonly discussed dimensions, focus/background and new/given, are useful and informative information-structural dimensions that complement each other in assessing the meaning of answers.

One can see that generally, CREG-1032 is an easier testbed for CoMiC than the bigger CREG-2155, which is likely due to the lower complexity of the reading texts we pointed out in section 4.2.3. Nevertheless, the improvement provided by focus annotation is stable across all different data sets. The improvements are also all statistically significant (established using McNemar's test with $\alpha = 0.05$).

Overall, the extrinsic evaluation of expert focus annotation demonstrates the practical relevance of information-structural notions in computational linguistic applications such as SAA.

8.4.2. Crowd Annotation

In externally establishing the relevance and quality of the crowd focus annotation, our goal is twofold: on the one hand, we want to find out whether the previously introduced Consensus Cost measure (see equation 6.1) is helpful in determining the quality of focus annotation as measured by its impact on SAA. On the other hand, it is interesting to determine whether the state of the art in automatic answer assessment can be advanced by integrating non-expert annotation of focus (as a step towards automatic focus annotation developed using the crowd-annotated data).

To cleanly separate the data used for testing CoMiC from the data used for training, we used the train/test split of CREG-5K that we already discussed in section 5.5, which splits the data approximately 80% to 20%.

In exploring the impact of different Consensus Costs, we used the same four cutoffs as for the annotation evaluation in section 6.4: 0.25, 0.5, 0.75 and the maximum value 1.0. For each cutoff, we picked the answers with crowd focus annotations satisfying the cutoff constraint in training and test set, and ran CoMiC on the resulting data excerpt, aligning only words in student and target answer that are focused. For the rest of the data, which did not meet the Consensus Cost criterion or for which no focus annotation was available, we used the standard version of CoMiC that only aligns words not previously mentioned in the question. We then calculated a weighted average (by number of test instances) of both system accuracies in order to arrive at an overall system result for the respective Consensus Cost value. The results are displayed in Table 8.2, obtained again using leave-one-out testing and TiMBL, as described in the previous subsection.

Cost	Focus		Given		Avg
$\leq$	train/test	%	train/test	%	%
base	–		4136/1001	81.5	81.5
0.25	1009/252	88.1	3127/749	80.4	82.3
0.5	2019/489	84.5	2117/512	80.7	82.5
0.75	3087/747	84.5	1049/254	79.5	**83.2**
1.0	3638/882	82.7	498/119	76.5	81.9

Table 8.2.: Results on the "unseen answers" test set

The 'train/test' column shows the number of training and test instances each system was run on, and the '%' column shows the classification accuracy achieved. The 'base' row gives the baseline resulting from using CoMiC as-is, without any focus information.

Looking at the results for the focus partition of the data, one can see that accuracy drops when taking into account focus annotation with higher Consensus Cost, even though thereby in principle more training data is becoming available.

For the 'Given' column, when data with higher Consensus Cost is used for the 'Focus' version of the system and thereby less data is available for training the 'Given' system, accuracy of the latter decreases.

Overall, a Consensus Cost cutoff of 0.75 gives the optimal trade-off between both system variants, yielding 83.2% classification accuracy.

Test with answers to unseen questions

In a second experiment, we also performed a question-based evaluation, meaning that for approximately 20% of randomly picked questions in CREG-5K, all answers were held out as the test set. As we explained in section 2.3, this is a much harder benchmark since the system in the test has to classify answers to previously unseen questions, providing some indication of the system's ability to learn something general rather than about specific question-answer pairs. The remainder of the testing procedure was the same as described above, yielding the results detailed in Table 8.3.

Cost	Focus		Given		Avg
≤	train/test	%	train/test	%	%
base	–		4016/1121	78.8	78.8
0.25	970/291	81.4	3046/830	78.2	79.0
0.5	1938/570	80.4	2078/551	78.2	79.3
0.75	2973/861	81.6	1043/260	76.9	**80.6**
1.0	3515/1005	79.6	501/116	78.4	79.5

Table 8.3.: Results on the "unseen questions" test set

The accuracies are generally lower due to the harder test scenario. Moreover, the clear trends observed above with regard to training and test size do not seem to apply as clearly here, likely again owing to the 'unseen questions' scenario. Given the many different types of potential questions and the relatively small number of different questions the system sees during training, it is more important for which questions the system has seen answers, than how many. However, despite the differences to the previous experiment, the optimal result is again achieved with a Consensus Cost of 0.75, supporting the conclusion that Consensus Cost supports a systematic characterization of annotation quality.

8.4.3. Automatic Annotation

For automatic annotation, we again use the same CREG-5K test sets as for evaluating the crowd annotation before, one based on the 'unseen answers' and one based on the 'unseen questions' test scenario. However, as already described in section 7.4, the training set differs: in order to arrive at a fair and generalizable testing setup, we removed all answers from the CREG-5K training set that occur also in CREG-ExpertFocus, the data set used to train the focus detection classifier in chapter 7. This means that neither the focus classifier nor CoMiC have seen any of the test set answers before.

The resulting smaller training set contains 1,606 student answers, while the test sets contain 1,002 (unseen answers) and 1,121 (unseen questions), respectively.

Table 8.4 summarizes the results for the different CoMiC variants and test sets in terms of accuracy. 'Basic givenness' again refers to the standard CoMiC system, 'Focus' to the version that only aligns focused tokens, and 'Combined' refers to the system that uses both feature versions. In addition to the two test sets introduced above, we tested the systems on the training set using 10-fold cross validation and the WEKA implementation of k-nearest-neighbor with $k = 5$.

Test set	Instances	Basic givenness	Focus	Combined
10-fold CV	1606	**83.19%**	80.95%	82.25%
Unseen answers	1002	**80.64%**	78.74%	80.54%
Unseen questions	1121	77.43%	77.34%	**78.41%**

Table 8.4.: CoMiC results using givenness and expert-based focus features

One can see that in general, the focus classifier seems to introduce too much noise to positively impact classification results. The standard CoMiC system outperforms the focus and the combined version for the cross validation case and the 'unseen answers' set. This is in contrast to the results we reported in section 8.4.1 using manual focus information, where the combined system significantly outperforms all other variants. This shows that while focus information is clearly useful in SAA, it needs to be reliable enough to be of actual benefit. Recall also that the way we use focus information in CoMiC implies a strong commitment: only focused words are aligned and included in feature extraction, which does not produce the desired result if the focus information is not accurate. A possible way of remedying this situation would be to use focus as an extra feature or less strict modifier of existing features, as we outlined in section 8.3. There is thus room for improvement both in the automatic detection of focus and its use in extrinsic tasks.

However, one result stands out encouragingly: in the 'unseen questions' case, the focus system almost reaches the performance of standard CoMiC, and the combined version tops both these systems by approximately 1%. This shows that even automatically determined information structural properties provide benefits when more concrete information, in the form of previously seen answers to the same questions, is not available. Our classifier thus successfully transfers general knowledge about focus to new question material. As an interesting side remark, the 'unseen questions' test set is also the only one to have a slight majority towards incorrect answers, so focus apparently was able to "correct" the bias of the training set here.

In conclusion, while the manner of employing focus in SAA and the quality of automatic annotation can still be improved, we can already report one positive result.

Explorations Using Crowd-Sourced Data

In section 7.7.3, we described and intrinsically evaluated several focus detection models we built using crowd-sourced annotation. In the intrinsic evaluation, some of these models came close to the performance of the model based on expert annotation using the same feature set.

In this section, we now want to explore how these models fare when plugged into the same extrinsic test that the expert-based model faced above. Two models were tested: one based on the crowd annotation experiment where only non-CREG-5K questions and answers were used (CREG-CrowdQuestions)[2], and one based on a combination of this crowd experiment and the expert data (CREG-CrowdQuestions-Expert). The results are shown in Table 8.5.

Test set	Instances	Basic givenness	Focus	Combined
Focus model based on CREG-CrowdQuestions				
10-fold CV	1606	**83.19%**	78.83%	82.12%
Unseen answers	1002	**80.64%**	76.85%	79.94%
Unseen questions	1121	**77.43%**	75.73%	76.98%
Focus model based on CREG-CrowdQuestions-Expert				
10-fold CV	1606	**83.19%**	81.26%	82.75%
Unseen answers	1002	**80.64%**	79.54%	80.04%
Unseen questions	1121	**77.43%**	76.27%	77.34%

Table 8.5.: CoMiC results using givenness and crowd-based focus features

The accuracies obtained show that none of the system variants including focus information beat the givenness baseline. Also, the performance of the CREG-CrowdQuestions model is lower than the one of the CREG-CrowdQuestions-Expert model, demonstrating again the generally lower quality of the crowd annotation. However, when comparing these results with the ones in Table 8.4, an interesting result emerges: the CREG-CrowdQuestions-Expert model beats the expert-only model from Table 8.4 in the cross-validation and 'unseen answers' settings, demonstrating that the additional crowd-annotated answers seen by the focus classifier during training do have the potential to make its classification more robust.

In sum, while the explorations here are quantitatively not beneficial, they are qualitatively interesting since they enable cross-model comparisons that reveal the impact of training data.

[2] We here use exactly the same filtering criteria for crowd annotations as described in section 7.7.3.

Summary

We presented an integration of focus into SAA based on the CoMiC system. Starting out by reminding ourselves what we expect from incorporating focus in SAA by concrete examples, we then took stock of how the alignment and classification process in CoMiC works, and discussed several possibilities of how focus could be included in the system. Having decided on using focus as a filter for alignment, we present a thorough quantitative evaluation of all three sources of focus information we have: expert annotation, crowd annotation, and automatic annotation.

Results show that focus is clearly beneficial in SAA, as demonstrated when using expert annotation to constrain alignment. A combination of the givenness- and focus-based features in CoMiC yields the best performance and significantly beats standard CoMiC on several test sets.

For crowd annotation, we could also show that focus positively impacts SAA results, beating the standard CoMiC baseline. Moreover, we demonstrated the usefulness of our measure for predicting the quality of crowd annotation, Consensus Cost, in selecting answers whose focus annotation should be used for extrinsic evaluation.

For automatic annotation, the picture was more diverse: while our focus detection classifier currently introduces too much noise to be of quantitative use in general, we did obtain a positive result in the test case of 'unseen questions' where the classifier successfully transfers general knowledge about focus to new question material.

We conclude that focus is indeed beneficial in SAA as demonstrated especially by the results based on expert annotation. However, both the performance of the detection model and the way focus information is currently incorporated leave room for improvement, and need to be addressed in future work.

Part IV.

Conclusion

9. Summary and Outlook

In this final chapter, we first review the main content of the thesis presented on a part-by-part basis in section 9.1. Based on this summary, we present our contributions in section 9.2, before finally pointing out future directions of this work in section 9.3.

9.1. Summary

In this thesis, we investigated the role of information-structural distinctions, specifically focus, in a concrete computational linguistic task, Short Answer Assessment (SAA). The overall aim was two-fold: in a computational linguistic research strand, we argued that information-structural notions such as focus are helpful and beneficial for SAA, and we showed this to be the case both quantitatively and qualitatively. Complementing this, in a more linguistic research strand we aimed to advance the state of the art in IS research by developing a focus annotation approach that is based on authentic data from the SAA setting, which enables i) an operationalization of meaning-based criteria discussed in the theoretical IS literature and ii) an independent external evaluation criterion for focus in the form of its impact in SAA.

Below, we provide a more detailed overview of the content presented in this thesis.

Part I: Background

In part I, we laid out the scene for our work in terms of the two fields it mainly draws on: Short Answer Assessment (SAA) in computational linguistics and Information Structure (IS) in theoretical linguistics.

In chapter 2, we gave an introduction and an overview of the field of Short Answer Assessment (SAA). We first defined the task, which is to classify an answer to a question in terms of whether it answers the question or not, with respect to a context and usually a reference answer. The main challenge of SAA is the variation in form and content that occurs in answers, i.e., the different well-formed and ill-formed possibilities in the language system to answer a question. In our overview of SAA approaches, we focused on those that make some use of the task context, concluding that most make rather peripheral use of questions and none but our own and related approaches (Bailey & Meurers, 2008; Hahn & Meurers, 2012) make the connection to IS explicit. We also briefly characterized the few publicly available data sets, all of which are English.

Complementing the computational background, in chapter 3, we gave an introduction into the general idea of information structure, which is to organize units of information (or meaning) so that they fit into the discourse. We characterized the three most important distinctions within IS: topic vs. comment, which deals with the entities information is organized around, given vs. new, which categorizes information according to how accessible it is in the discourse, and focus vs. background, which separates an utterance into a part that answers a current (implicit) question and one that does not. Having defined these notions, we reviewed how they are annotated in corpus data by several existing approaches, noting that topic and focus seem to be somewhat harder to annotate than givenness. Equipped with both theoretical and practical knowledge of IS notions, we asked ourselves which notion is most relevant for SAA, and after ruling out topic/comment, we discussed what given/new and focus/background can contribute to meaning assessment in the context of concrete examples, concluding that focus represents most accurately the part of the answer we are interested in. Returning to the annotation of focus, we characterized the two main issues that focus annotation approaches are faced with: (1) determining relevant alternatives in the context, so focus can be pinpointed, and (2) determining the extent of the focus, i.e., its borders. We noted that problem (1) can be alleviated by using a data basis including more explicit context, such as questions.

Part II: The Empirical Basis and Experimental Sandbox

In part II, we described the foundation this thesis builds on: the Corpus of Reading Comprehension Exercises in German (CREG), our empirical basis and the Comparing Meaning in Context (CoMiC) system, our SAA approach. The corpus and the system were essential for carrying out our subsequent research, and we made substantial contributions to both, as demonstrated by co-authorship in several peer-reviewed publications (Meurers, Ott & Ziai, 2010; Meurers, Ziai, Ott & Bailey, 2011a; Meurers, Ziai, Ott & Kopp, 2011b; Ott, Ziai & Meurers, 2012).

In chapter 4, we presented the empirical basis for both our analysis of focus in answers and for our SAA experiments. We first described some desirable characteristics for an authentic data source for SAA: an explicit, linguistically encoded task context and questions that require free-text answers, so that form variation occurs and can be studied. Having decided that reading comprehension exercises fit our needs, we proceeded to describing CREG itself. We started with the collection process which included the development of WELCOME, a web-based corpus collection tool that allows distributed entry of richly structured reading comprehension data by non-technical users, allowing the incremental creation of reading comprehension corpora. Turning from the process to the end result, we characterized the corpus structure, which includes reading texts, questions, target answers, student answers and two meaning assessments for each student answer. Based on this discussion of CREG's components, we described several subsets of CREG that were created for evaluation or annotation purposes. The most important of these are the balanced CREG-1032 and CREG-5K sets, and the CREG-2155 set sampled for focus annotation.

In chapter 5, we presented the Comparing Meaning in Context (CoMiC) SAA system, which forms the basis for the SAA experiments reported in 8. CoMiC is conceptually based on the English CAM (Bailey & Meurers, 2008) and aligns answers to pre-specified target answers on several linguistic abstraction levels, from surface forms to semantic relatedness and synonymy. It then extracts features in the form of summary statistics on the number and type of alignments found, and uses these features to classify an answer with regard to content. Going well beyond CAM, we described how CoMiC was designed to support parallel annotation layers, flexible marking of linguistic units (such as focus and givenness marking), and straightforward extensibility with new components, using the UIMA architecture (Ferrucci & Lally, 2004). Having replicated the performance of CAM for a small English data set, we then turned to the transfer of the system to German. We showed that CoMiC achieves robust performance in the $> 80\%$ range on balanced data sets such as CREG-1032 and CREG-5K. Further demonstrating CoMiC's adaptability, we briefly discussed applications of the system to tasks such as QA (Rudzewitz & Ziai, 2015; Rudzewitz, 2016b) and plagiarism detection (Rudzewitz, 2016a).

Part III: Focus: Internal and External Relevance

Part III is the main part of this thesis, where we describe manual focus annotation, automatic focus detection, and extrinsic evaluation of focus in the SAA context.

In chapter 6, we presented our work on manual focus annotation, both by experts and by crowd workers. We first discussed possible sources of evidence for focus annotation in corpus data, pointing out that surface criteria are not sufficient and settling on meaning-based criteria instead. With this view in mind, we described our iterative annotation scheme, which operationalizes focus by making use of explicit questions: annotators first determine surface question forms and the alternative set, before identifying instances of focus in answers. In order to pin down the extent of the focus, we use a meaning-based substitution test with which the focus status of an individual surface word can be determined.

Having defined the annotation scheme, we described our expert annotation experiment, where two student annotators were trained to identify focus in student and target answers of CREG using the brat annotation tool. The experiment was carried out in two phases, the first using the CREG-1032 and the second using the CREG-2155 subset and updated annotation guidelines resulting from insights of the first phrase. The agreement across both annotation phases computed to $\kappa = .7$ on 4,177 student and target answers, which is very competitive with regard to the state of the art, where results reported were generally lower (cf., e.g., Ritz et al. 2008; Calhoun et al. 2010). Having done the agreement study, we trained a third annotator to adjudicate the results of both phases, resulting in our merged gold standard CREG-ExpertFocus.

In order to investigate whether focus can be annotated reliably by non-experts, we conducted a crowd-sourcing experiment, showing that crowd workers do provide reliable focus annotation for some types of data, which is especially apparent in an analysis according to question types: in our experiment, crowd workers reach near-expert level for *who-*, *when-* and *where*-questions, produce acceptable results for *which-*, *what-* and *how*-questions and perform poorly on *why*-questions. We also showed how the quality of crowd annotation may be predicted independently of a comparison with experts, using Consensus Cost, an agreement-based measure we defined.

Equipped with a high-quality focus-annotated data set, we presented our focus detection approach in chapter 7. We first reviewed existing work in focus detection, pointing out that it exclusively targets English and is mostly based on gold-standard annotation of spoken language including prosodic information, instead of automatic annotation on written language. Taking a step back, we asked ourselves which relevant observable properties of questions and answers one can automatically derive for focus detection in written language. Besides discussing a range of lexical, syntactic, positional and Givenness features that we experimented with, we elaborated on classification-related issues such as algorithm selection and training/testing partitioning of our data in the context of focus detection.

Having settled on a feature set, we trained a logistic regression model and evaluated it using 10-fold cross-validation on our gold standard data. Quantitatively, our approach performs robustly, reaching 77.4% classification accuracy in a data set with 26,980 words and beating several baselines by a large margin (majority: 58.1%, givenness: 65.1%, POS: 68.4%). A performance analysis by feature sets reveals that the biggest incremental gains come from givenness and positional features, when added to a model already including syntactic question and answer features. Complementing the quantitative evaluation, we qualitatively investigate several characteristic examples of detection behavior, indicating that focus detection could benefit both from top-down well-formedness constraints, and more accurate givenness recognition.

In an attempt to address some of these issues, we presented three extensions to the initial model: a distributional approach to givenness, the incorporation of constituency and sentence topology, and the use of crowd-annotated data for focus detection. For the former two, we showed that they further increase the accuracy of our model, pushing it to 78.1%.

In chapter 8, we finally presented the integration of focus into SAA based on the CoMiC system. Starting out by reminding ourselves what we expect from incorporating focus in SAA using concrete examples, we then took stock of how the alignment and classification process in CoMiC works, and discussed several possibilities of how focus could be included in the system. Having decided on using focus as a filter for alignment, we present a thorough quantitative evaluation of all three sources of focus information we have: expert annotation, crowd annotation, and automatic annotation.

Results show that focus is clearly beneficial in SAA, as demonstrated when using expert annotation to constrain alignment. A combination of the givenness- and focus-based features in CoMiC yields 85.6% on CREG-ExpertFocus in leave-one-out testing, significantly beating standard CoMiC (83.2%).

For crowd annotation, we could also show that focus positively impacts SAA results. Using an 80/20 train/test split of CREG-5K and the Consensus Cost measure for crowd annotation quality we defined in chapter 6, we explored the impact of different Consensus Cost cutoffs in training and testing CoMiC. Both the results on the 'unseen answers' test set and the ones on the 'unseen questions' test set show the best accuracy for a Consensus Cost ≤ 0.75, beating the baseline 83.2% to 81.5% in the former case and 80.6% to 78.8% in the latter case in leave-one-out testing.

For automatic annotation, the picture was more diverse: while our focus detection classifier currently introduces too much noise to be of quantitative use in general, we did obtain a positive result in the test case of 'unseen questions' where the classifier successfully transfers general knowledge about focus to new question material, beating the standard CoMiC baseline 78.4% to 77.4% in 10-fold cross-validation.

We conclude that focus is indeed beneficial in SAA as demonstrated especially by the results based on expert annotation. However, both the performance of the detection model and the way focus information is currently used leave room for improvement, and need to be addressed in future work.

9.2. Contributions

Based on the summary in the previous section, we now list all contributions that are part of this thesis. They are structured into three areas: research results, software, and resources.

9.2.1. Research Results

We start by listing the research results of this thesis in terms of the fields we have contributed to: Information Structure and Short Answer Assessment.

Information Structure

Annotation and analysis of focus We have developed a new approach to annotating and analyzing focus in authentic data. The approach builds on current meaning-based views of focus and operationalizes them in an incremental annotation scheme, resulting in substantial inter-annotator agreement on a data set involving non-wellformed language.

Crowd annotation of focus We demonstrated that focus annotation is feasible using non-experts, i.e., ordinary speakers of a language, establishing the quality of crowd annotation both by comparing it to our expert annotation and by independently predicting it using Consensus Cost, a new measure we defined.

Automatic focus detection Building on our successful meaning-based focus annotation work, we have developed the first automatic focus detection approach for German. It combines a range of linguistically well-motivated features based on both questions and answers, including syntactic cues, positional properties and surface question forms. We showed that our approach outperforms several baselines by a large margin.

Short Answer Assessment

Impact of information-structural properties we have shown that focus can be integrated into alignment-based SAA systems as a filter and performed an extrinsic evaluation of manual and automatic annotation within the CoMiC system, revealing that both manually and automatically determined focus have the potential to result in quantitative gains.

9.2.2. Software

The research presented in this thesis has led to two major software outcomes: the corpus collection tool WELCOME and the SAA system CoMiC.

WELCOME

We presented the WELCOME system to which we significantly contributed, a web-based application which enables distributed data entry for the purpose of creating richly structured reading comprehension corpora, and show how this system was used to collection a large German corpus. It is the only system of its kind that we know of, and its architecture has already influenced and facilitated the design of at least one new project, FeedBook, where the goal is to develop an interactive workbook to support individualized instruction for 7th grade English language learners[1].

WELCOME was the result of a team effort also involving Niels Ott, Georgi Boychev and Detmar Meurers.

CoMiC

We presented CoMiC, an alignment-based SAA system to which we significantly contributed, which achieves state-of-the-art performance for both English and German. Its modular and extensible architecture, which builds on the industry-grade Unstructured Information Management Architecture (UIMA), supports the implementation of innovative linguistically motivated components resulting in new insights, as is the case with the incorporation of focus annotation and detection in this thesis. Others have also extended or adapted the approach to be applicable to other tasks or include new evidence (Horbach et al., 2013; Rudzewitz & Ziai, 2015; Rudzewitz, 2015, 2016a).

Like WELCOME, CoMiC was a team effort, including contributions by Niels Ott, Björn Rudzewitz and Detmar Meurers.

[1] `http://purl.org/feedbook`

9.2.3. Resources

Finally, we present the resources created as part of this thesis: the Corpus of Reading Comprehension Exercises in German (CREG) and the part of it that is annotated with focus.

CREG

CREG is the largest reading comprehension corpus publicly available, and one of the few data sets publicly available to SAA researchers at all (cf., e.g., Burrows et al. 2015). It contains more than 35,000 student answers and 1,600 target answers to over 1,500 questions on about 150 reading texts. While in SAA the interest is usually in the answers produced by the students and the task they perform, CREG also contains multiple meta-data records for each of the students who produced answers, enabling researchers to model student's development over time.

CREG was created in collaboration with Niels Ott and Detmar Meurers (Ott et al., 2012).

Focus-annotated CREG

The focus-annnotated version of CREG is our main contribution in terms of resources. The expert-annotated portion CREG-ExpertFocus comprises 4,177 student and target answers annotated with focus by two annotators. The full annotation also includes surface questions forms and semantic Answer Types for each focus instance. Additionally, there is a gold standard version where a third annotator served as judge to merge the two annotation versions into one definite version.

Complementing CREG-ExpertFocus, we created a crowd-annotated version of the CREG-5K corpus. Besides CREG-5K, which includes over 5,500 annotated student and target answers, we also ran a second experiment to cover more questions not part of CREG-5K. This second data set is called CREG-CrowdQuestions and contains over 3,300 student and target answers.

CREG-ExpertFocus was created in collaboration with Detmar Meurers, and the crowd-annotation data sets were created in collaboration with Kordula De Kuthy and Detmar Meurers (Ziai & Meurers, 2014; De Kuthy et al., 2016a).

9.3. Outlook

In this final section, we discuss what directions emerge from our work for future research. Parallel to the chapters in part III of this thesis, future research can be divided into three broad areas, which we outline below.

9.3.1. Focus Annotation

In the area of focus annotation, there is still work left to be done besides the areas we already discussed in section 6.3.7. In general, for both theory validation and training computational approaches, more high-quality annotated data would be very valuable. While such further focus annotation efforts could take the shape of annotating more SAA data, theoretical IS research would benefit more from extending the annotation approach to data sources with more natural discourses, such as dialogue or interview corpora. Such data sources pose the problem that they typically have few explicit questions on which our focus annotation scheme crucially depends.

However, recent research (Riester, Brunetti & De Kuthy, in press) has shown that it is possible to reliably determine QUDs and make them explicit in discourse annotation, which would in turn enable reliable focus annotation. Following this research strand, the approach presented in this thesis could be scaled up beyond explicit question-answer pairs: De Kuthy, Reiter & Riester (2018) spell out an explicit analysis of text in terms of QUDs and show that it is possible to annotate explicit QUDs with high inter-annotator agreement. Combined with an automated approach to question generation, it could thus be possible to recover implicit QUDs from text and subsequently apply our current approach to any text, based on an independently established, general formal pragmatic analysis.

9.3.2. Focus Detection

Concerning focus detection, there are at least two routes that should be followed in the future.

First, while linguistically well-motivated, the feature set we have explored in this thesis is limited and could be extended. For example, more complex syntactic features are possible, such as the role a word plays in its parent constituent.

Moreover, features modeling language well-formedness in answers could be integrated, since ungrammatical answers are less likely to produce a usable interpretation and hence a valid focus instance. In modeling alternatives, one could explore the automatic detection of Answer Types, as already mentioned at the end of section 7.7.2: since Answer Types determine the syntactic category of focus much more reliably than surface questions forms, they could potentially be very helpful in detecting focus. Such an approach should then build a bridge to Answer Typing in QA literature (cf., e.g., Li & Roth 2002; Pinchak & Lin 2006), where automatically determining the type of factoid questions has been investigated.

Second, our approach so far does not leverage the power of sequence classification approaches for focus detection, instead classifying each word on its own. While we have shown robust performance using a word-based focus classifier and almost all related approaches also follow this route, more recent research (Zang et al., 2014) has shown that it is possible to exploit contextual properties by using CRF models, even though it seems that a fair amount of experimentation with feature templates is required to benefit from sequence classification, as our own brief explorations were unsuccessful in improving classification performance.

9.3.3. Focus in Computational Linguistic Tasks

The use of focus information in computational linguistic applications is an even less-researched area than focus detection itself, so there is ample opportunity for improvement, as we outline below.

As far as focus in SAA is concerned, we have already noted that quantifiable benefits are constrained by the way focus information can currently be used and integrated: focus as a filter for alignment implies a strong commitment to the focus information, which backfires when the information is less than perfect. A possible solution would be to introduce a smaller commitment, such as the use of a weighting approach for alignments along the lines of Rudzewitz (2015). However, besides the method of integrating focus, it is also possible that the benefit is constrained by the data: if new information is identical to focused information in most cases, focus offers little benefit over recognizing new (or non-given) material.

Going forward from classification to diagnosis in real-life applications making use of SAA, such as Intelligent Language Tutoring Systems, the Answer Types mentioned above provide the added benefit that besides an overall classification of responses, one can give more fine-grained feedback: instead of telling the learner that the response is incorrect, a feedback message informing them about e.g. a missing `Location` would be within range of the system.

Finally, given the reliable training basis and the robust detection approach we have provided in this work, we envisage the use of focus in other computational linguistic tasks as well. For instance, QA and query-focused summarization are meaning-centered tasks where an information requirement is often explicitly formulated in natural language, making them primary candidates for exploring the benefit of incorporating focus. Given that they target extraction of information from potentially much longer texts than the short answers we have tackled, the possible benefit of detecting focus given an explicit question may in fact be even greater than in SAA.

Bibliography

Abney, Steven. 1997. Partial Parsing via Finite-State Cascades. *Natural Language Engineering* 2. 337–344. URL http://purl.org/net/Abney-97.pdf.

Aldabe, Itziar, Montse Maritxalar & Oier Lopez de Lacalle. 2013. EHU-ALM: Similarity-Feature Based Approach for Student Response Analysis. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 580–584. Atlanta, Georgia, USA: Association for Computational Linguistics. URL http://aclweb.org/anthology/S13-2097.

Amaral, Luiz, Detmar Meurers & Ramon Ziai. 2011. Analyzing Learner Language: Towards A Flexible NLP Architecture for Intelligent Language Tutors. *Computer-Assisted Language Learning* 24(1). 1–16. URL http://purl.org/dm/papers/amaral-meurers-ziai-10.html.

Antworth, Evan L. 1993. Glossing Text with the PC-KIMMO Morphological Parser. *Computers and the Humanities* 26. 475–484. URL http://www.springerlink.com/content/r20w66k70976ur91/fulltext.pdf.

Artstein, Ron & Massimo Poesio. 2008. Survey Article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4). 555–596. URL http://aclweb.org/anthology/J08-4004.pdf.

Artzi, Yoav, Tom Kwiatkowski & Jonathan Berant (eds.). 2014. *Proceedings of the ACL 2014 Workshop on Semantic Parsing*. Baltimore, MD: Association for Computational Linguistics. URL http://aclweb.org/anthology/W14-2400.

Atkinson, Kevin. 2004. Spell Checking Oriented Word Lists (SCOWL). Web resource. URL http://wordlist.sourceforge.net/.

Bachman, Lyle, Nathan Carr, Greg Kamei, Mikyung Kim, Michael Pan, Chris Salvador & Yasuyo Sawaki. 2002. A Reliable Approach to Automatic Assessment of Short Answer Free Responses. In *Proceedings of the*

19th International Conference on Computational Linguistics (COLING 2002), 1–4. URL http://portal.acm.org/ft_gateway.cfm?id=1071907&type=pdf&coll=GUIDE&dl=GUIDE&CFID=47019848&CFTOKEN=16425462.

Badino, Leonardo & Robert A. J. Clark. 2008. Automatic labeling of contrastive word pairs from spontaneous spoken English. In *Proceedings of the 2008 IEEE Spoken Language Technology Workshop*, 101–104. doi: 10.1109/SLT.2008.4777850.

Bailey, Stacey. 2008. *Content assessment in intelligent computer-aided language learning: Meaning error diagnosis for English as a second language*: The Ohio State University dissertation. URL http://purl.org/net/Bailey-08.pdf.

Bailey, Stacey & Detmar Meurers. 2008. Diagnosing meaning errors in short answers to reading comprehension questions. In Joel Tetreault, Jill Burstein & Rachele De Felice (eds.), *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*, 107–115. Columbus, Ohio. URL http://aclweb.org/anthology/W08-0913.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Journal of Language Resources and Evaluation* 3(43). 209–226. URL http://wacky.sslmit.unibo.it/lib/exe/fetch.php?media=papers:wacky_2008.pdf.

Baumann, Stefan, Caren Brinckmann, Silvia Hansen-Schirra, Geert-Jan Kruijff, Ivana Kruijff-Korbayova, Stella Neumann, Erich Steiner, Elke Teich & Hans Uszkoreit. 2004a. The MULI Project. Annotation and Analysis of Information Structure in German and English. In *Proceedings of LREC 2004*, Lisbon. URL www.coli.uni-saarland.de/~cabr/papers/muli.LREC2004main.pdf.

Baumann, Stefan, Caren Brinckmann, Silvia Hansen-Schirra, Geert-Jan Kruijff, Ivana Kruijff-Korbayová, Stella Neumann & Elke Teich. 2004b. Multi-Dimensional Annotation of Linguistic Corpora for Investigating Information Structures. In *Proceedings of NAACL/HLT 2004 Conference Workshop Frontiers in Corpus Annotation*, Boston, MA. URL http://aclweb.org/anthology/W/W04/W04-2707.pdf.

Baumann, Stefan & Arndt Riester. 2010. Annotating Information Status in Spontaneous Speech. In *Proceedings of the Fifth International Conference on Speech Prosody*, Chicago. URL http://www.ims.uni-stuttgart.de/~arndt/doc/baumannRiesterSpeechPros2010.

Baumann, Stefan & Arndt Riester. 2012. Referential and Lexical Givenness: Semantic, Prosodic and Cognitive Aspects. In Gorka Elordieta & Pilar Prieto (eds.), *Prosody and Meaning*, vol. 25 Interface Explorations, Berlin: Mouton de Gruyter. URL http://www.ims.uni-stuttgart.de/~arndt/doc/baumannRiesterBarcelonaPrefinal.pdf.

Becker, M. & A. Frank. 2002. A stochastic topological parser for German. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, 71–77. Kluwer.

Bethard, Steven, Philip Ogren & Lee Becker. 2014. ClearTK 2.0: Design Patterns for Machine Learning in UIMA. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 3289–3293. Reykjavik, Iceland: European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/218_Paper.pdf.

Bicici, Ergun & Josef van Genabith. 2013. CNGL: Grading Student Answers by Acts of Translation. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 585–591. Atlanta, Georgia, USA: Association for Computational Linguistics. URL http://aclweb.org/anthology/S13-2098.

Bird, Steven & Mark Liberman. 2000. A Formal Framework for Linguistic Annotation. *Speech Communication* 33(1-2). 23–60.

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3. 993–1022.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius & George Smith. 2002. The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol. URL http://bultreebank.s481.sureserver.com/proceedings/paper03.pdf.

Brockett, Chris & William B. Dolan. 2005. Support Vector Machines for Paraphrase Identification and Corpus Construction. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 1–8. URL http://aclweb.org/anthology/I05-5001.

Büring, Daniel. 2003. On D-trees, beans, and B-accents. *Linguistics and Philosophy* 26(5). 511–545.

Büring, Daniel. 2007. Intonation, Semantics and Information Structure. In Gillian Ramchand & Charles Reiss (eds.), *The Oxford Handbook of Linguistic Interfaces*, Oxford University Press. URL http://semanticsarchive.net/Archive/GQ0YjgxM/buring.information.structure.v2005.pdf.

Burrows, Steven, Iryna Gurevych & Benno Stein. 2015. The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education* 25(1). 60–117.

Calhoun, Sasha. 2007. Predicting Focus through Prominence Structure. In *Proceedings of Interspeech*, Antwerp, Belgium. URL http://www.cstr.inf.ed.ac.uk/downloads/publications/2007/calhounIS07.pdf.

Calhoun, Sasha, Jean Carletta, Jason Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman & David Beaver. 2010. The NXT-format Switchboard Corpus: A Rich Resource for Investigating the Syntax, Semantics, Pragmatics and Prosody of Dialogue. *Language Resources and Evaluation* 44. 387–419. URL http://link.springer.com/article/10.1007%2Fs10579-010-9120-1.

Calhoun, Sasha, Malvina Nissim, Mark Steedman & Jason Brenier. 2005. A Framework for Annotating Information Structure in Discourse. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, 45–52. Ann Arbor, Michigan: Association for Computational Linguistics. URL http://aclweb.org/anthology/W/W05/W05-0307.

Carreras, Xavier & Lluis Marquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *CoNLL '05: Proceedings of the Ninth Conference on Computational Natural Language Learning*, URL http://www.lsi.upc.edu/~srlconll/st05/papers/intro.pdf.

de Castilho, Richard Eckart & Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT)*, 1–11. Dublin, Ireland.

Chafe, Wallace. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Charles N. Li (ed.), *Subject and topic*, 27–55. New York: Academic Press.

Chandrasekar, R., Christine Doran & B. Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, 1041–1044.

Charniak, Eugene. 1996. Tree-bank grammars. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 1031–1036.

Cheung, Jackie Chi Kit & Gerald Penn. 2009. Topological field parsing of German. In *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, 64–72. Morristown, NJ, USA: Association for Computational Linguistics. URL `http://aclweb.org/anthology/P09-1008`.

Chiarcos, Christian, Ines Fiedler, Mira Grubic, Andreas Haida, Katharina Hartmann, Julia Ritz, Anne Schwarz, Amir Zeldes & Malte Zimmermann. 2009. Information Structure in African Languages: Corpora and Tools. In *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages – AfLaT 2009*, 17 – 24. Athens, Greece. URL `http://aflat.org/files/W09-0703.pdf`.

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20(1). 37–46.

Collins, Michael John. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, 184–191. Morristown, NJ, USA: Association for Computational Linguistics. doi: 10.3115/981863.981888. URL `http://portal.acm.org/citation.cfm?id=981863.981888`.

Cook, Philippa & Felix Bildhauer. 2013. Identifying “aboutness topics”: two annotation experiments. *Dialogue and Discourse* 4(2). 118–141.

Cox, David R. 1958. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)* 20(2). 215–242.

Culicover, Peter & Michael Rochemont. 1983. Stress and focus in English. *Language* 59. 122–165.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot & Antal van den Bosch. 2007. *Timbl: Tilburg memory-based learner reference guide, ilk technical report ilk 07-03*. Induction of Linguistic Knowledge Research Group Department of Communication and Information Sciences, Tilburg University Tilburg, The Netherlands. URL `http://ilk.uvt.nl/downloads/pub/papers/ilk.0703.pdf`. Version 6.0.

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

Dagan, Ido, Oren Glickman & Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In Joaquin Quiñonero Candela, Ido Dagan, Bernardo Magnini & Florence d’Alché Buc (eds.), *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment*, vol. 3944 Lecture Notes in Computer Science, 177–190. Springer. URL `http://u.cs.biu.ac.il/~dagan/publications/RTEChallenge.pdf`.

Daume III, Hal. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 256–263. Prague, Czech Republic: Association for Computational Linguistics. URL `http://aclweb.org/anthology/P07-1033`.

Daumé III, Hal & Daniel Marcu. 2006. Bayesian Query-Focused Summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 305–312. Association for Computational Linguistics. URL `http://aclweb.org/anthology/P06-1039`.

De Kuthy, Kordula, Nils Reiter & Arndt Riester. 2018. QUD-Based Annotation of Discourse Structure and Information Structure: Tool and Evaluation. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, JP.

De Kuthy, Kordula, Ramon Ziai & Detmar Meurers. 2016a. Focus Annotation of Task-Based Data: a comparison of expert and crowd-sourced annotation in a reading comprehension corpus. In *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference (LREC)*, 3928–3934. Portorož, Slovenia. URL `http://www.lrec-conf.org/proceedings/lrec2016/pdf/1083_Paper.pdf`.

De Kuthy, Kordula, Ramon Ziai & Detmar Meurers. 2016b. Focus Annotation of Task-Based Data: Establishing the Quality of Crowd Annotation. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, 110–119. Berlin, Germany: ACL. URL `http://aclweb.org/anthology/W16-1713.pdf`.

Dima, Corina. 2015. Reverse-engineering Language: A Study on the Semantic Compositionality of German Compounds. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1637–1642. Lisbon, Portugal: Association for Computational Linguistics. URL `http://aclweb.org/anthology/D15-1188`.

Dzikovska, Myroslava, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan & Hoa Trang Dang. 2013. SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 263–274. Atlanta, Georgia, USA: Association for Computational Linguistics. URL `http://aclweb.org/anthology/S13-2045`.

Dzikovska, Myroslava O., Amy Isard, Peter Bell, Johanna D. Moore, Natalie B. Steinhauser, Gwendolyn E. Campbell, Leanne S. Taylor, Simon Caine & Charlie Scott. 2011. Adaptive Intelligent Tutorial Dialogue in the BEETLE II System. In *Proceedings of the 15th International Conference on Artificial Intelligence* *in Education* AIED, 621–621. Berlin, Heidelberg: Springer. URL http://dl.acm.org/citation.cfm?id=2026506.2026633.

Dzikovska, Myroslava O., Rodney D. Nielsen & Chris Brew. 2012. Towards Effective Tutorial Feedback for Explanation Questions: A Dataset and Baselines. In *Proceedings of Human Language Technologies: The 2012 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2012)*, Montréal, Canada: Association for Computational Linguistics.

Eason, Sarah H., Lindsay F. Goldberg, Katherine M. Young, Megan C. Geist & laurie E. Cutting. 2012. Reader-text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology* 104(3). 515–528.

Eckart, Kerstin, Arndt Riester & Katrin Schweitzer. 2012. *A discourse information radio news database for linguistic analysis* 65–76. Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-28249-2_7. URL http://dx.doi.org/10.1007/978-3-642-28249-2_7.

Eckert, Miriam & Michael Strube. 2000. Dialogue Acts, Synchronizing Units, and Anaphora Resolution. *Journal of Semantics* 17(1). 51–89. doi: 10.1093/jos/17.1.51. URL http://jos.oxfordjournals.org/content/17/1/51.abstract.

Ferrucci, David & Adam Lally. 2004. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering* 10(3–4). 327–348. URL http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=252253&fulltextType=RA&fileId=S1351324904003523.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5). 378–382.

Foth, Kilian. 2006. Eine umfassende Constraint-Dependenz-Grammatik des Deutschen. Tech. rep. Universität Hamburg.

Foth, Kilian A., Arne Köhn, Niels Beuck & Wolfgang Menzel. 2014. Because Size Does Matter: The Hamburg Dependency Treebank. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2326–2333. Reykjavik, Iceland: European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/860_Paper.pdf`.

Gabrilovich, Evgeniy & Shaul Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 6–12. URL `http://www.bradblock.com/Computing_Semantic_Relatedness_using_Wikipedia_based_Explicit_Semantic_Analysis.pdf`.

Gale, David & Lloyd S. Shapley. 1962. College Admissions and the Stability of Marriage. *American Mathematical Monthly* 69. 9–15. URL `http://www.econ.ucsb.edu/~tedb/Courses/Ec100C/galeshapley.pdf`.

Götz, Thilo & Oliver Suhre. 2004. Design and implementation of the UIMA Common Analysis System. *IBM Systems Journal* 43(3). 476–489. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.5824&rep=rep1&type=pdf`.

Götze, Michael, Thomas Weskott, Cornelia Endriss, Ines Fiedler, Stefan Hinterwimmer, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas & Ruben Stoel. 2007. Information structure. In Stefanie Dipper, Michael Götze & Stavros Skopeteas (eds.), *Information Structure in Cross-Linguistic Corpora*, vol. 7 Interdisciplinary Studies on Information Structure, 147–187. Universitätsverlag Potsdam.

Grice, Martine & Stefan Baumann. 2002. Deutsche Intonation und GToBI. *Linguistische Berichte* 1991(191). 267–298.

Hahn, Michael & Detmar Meurers. 2012. Evaluating the Meaning of Answers to Reading Comprehension Questions: A Semantics-Based Approach. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational*

Applications (BEA-7) at NAACL-HLT 2012, 94–103. Montreal. URL http://purl.org/dm/papers/hahn-meurers-12.html.

Hahn, Michael & Detmar Meurers. 2013. On Deriving Semantic Representations from Dependencies: A Practical Approach for Evaluating Meaning in Learner Corpora. In Kim Gerdes, Eva Hajičová & Leo Wanner (eds.), *Computational Dependency Theory*, 62–77. Amsterdam: IOS Press. URL http://purl.org/dm/papers/hahn-meurers-13.html. Revised version of paper presented at DEPLING 2011.

Hajič, Jan, Eva Hajičová, Jaroslava Hlaváčová, Václav Klimeš, Jiří Mírovský, Petr Pajas, Jan Štěpánek, Barbora Vidová Hladká & Zdeněk Žabokrtský. 2006. *The prague dependency treebank 2.0*. URL http://ufal.mff.cuni.cz/pdt2.0/.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann & Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. In *The SIGKDD Explorations*, vol. 11, 10–18.

Halliday, Michael. 1967. Notes on Transitivity and Theme in English. Part 1 and 2. *Journal of Linguistics* 3. 37–81, 199–244.

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

Hantke, Simone, Erik Marchi & Björn Schuller. 2016. Introducing the Weighted Trustability Evaluator for Crowdsourcing Exemplified by Speaker Likability Classification. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France: European Language Resources Association (ELRA).

Hearst, Marti A., Susan T. Dumais, Edgar Osman, John Platt & Bernhard Schölkopf. 1998. Support Vector Machines. *IEEE Intelligent Systems and their Applications* 13(4). 18–28.

Heilman, Michael & Nitin Madnani. 2013. ETS: Domain Adaptation and Stacking for Short Answer Scoring. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 275–279. Atlanta, Georgia, USA: Association for Computational Linguistics. URL `http://aclweb.org/anthology/S13-2046`.

Hempelmann, Christian F., David Dufty, Philip M. McCarthy, Arthur C. Graesser, Zhiqiang Cai & Danielle S. McNamara. 2005. Using LSA to Automatically Identify Givenness and Newness of Noun Phrases in Written Discourse. In B. G. Bara, L. Barsalou & M. Bucciarelli (eds.), *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, 941–949. Stresa, Italy: Erlbaum. doi: 10.1.1.116.5716.

Hockett, Charles. 1958. *A course in modern linguistics*. New York: McMillan.

Höhle, Tilman N. 1986. Der Begriff 'Mittelfeld'. Anmerkungen über die Theorie der topologischen Felder. In A. Schöne (ed.), *Kontroversen alte und neue. Akten des VII. Internationalen Germanistenkongresses Göttingen 1985*, 329–340. Tübingen: Niemeyer. Bd. 3.

Horbach, Andrea, Alexis Palmer & Manfred Pinkal. 2013. Using the text to evaluate short answers for reading comprehension exercises. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, 286–295. Atlanta, Georgia, USA: Association for Computational Linguistics. URL `http://aclweb.org/anthology/S13-1041`.

Hsueh, Pei-Yun, Prem Melville & Vikas Sindhwani. 2009. Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing* HLT '09, 27–35. Stroudsburg, PA, USA: Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1564131.1564137`.

Iacobacci, Ignacio, Mohammad Taher Pilehvar & Roberto Navigli. 2015. SensEmbed: learning sense embeddings for word and relational similarity. In *Proceedings of ACL*, 95–105.

Jacobs, Joachim. 2001. The dimensions of topic-comment. *Linguistics* 39. 641–681. URL http://www.degruyter.com/dg/viewarticle.fullcontentlink:pdfeventlink/contentUri?format=INT&t:ac=j$002fling.2001.39.issue-4$002fling.2001.027$002fling.2001.027.xml.

Jimenez, Sergio, Claudia Becerra & Alexander Gelbukh. 2013. SOFTCARDINALITY: Hierarchical Text Overlap for Student Response Analysis. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 280–284. Atlanta, Georgia, USA: Association for Computational Linguistics. URL http://aclweb.org/anthology/S13-2047.

Kim, Sung-Suk. 1998. Time-delay recurrent neural network for temporal correlations and prediction. *Neurocomputing* 20(1–3). 253–263. doi: 10.1016/S0925-2312(98)00018-6.

Klein, Dan & Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL 2003)*, 423–430. Sapporo, Japan. URL http://aclweb.org/anthology/P03-1054.

Kohavi, Ron. 1995. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2* IJCAI'95, 1137–1143. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Köper, Maximilian, Christian Scheible & Sabine Schulte im Walde. 2015. Proceedings of the 11th International Conference on Computational Semantics, 40–45. Association for Computational Linguistics. URL http://aclweb.org/anthology/W15-0105.

Kordon, Klaus. 2006. *Ein Trümmersommer*. Beltz & Gelberg.

Krifka, Manfred. 1992. A Compositional Semantics for Multiple Focus Constructions. In Joachim Jacobs (ed.), *Informationsstruktur und Grammatik*, 17–54. Opladen: Westdeutscher Verlag.

Krifka, Manfred. 2001. For a structured meaning account of questions and answers. In C. Fery & W. Sternefeld (eds.), *Audiatur Vox Sapientia. A Festschrift* *for Arnim von Stechow*, vol. 52 studia grammatica, 287–319. Berlin: Akademie Verlag.

Krifka, Manfred. 2007. Basic Notions of Information Structure. In Caroline Fery, Gisbert Fanselow & Manfred Krifka (eds.), *The notions of information structure*, vol. 6 Interdisciplinary Studies on Information Structure (ISIS), 13–55. Potsdam: Universitätsverlag Potsdam. URL http://opus.kobv.de/ubp/volltexte/2008/1960/.

Krifka, Manfred & Renate Musan. 2012. Information structure: overview and linguistic issues. In Manfred Krifka & Renate Musan (eds.), *The Expression of Information Structure*, vol. 5 The Expression of Cognitive Categories, 1–43. Berlin/Boston: De Gruyter Mouton.

Krippendorff, Klaus. 1980. *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage Publications.

Kruijff-Korbayová, Ivana & Geert-Jan M. Kruijff. 2004. Discourse-level Annotation for Investigating Information Structure. In Bonnie Webber & Donna K. Byron (eds.), *ACL 2004 Workshop on Discourse Annotation*, 41–48. Barcelona, Spain: Association for Computational Linguistics. URL http://aclweb.org/anthology-new/W/W04/W04-0206.pdf.

Kupietz, Marc, Cyril Belica, Holger Keibel & Andreas Witt. 2010. The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta: European Language Resources Association (ELRA).

Lafferty, John, Andrew McCallum & Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, 282–289.

Lambrecht, Knud. 1994. *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents* Cambridge Studies in Linguistics. Cambridge University Press.

Landauer, Thomas, Peter Foltz & Darrell Laham. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes* 25. 259–284. URL `http://lsa.colorado.edu/papers/dp1.LSAintro.pdf`.

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

Lesk, Michael. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, 24–26. Toronto, Ontario, Canada. URL `http://portal.acm.org/citation.cfm?id=318728`.

Levenshtein, Vladimir I. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady* 10(8). 707–710. URL `http://www.mendeley.com/research/binary-codes-capable-of-correcting-insertions-and-reversals/`.

Li, Xin & Dan Roth. 2002. Learning Question Classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, 1–7. Taipei, Taiwan. URL `http://aclweb.org/anthology/C02-1150`.

Liu, Hugo. 2004. MontyLingua: An End-to-End Natural Language Processor with Common Sense. Software Website. Media Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts. URL `http://web.media.mit.edu/~hugo/montylingua/`.

Lüdeling, Anke, Maik Walter, Emil Kroymann & Peter Adolphs. 2005. Multilevel error annotation in learner corpora. In *Proceedings of Corpus Linguistics*, Birmingham. URL `http://www.corpus.bham.ac.uk/PCLC/Falko-CL2006.doc`.

Lüdeling, Anke, Julia Ritz, Manfred Stede & Amir Zeldes. 2016. Corpus Linguistics and Information Structure Research. In Caroline Féry & Shinichiro Ishihara (eds.), *The Oxford Handbook of Information Structure*, Oxford University Press. URL `http://www.oxfordhandbooks.com/10.1093/oxfordhb/9780199642670.001.0001/oxfordhb-9780199642670-e-013`.

Makatchev, Maxim & Kurt VanLehn. 2007. Combining Baysian Networks and Formal Reasoning for Semantic Classification of Student Utterances.

In *Proceedings of the International Conference on AI in Education (AIED)*, Los Angeles.

Marcus, M., Beatrice Santorini & M. A. Marcinkiewicz. 1993a. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2). 313–330. URL `ftp://ftp.cis.upenn.edu/pub/treebank/doc/cl93.ps.gz`.

Marcus, Mitchell, Kim Grace, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Vies, Mark Ferguson, Karen Katz & Britta Schasberger. 1994. The Penn Treebank: annotating predicate argument structure. In *Proceedings of the workshop on Human Language Technology (HLT '94)*, 114–119. Plainsboro, NJ: Morgan Kaufmann Publishers Inc. doi: http://dx.doi.org/10.3115/1075812.1075835. URL `http://citeseer.ist.psu.edu/marcus94penn.html`.

Marcus, Mitchell P., Mary Ann Marcinkiewicz & Beatrice Santorini. 1993b. Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics* 19(2). 273–290. URL `http://aclweb.org/anthology/J93-2004`.

McCarthy, P. M., V. Rus, S. S. Crossley, A. C. Graesser & D. S. McNamara. 2008. Assessing forward-, reverse-, and average-entailment indeces on natural language input from the intelligent tutoring system, iSTART. In D. Wilson & G. Sutcliffe (eds.), *Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference*, 165–170.

Meurers, Detmar. 2005. On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German. *Lingua* 115(11). 1619–1639. URL `http://purl.org/dm/papers/meurers-03.html`.

Meurers, Detmar & Markus Dickinson. 2017. Evidence and Interpretation in Language Learning Research: Opportunities for Collaboration with Computational Linguistics. *Language Learning* 67(2). `http://dx.doi.org/10.1111/lang.12233`.

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

Meurers, Detmar, Ramon Ziai, Niels Ott & Stacey Bailey. 2011a. Integrating Parallel Analysis Modules to Evaluate the Meaning of Answers to Reading Comprehension Questions. *IJCEELL. Special Issue on Automatic Free-text Evaluation* 21(4). 355–369. URL `http://www.inderscience.com/info/inarticle.php?artid=42793`.

Meurers, Detmar, Ramon Ziai, Niels Ott & Janina Kopp. 2011b. Evaluating Answers to Reading Comprehension Questions in Context: Results for German and the Role of Information Structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, 1–9. Edinburgh. URL `http://aclweb.org/anthology/W11-2401.pdf`.

Mihalcea, Rada, Courtney Corley & Carlo Strapparava. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of the National Conference on Artificial Intelligence*, vol. 21(1), 775–780. Menlo Park, CA: American Association for Artificial Intelligence (AAAI) Press. URL `http://www.cse.unt.edu/~rada/papers/mihalcea.aaai06.pdf`.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Miller, George. 1995. WordNet: a lexical database for English. *Communications of the ACM* 38(11). 39–41. URL `http://aclweb.org/anthology/H94-1111`.

Miller, Tristan, Darina Benikova & Sallam Abualhaija. 2015. GermEval 2015: LexSub – A Shared Task for German-language Lexical Substitution. In *Proceedings of GermEval 2015: LexSub*, 1–9. URL `https://sites.google.com/site/germeval2015/program/2015_GermEval_LexSub.pdf?attredirects=0&d=1`.

Minnen, Guido, John Carroll & Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering* 7(3). 207–233.

Mitchell, Tom, Nicola Aldrige & Peter Broomhead. 2003. Computerized Marking of Short-Answer Free-Text Responses. Paper presented at the 29th annual conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

Mohler, Michael, Razvan Bunescu & Rada Mihalcea. 2011. Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 752–762. Portland, Oregon, USA: Association for Computational Linguistics. URL `http://aclweb.org/anthology/P11-1076`.

Müller, Christoph & Michael Strube. 2006. Multi-Level Annotation of Linguistic Data with MMAX2. In Sabine Braun, Kurt Kohn & Joybrato Mukherjee (eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, vol. 3 English Corpus Linguistics, 197–214. Frankfurt: Peter Lang.

Màrquez, Lluís, James Glass, Walid Magdy, Alessandro Moschitti, Preslav Nakov & Bilal Randeree. 2015. SemEval-2015 Task 3: Answer Selection in Community Question Answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Association for Computational Linguistics.

Nenkova, Ani & Dan Jurafsky. 2007. Automatic detection of contrastive elements in spontaneous speech. In *Proceedings of the 2007 IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, 201–206. doi: 10.1109/ASRU.2007.4430109.

Nielsen, Rodney D., Wayne Ward & James H. Martin. 2009. Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering* 15(4). 479–501. doi: 10.1017/S135132490999012X.

Nielsen, Rodney D., Wayne Ward, James H. Martin & Martha Palmer. 2008. Annotating Students' Understanding of Science Concepts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC '08)*, 3441–3448. European Language Resources Association (ELRA).

Nissim, Malvina, Shipra Dingare, Jean Carletta & Mark Steedman. 2004. An annotation scheme for information status in dialogue. In *Proceedings of the 4th Conference on Language Resources and Evaluation*, Lisbon, Portugal. URL `http://www.lrec-conf.org/proceedings/lrec2004/pdf/638.pdf`.

Nivre, Joakim, Jens Nilsson, Johan Hall, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov & Erwin Marsi. 2007. MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. *Natural Language Engineering* 13(1). 1–41. URL http://w3.msi.vxu.se/~nivre/papers/nle07.pdf.

Och, Franz Josef. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, URL http://acl.ldc.upenn.edu/acl2003/main/pdfs/Och.pdf.

Ogren, Philip & Steven Bethard. 2009. Building Test Suites for UIMA Components. In *Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing (SETQA-NLP 2009)*, 1–4. Boulder, Colorado: Association for Computational Linguistics. URL http://aclweb.org/anthology/W/W09/W09-1501.

Ott, Niels, Ramon Ziai, Michael Hahn & Detmar Meurers. 2013. CoMeT: Integrating different levels of linguistic modeling for meaning assessment. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval)*, 608–616. Atlanta, GA: ACL. URL http://aclweb.org/anthology/S13-2102.pdf.

Ott, Niels, Ramon Ziai & Detmar Meurers. 2012. Creation and Analysis of a Reading Comprehension Exercise Corpus: Towards Evaluating Meaning in Context. In Thomas Schmidt & Kai Wörner (eds.), *Multilingual Corpora and Multilingual Corpus Analysis* Hamburg Studies in Multilingualism (HSM), 47–69. Amsterdam: Benjamins. URL https://benjamins.com/#catalog/books/hsm.14.05ott.

Pado, Ulrike & Cornelia Kiefer. 2015. Short Answer Grading: When Sorting Helps and When it Doesn't. In *Proceedings of the 4th workshop on NLP for Computer Assisted Language Learning at NODALIDA*, 43.

Paggio, Patrizia. 2006. Annotating Information Structure in a Corpus of Spoken Danish. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, 1606 – 1609. Genoa, Italy. URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/639_pdf.pdf.

Pearson, Karl. 1895. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London* 58(347-352). 240–242. doi: 10.1098/rspl.1895.0041. URL `http://rspl.royalsocietypublishing.org/content/58/347-352/240.short`.

Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Association for Computational Linguistics. URL `http://aclweb.org/anthology/D14-1162`.

Pérez, Diana, Enrique Alfonseca, Pilar Rodríguez, Alfio Gliozzo, Carlo Strapparava & Bernardo Magnini. 2005. About the effects of combining Latent Semantic Analysis with natural language processing techniques for free-text assessment. *Revista signos* 38(59). 325–343.

Perez-Marin, Diana & Ismael Pascual-Nieto. 2011. Willow: a system to automatically assess students' free-text answers by using a combination of shallow NLP techniques. *International Journal of Continuing Engineering Education and Life-Long Learning* 21(2/3). 155–169. doi: 10.1504/IJCEELL.2011.040196.

Petrov, Slav & Dan Klein. 2007. Improved Inference for Unlexicalized Parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 404–411. Rochester, New York.

Pinchak, Christopher & Dekang Lin. 2006. A Probabilistic Answer Type Model. In *Proceedings of the 11th Conference of the European Chapter of the Association of Computational Lingustics (EACL)*, 393–400.

Pradhan, Sameer S., Wayne H. Ward, Kadri Hacioglu, James H. Martin & Dan Jurafsky. 2004. Shallow Semantic Parsing using Support Vector Machines. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, 233–240. Association for Computational Linguistics. URL `http://aclweb.org/anthology/N04-1030`.

Prince, Ellen F. 1981. Toward a taxonomy of given-new information. In Peter Cole (ed.), *Radical Pragmatics*, 223–256. New York: Academic Press. URL http://www.ling.upenn.edu/~ellen/givennew.pdf.

Prince, Ellen F. 1992. The ZPG Letter: Subjects, Definiteness, and Information-status. In William C. Mann & Sandra A. Thompson (eds.), *Discourse Description: Diverse linguistic analyses of a fund-raising text*, 295–325. Philadelphia/Amsterdam: John Benjamins.

Pulman, Stephen G. & Jana Z. Sukkarieh. 2005. Automatic Short Answer Marking. In Jill Burstein & Claudia Leacock (eds.), *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, 9–16. Ann Arbor, Michigan: Association for Computational Linguistics. URL http://aclweb.org/anthology/W05-0202.

Qing, Ciyang, Ulle Endriss, Raquel Fernandez & Justin Kruger. 2014. Empirical Analysis of Aggregation Methods for Collective Annotation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1533–1542. Dublin, Ireland: Dublin City University and Association for Computational Linguistics. URL http://aclweb.org/anthology/C14-1145.

Quixal, Martí & Detmar Meurers. 2016. How can writing tasks be characterized in a way serving pedagogical goals and automatic analysis needs? *CALICO Journal* 33. 19–48. URL http://purl.org/dm/papers/Quixal.Meurers-16.html.

Razagifard, Parisa & Massoud Rahimpour. 2010. The effect of computer-mediated corrective feedback on the development of second language learners' grammar. *International Journal of Instructional Technology and Distance Learning* 7(5). 11–30.

Reinhart, Tanya. 1981. Pragmatics and Linguistics: An Analysis of Sentence Topics. *Philosophica* 27. 53–94.

Richter, Frank & Manfred Sailer. 2003. Basic Concepts of Lexical Resource Semantics. In Arnold Beckmann & Norbert Preining (eds.), *ESSLLI 2003*

– *Course Material I*, vol. 5 Collegium Logicum, 87–143. Wien: Kurt Gödel Society.

Riester, Arndt & Stefan Baumann. 2011. Information Structure Annotation and Secondary Accents. In Stefanie Dipper & Heike Zinsmeister (eds.), *Beyond Semantics. Corpus-based Investigations of Pragmatic and Discourse Phenomena*, vol. 3 Bochumer Linguistische Arbeitsberichte, 111–127. Ruhr Universität Bochum. URL `http://www.ims.uni-stuttgart.de/~arndt/doc/baumannRiesterBeyondSem.pdf`.

Riester, Arndt & Stefan Baumann. 2013. Focus Triggers and Focus Types from a Corpus Perspective. *Dialogue & Discourse* 4(2). 215–248. doi: 10.5087/dad.2013.210.

Riester, Arndt, Lisa Brunetti & Kordula De Kuthy. in press. Annotation Guidelines for Questions under Discussion and Information Structure. In Evangelia Adamou, Katharina Haude, & Martine Vanhove (eds.), *Information structure in lesser-described languages: Studies in prosody and syntax* Studies in Language Companion Series, John Benjamins.

Riester, Arndt, David Lorenz & Nina Seemann. 2010. A Recursive Annotation Scheme for Referential Information Status. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valletta, Malta. URL `http://www.lrec-conf.org/proceedings/lrec2010/pdf/764_Paper.pdf`.

Riezler, Stefan. 2014. On the Problem of Theoretical Terms in Empirical Computational Linguistics. *Computational Linguistics* 40(1). 235–245.

Ritz, Julia, Stefanie Dipper & Michael Götze. 2008. Annotation of Information Structure: An Evaluation Across Different Types of Texts. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 2137–2142. Marrakech, Morocco. URL `http://www.lrec-conf.org/proceedings/lrec2008/pdf/543_paper.pdf`.

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

Roberts, Craige. 2012. Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics. *Semantics and Pragmatics* 5(6). 1–69. doi: 10.3765/sp.5.6.

Rooth, Mats. 1985. *Association with focus*. Amherst, MA: University of Massachusetts dissertation.

Rooth, Mats. 1992. A theory of focus interpretation. *Natural Language Semantics* 1(1). 75–116.

Rosé, Carolyn Penstein, Antonio Roque, Dumisizwe Bhembe & Kurt VanLehn. 2003. A Hybrid Approach to Content Analysis for Automatic Essay Grading. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers - Volume 2* NAACL-Short '03, 88–90. Edmonton, Canada: Association for Computational Linguistics.

Rudzewitz, Björn. 2015. *Alignment Weighting for Short Answer Assessment*. University of Tübingen Bachelor's thesis. URL `www.sfs.uni-tuebingen.de/~brzdwtz/resources/BA_Thesis.pdf`.

Rudzewitz, Björn. 2016a. Exploring the Intersection of Short Answer Assessment, Authorship Attribution, and Plagiarism Detection. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 235–241. San Diego, CA. URL `https://aclweb.org/anthology/W16-0527.pdf`.

Rudzewitz, Björn. 2016b. *An integrated approach to answer selection in question answering: Exploring multiple information sources and domain adaptation*. Department of Linguistics, University of Tübingen Master thesis in computational linguistics. URL `http://www.sfs.uni-tuebingen.de/~brzdwtz/resources/MA_Thesis.pdf`.

Rudzewitz, Björn & Ramon Ziai. 2015. CoMiC: Adapting a Short Answer Assessment System for Answer Selection. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 247–251. ACL. URL `http://aclweb.org/anthology/S15-2044`.

Salton, Gerard & Michael J. McGill. 1983. *Introduction to modern information retrieval*. New York: McGraw-Hill.

Schiller, Anne, Simone Teufel & Christine Thielen. 1995. The Stuttgart-Tübingen Tagset (STTS). Tech. rep. Universität Stuttgart, Universität Tübingen Germany. URL `http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html`.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, 44–49. Manchester, UK. URL `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf`.

Schmidt, Thomas. 2004. Transcribing and annotating spoken language with EXMARaLDA. In *Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004*, Paris: ELRA. URL `http://www.exmaralda.org/files/Paper_LREC.pdf`. EN.

Schölkopf, Bernhard & Alexander J. Smola. 2003. A Short Introduction to Learning with Kernels. In *Advanced Lectures on Machine Learning*, Springer-Verlag.

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

Selkirk, Elisabeth. 1984. *Phonology and syntax. The relation between sound and structure*. Cambridge, MA: MIT Press.

Selkirk, Elisabeth. 2002. Contrastive FOCUS vs. presentational focus: prosodic evidence from right node raising in English. In *Proceedings of Speech Prosody*, 643–646. Aix-en-Provence, France.

Sgall, Petr, Hajicová & Jarmila Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Academia.

Siddiqi, R., C.J. Harrison & R. Siddiqi. 2010. Improving Teaching and Learning through Automated Short-Answer Marking. *IEEE Transactions on Learning Technologies* 3(3). 237–249. doi: 10.1109/TLT.2010.4.

Silverman, Kim, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert & Julia Hirschberg. 1992. ToBI: A Standard for Labeling English Prosody. In *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP '92)*, 867–870. Banff, Alberta, Canada.

Skopeteas, Stavros, Ines Fiedler, Samantha Hellmuth, Anne Schwarz, Ruben Stoel, Gisbert Fanselow, Caroline Féry & Manfred Krifka. 2006. *Questionnaire on information structure (QUIS): reference manual*, vol. 4 Interdisciplinary studies on information structure (ISIS). Universitätsverlag Potsdam.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky & Andrew Y. Ng. 2008. Cheap and Fast—but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* EMNLP '08, 254–263. Stroudsburg, PA, USA: Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1613715.1613751`.

Soon, Wee Meng, Hwee Tou Ng & Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics* 27(4). 521–544.

Sridhar, Vivek Kumar Rangarajan, Ani Nenkova, Shrikanth Narayanan & Dan Jurafsky. 2008. Detecting prominence in conversational speech: pitch accent, givenness and focus. In *Proceedings of Speech Prosody*, 380–388. Campinas, Brazil.

Stechow, Arnim von. 1991. Focusing and backgrounding operators. In W. Abraham (ed.), *Discourse Particles*, 37–84. Amsterdam/Philadelphia: John Benjamins Publishing Co.

Stede, Manfred. 2004. The Potsdam Commentary Corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, 96–102. Stroudsburg, PA, USA: Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1608938.1608951`.

Stede, Manfred. 2012. Computation and modeling of information structure. In Manfred Krifka & Renate Musan (eds.), *The Expression of Information Structure*, vol. 5 The Expression of Cognitive Categories, 363–408. Berlin/Boston: De Gruyter Mouton.

Stede, Manfred & Sara Mamprin. 2016. Information structure in the Potsdam Commentary Corpus: Topics. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France: European Language Resources Association (ELRA).

Sukkarieh, Jana Z. & Svetlana Stoyanchev. 2009. Automating Model Building in c-rater. In R. Barzilay, J.-S. Chang & C. Sauper (eds.), *Proceedings of the 2009 Workshop on Applied Textual Inference (TextInfer)*, 61–69. Singapore: Association for Computational Linguistics. URL `http://aclweb.org/anthology/W/W09/W09-2509.pdf`.

Tandalla, Luis. 2012. Scoring short answer essays. ASAP Short Answer Scoring Competition System Description. Tech. rep. Downloaded from http://kaggle.com/asap-sas/.

Telljohann, Heike, Erhard Hinrichs & Sandra Kübler. 2004. The TüBa-D/Z Treebank: Annotating German with a Context-Free Backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lissabon.

Telljohann, Heike, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister & Kathrin Beck. 2015. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Tech. rep. Seminar für Sprachwissenschaft, Universität Tübingen Germany. URL `http://www.sfs.uni-tuebingen.de/fileadmin/user_upload/ascl/tuebadz-stylebook-1508.pdf`.

Tjong Kim Sang, Erik F. & Sabine Buchholz. 2000. Introduction to the CoNLL-2000 Shared Task: Chunking. In *Proceedings of the Fourth Conference on Computational Language Learning (CoNLL-2000) and the Second Learning Language in Logic Workshop (LLL-2000)*, 127–132. Lisbon, Portugal. URL `http://lcg-www.uia.ac.be/conll2000/ps/12732tjo.ps`.

Tjong Kim Sang, Erik F. & Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, vol. 4, 142–147. Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.3115/1119176.1119195. URL `http://dx.doi.org/10.3115/1119176.1119195`.

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

Uriarte, Adrian Bello. 2013. Vocabulary Teaching: Focused Tasks for Enhancing Acquisition in EFL Contexts. *MEXTESOL Journal* 37(2).

VanLehn, Kurt, Pamela W. Jordan, Carolyn Penstein Rosé, Dumisizwe Bhembe, Michael Boettner, Andy Gaydos, Maxim Makatchev, Umarani Pappuswamy, Micheal Ringenberg, Antonio Roque, Stephanie Siler & Ramesh Srivastava. 2002. The Architecture of Why2-Atlas: A Coach for Qualitative Physics Essay Writing. In *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, vol. 2363, 158–167. Biarritz, France and San Sebastian, Spain: Springer LNCS.

Vesela, Katerina, Jirí Havelka & Eva Hajicová. 2004. Annotators Agreement: The Case of Topic-Focus Articulation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*, European Language Resources Association. URL `http://www.lrec-conf.org/proceedings/lrec2004/pdf/350.pdf`.

Weiss, Sholom M. & Casimir A. Kulikowski. 1991. *Computer systems that learn: Classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. San Mateo, CA: Morgan Kaufmann.

Wolpert, David H. 1992. Stacked generalization. *Neural Networks* 5(2). 241–259.

Zang, Xiao, Zhiyong Wu, Helen Meng, Jia Jia & Lianhong Cai. 2014. Using Conditional Random Fields to Predict Focus Word Pair in Spontaneous Spoken English. In *Proceedings of INTERSPEECH 2014*, 756–760. International Speech Communication Association.

Zhang, Tong, Mark Hasegawa-Johnson & Stephen E. Levinson. 2006. Extraction of pragmatic and semantic salience from spontaneous spoken English. *Speech Communication* 48(3–4). 437–462. doi: http://dx.doi.org/10.1016/j.specom.2005.07.007. URL `http://www.sciencedirect.com/science/article/pii/S0167639305001743`. Spoken Language Understanding in Conversational Systems.

Ziai, Ramon, Kordula De Kuthy & Detmar Meurers. 2016. Approximating Givenness in Content Assessment Through Distributional Semantics. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM)*, 209–218. Berlin, Germany: ACL. URL `http://aclweb.org/anthology/S16-2026.pdf`.

Ziai, Ramon & Detmar Meurers. 2014. Focus Annotation in Reading Comprehension Data. In *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII, 2014)*, 159–168. COLING Dublin, Ireland: ACL. URL `http://aclweb.org/anthology/W14-4922.pdf`.

Ziai, Ramon & Detmar Meurers. 2018. Automatic Focus Annotation: Bringing Formal Pragmatics Alive in Analyzing the Information Structure of Authentic Data. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, New Orleans, LA: ACL. URL `http://www.sfs.uni-tuebingen.de/~rziai/papers/Ziai.Meurers-18.pdf`. To appear.

Ziai, Ramon, Niels Ott & Detmar Meurers. 2012. Short Answer Assessment: Establishing Links Between Research Strands. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7) at NAACL-HLT 2012*, 190–200. Montreal. URL `http://aclweb.org/anthology/W12-2022.pdf`.

A. Example Corpus Data From CREG

Example with a short text from CREG-1032

Schloss Pillnitz (Source: http://www.geo.de/reisen/community/reisebericht/42815/2/Elbflorenz-Dresden-und-Umgebung)
Das Schloss, das im Osten Dresdens liegt, ist für mich das schönste Schloss in Dresdens Umgebung. Beim Bummel durch den weitläufigen Park, vorbei am Palmenhaus und der Orangerie kann man allen Stress und alle Sorgen vergessen und einfach die Schönheit der Umgebung genießen. Hier findet man im Sommer zahlreiche Hochzeitspaare – wegen seiner Schönheit ist Pillnitz einer der beliebtesten Szenerien zum Heiraten in Dresden. 1768, mit dem Beginn der Regierungszeit von Kurfürst Friedrich August III. (1750-1827), wurde Pillnitz zur Sommerresidenz der sächsischen Könige.

Eine besondere Attraktion im Park ist die Kamelie. Die mittlerweile über 230 Jahre alte und 8,90 m hohe Kamelie bekam 1992 ein fahrbares Haus, in dem Temperatur, Belüftung, Luftfeuchte und Beschattung durch einen Klimacomputer geregelt werden . In der warmen Jahreszeit wird das Haus neben die Kamelie gerollt. Während der Blütezeit von Mitte Februar bis April trägt sie zehntausende karminrote Blüten. Ableger der Pillnitzer Kamelie werden jedes Jahr in begrenzter Zahl während der Blütezeit verkauft, dann ist ein Besuch besonders lohnend.

Example question, reference and learner answer:

Question:	Wie kann man sich in Pillnitz erholen und den Stress vergessen?
Ref. answer:	Bei einem Bummel durch den Park kann man den Stress vergessen.
Learner answer:	Beim Bummel durch den weitläufigen Park, vorbei am Palmenhaus und der Orangerie kann man sich erholen und den Stress vergessen.

Example with a longer text from CREG-5K

Schwarzmarkt (Source: Kordon 2006)

Der schwarze Markt lag in einer Seitenstraße, die sich von all den anderen heilgebliebenen Straßen ringsherum nur darin unterschied, daß sie belebter war. Frauen, Männer und junge Burschen gingen in ihr auf und ab und murmelten dabei ständig etwas vor sich hin, was sich wie Selbstgespräche anhörte. Aber sie sprachen nicht mit sich selbst, ihre Augen blickten wach und fragend. Ging ein Entgegenkommender näher heran, hörte er, daß ihm ein Angebot gemacht wurde. "Leberwurst, frisch vom Land!" flüsterte da ein älterer Mann, und ein junger Bursche hatte "Nylons! Echt amerikanische Nylon-Strümpfe mit Naht" anzubieten. Eine Frau bot "Rasierklingen, extra fein", eine andere "Meißner Porzellan, Tassen Teller, Untertassen" an.

Frau Kagelmann kannte die Regeln des schwarzen Marktes, aber sie hatte Hemmungen, es den anderen gleichzutun. Sie brauchte jedesmal eine Anlaufzeit.

Ein Kriegsinvalide ohne Beine, der auf einem Brett mit Rollen hockte und sich mit den Händen vom Pflaster abstieß, hielt vor Frau Kagelmann. "Brauchen Sie wieder Garn?" fragte er. Er hatte in ihr eine ehemalige Kundin erkannt.

"Diesmal nicht", antwortete Frau Kagelmann. "Diesmal brauche ich was zu essen. Mein Sohn ist heimgekehrt."

"Gesund?" fragte der Mann, der zu Frau Kagelmann aufschauen mußte.

"Ja."

"Herzlichen Glückwunsch!" Der Invalide rollte weiter. "Und wenn Sie mal wieder Garn brauche, Sie wissen ja!"

Frau Kagelmann sah dem Mann auf seinem Brett nach. Das hätte Uli auch passieren können. Sie durfte sich wirlich nicht beschweren.

"Eine Uhr! Eine silberne Spieluhr!" Frau Kagelmann begann nun ebenfalls zu flüstern. Der freundliche Invalide hatte ihr Mut gemacht. Und sie hatte Erfolg. Ein junger Mann mit einem ein wenig zu großen Hut auf dem Kopf machte sich an sie heran. "Zeigen", sagte er.

Frau Kagelmann ging in einen Hausflur, holte die Uhr heraus und zeigte sie dem jungen Mann. Doch als er danach greifen wollte, zog sie sie wieder zurück. "So nicht! Womit zahlen Sie?"

"Zigaretten." Der junge Mann öffnete seine Jacke: In seinem Hosenbund steckte eine Stange amerikanische Zigaretten.

Frau Kagelmann blieb mißtrauisch. Sie hielt dem Mann die Uhr wieder hin, wickelte sich die Kette aber um die Hand, damit er ihr die Uhr nicht entreißen konnte.

Der junge Mann besah sich die Uhr. "Hundert", sagte er dann.

"Hundert was?"

"Hundert Zigaretten. Fünf Päckchen. Eine halbe Stange!"

Frau Kagelmann ließ sich die Zigaretten geben und verstaute die Päckchen einzeln in ihrer Bluse. Erst dann gab sie dem jungen Mann die Uhr.

Der junge Mann zog sie auf und hielt sie an sein Ohr. "Eine Erinnerung an Ihren Mann?" fragte er. Und als Frau Kagelmann nickte, schob er sich den Hut ins Genick und grinste: "Wenn sie sie wiederhaben wollen – Preis: eine Stange Amis."

Frau Kagelmann erwiderte nichts. Sie verließ den Hausflur und ging weiter die Straße entlang. Was jetzt noch kam, war leicht. Zigaretten waren die beste Währung, Zigaretten nahm einem jeder ab, und man konnte sie päckchen- oder stückweise eintauschen.

Es dauerte nicht lange und Frau Kagelmann besaß anstelle der fünf Päckchen Zigaretten ein halbes Pfund Trockengemüse, ein Pfund Graupen, ein halbes Brot, ein Glas Marmelade und ein viertel Pfund Trockenmilch. Sie wußte nicht, ob sie für die Uhr und danach für die fünf Päckchen mehr hätte herausschlagen können, aber das wußte sie nie, wenn sie den schwarzen Markt verließ, deshalb war sie zufrieden.

Sie hatte die Flüsterstraße noch nicht verlassen, als ein Pfiff ertönte und drei Jungen an ihr vorbeiliefen und "Razzia!" schrien.

Polizei! Wenn die fanden, was sie bei sich trug, würden sie es ihr abnehmen. Frau Kagelmann schaltete schnell: Zum Fortlaufen war sie nicht flink genug, also mußte sie sich verstecken. Ganz langsam, als gingen sie die fliehenden Schwarzhändler, die in immer größerer Anzahl an ihr vorüberliefen nichts an, ging sie auf einen der Hauseingänge zu und hinein. Durch den Hausflur gelangte sie auf den Hof und betrat dort die Kellertreppe. Es war dunkel in dem Keller, aber sie machte kein Licht. Sie tastete sich bis an das Ende des Kellerganges und lehnte sich an einen der Holzversschläge.

Example question, reference and learner answer:

Question: Warum nannte man den schwarzen Markt “die Flüsterstrasse”?

Ref. answer: Er wurde so genannt, weil die Leute ständig etwas vor sich hin murmelten. Näherte sich eine Person den Männern und Frauen auf dem Markt, dann flüsterten diese ihr ihre Angebote zu.

Learner answer: Die Straße hieß die Flüsterstrasse, weil diese Straße ein Schwarzmarkt hatte und man flüstern musste.

B. German Tagsets

STTS POS Tagset (Schiller et al., 1995)

This table lists the part-of-speech categories of the Stuttgart-Tübingen Tagset (STTS) as used in the TüBa-D/Z. It is a verbatim copy of the table in Telljohann et al. (2015).

POS	**Description**	**Examples**
ADJA	attributive adjective	*[das] große [Haus]*
ADJD	adverbial or predicative adjective	*[er fährt] schnell, [er ist] schnell*
ADV	adverb	*schon, bald, doch*
APPR	preposition; left circumposition	*in [der Stadt], ohne [mich]*
APPRART	preposition + article	*im [Haus], zur [Sache]*
APPO	postposition	*[ihm] zufolge, [der Sache] wegen*
APZR	right circumposition	*[von jetzt] an*
ART	definite or indefinite article	*der, die, das, ein, eine*
CARD	cardinal number	*zwei [Männer], [im Jahre] 1994*
FM	foreign language material	*[Er hat das mit "]* *A big fish [" übersetzt]*
ITJ	interjection	*mhm, ach, tja*
KOUI	subordinating conjunction with zu + infinitive	*um [zu leben], anstatt [zu fragen]*
KOUS	subordinating conjunction with clause	*weil, daß, damit, wenn, ob*
KON	coordinative conjunction	*und, oder, aber*
KOKOM	particle of comparison, no clause	*als, wie*
NN	noun	*Tisch, Herr, [das] Reisen*
NE	proper noun	*Hans, Hamburg, HSV*

POS	Description	Examples
PDS	substituting demonstrative pronoun	*dieser, jener*
PDAT	attributive demonstrative pronoun	*jener [Mensch]*
PIS	substituting indefinite pronoun	*keiner, viele, man, niemand*
PIAT	attributive indefinite pronoun without determiner	*kein [Mensch], irgendein [Glas]*
PIDAT	attributive indefinite pronoun with determiner	*[ein] wenig [Wasser],* *[die] beiden [Brüder]*
PPER	irreflexive personal pronoun	*ich, er, ihm, mich, dir*
PPOSS	substituting possessive pronoun	*meins, deiner*
PPOSAT	attributive possessive pronoun	*mein [Buch], deine [Mutter]*
PRELS	substituting relative pronoun	*[der Hund,] der*
PRELAT	attributive relative pronoun	*[der Mann ,] dessen [Hund]*
PRF	reflexive personal pronoun	*sich, einander, dich, mir*
PWS	substituting interrogative pronoun	*wer, was*
PWAT	attributive interrogative pronoun	*welche [Farbe], wessen [Hut]*
PWAV	adverbial interrogative or relative pronoun	*warum, wo, wann, worüber, wobei*
PROP	pronominal adverb	*dafür, dabei, deswegen, trotzdem*
PTKZU	zu + infinitive	*zu [gehen]*
PTKNEG	negation particle	*nicht*
PTKVZ	separated verb particle	*[er kommt] an, [er fährt] rad*
PTKANT	answer particle	*ja, nein, danke, bitte*
PTKA	particle with adjective or adverb	*am [schönsten], zu [schnell]*
TRUNC	truncated word - first part	*An- [und Abreise]*

POS	Description	Examples
VVFIN	finite main verb	*[du] gehst, [wir] kommen [an]*
VVIMP	imperative, main verb	*komm [!]*
VVINF	infinitive, main	*gehen, ankommen*
VVIZU	infinitive + zu, main	*anzukommen, loszulassen*
VVPP	past participle, main	*gegangen, angekommen*
VAFIN	finite verb, aux	*[du] bist, [wir] werden*
VAIMP	imperative, aux	*sei [ruhig !]*
VAINF	infinitive, aux	*werden, sein*
VAPP	past participle, aux	*gewesen*
VMFIN	finite verb, modal	*dürfen*
VMINF	infinitive, modal	*wollen*
VMPP	past participle, modal	*[er hat] gekonnt*
XY	non-word containing special characters	*D2XW3, letters*
$,	comma	*,*
$.	sentence-final punctuation	*. ? ! ; :*
$(	other sentence-internal punctuation	*- [ ] ( )*

Hamburg Dependency Tagset (Foth, 2006)

This table lists the dependency categories of the Hamburg Dependency Treebank. It is a tabular version of the list in Foth et al. (2014).

Label	**Description**
ADV	Denotes adverbial modification by proper adverbs or words from related classes (predicative adjectives and various particles that the STTS assigns to their own class)
APP	(apposition, always subordinated strictly left to right) Relates adjacent nominal words in the same NP (headline phrases) or in proper appositions (I, Robot)
ATTR	Attributive adjectives or numbers modifying a noun
AUX	Auxiliary, connects verbs in the same verb group, the finite verb is always the head of such a chain
AVZ	(Abtrennbarer VerbZusatz) separable verb particle, attaches a separated verb particle to its verb
CJ	Conjunct, complement of a conjunction, i. e. connected to a word like 'und'
DET	Determiner of a noun
ETH	Ethic dative, i. e. a nominal adjunct in the dative case that is not licensed by a verb frame
EXPL	(expletive) only used for the expletive use of the pronoun 'es'
GMOD	Genitive modification, the dependent word is in the genitive case and modifies a nominal
GRAD	Gradual, an NP indicating a measurement as in "three meters deep"
KOM	Comparison words modifying a noun or a verb, typically 'wie' or 'als'
KON	Coordination connecting words in a coordination chain (except the final word below a coordination, which is CJ). In coordinations, the word to the left is always the head of the word to the right
KONJ	Conjunction modifying a verb signalling an SOV subclause

Label	Description
NEB	(Nebensatz) Subordinate clause, connecting the finite verb of the subordinate clause to the verb in the superordinate main clause. (For some types of subclauses, such as relative clauses, there are special labels.)
NP2	A rare label for logical subjects in elliptical coordinations
OBJA	Accusative object
OBJA2	Second accusative object, for the rare case where a verb has a valency for two accusative objects
OBJC	Object clause, for the finite verb in a subclause that is attached to a verb as a complement
OBJD	Dative object
OBJG	Genitive object
OBJI	Infinitive verb used as a complement to another verb
OBJP	Prepositional object, for prepositions that are a complement to a verb. In contrast to a PP, it cannot be omitted.
PAR	Parenthesis, superior clause that is inserted into its subclause. In such a case, to prevent a non-projective structure, the finite verb of the subclause is attached to the last word before the inserted clause.
PART	Particle, for example 'zu' modifying an infinite verb, or the second part of a circumposition modifying the respective preposition
PN	The complement of a preposition (or post-position)
PP	Prepositional phrase, for the attachment of prepositions
PRED	Predicative complement, mostly for the verb 'sein'
REL	(relative clause) Connects the finite verb of a relative clause to its (nominal or verbal) antecedent. Often non-projective.
S	(sentence) the label for the root node of SVO sentences and phrase fragments, or an SVO sentence subordinated to a verb as a complement.
SUBJ	(surface subject) Any nominal material filling the subject slot of a verb (not necessarily the Vorfeld position, see 'EXPL')
SUBJC	(subject clause) Any verbal material filling a subject slot

Label	Description
VOK	(Vokativ) Salutation, usually a proper name, arbitrarily attached to the nearest word because of its tenuous connection with the syntax tree
ZEIT	(time) Time information in the form of (usually four- digit) year numbers attached without a preposition
''	(the empty label) for punctuation marks
REF	The only label for the separate reference level: the label of pronouns attached to their antecedent.

TüBa-D/Z Constituent Tagset (Telljohann et al., 2004)

This table lists the constituent categories of the TüBa-D/Z. It is a verbatim copy of the table in Telljohann et al. (2015).

Node Label	**Description**
	Phrase Node Labels
ADJX	adjectival phrase
ADVX	adverbial phrase
DP	determiner phrase (e.g. *gar keine*)
FX	foreign language phrase
NX	noun phrase
PX	prepositional phrase
VXFIN	finite verb phrase
VXINF	non-finite verb phrase
	Topological Field Node Labels
LV	resumptive construction (Linksversetzung)
C	complementizer field (C-Feld)
FKOORD	coordination consisting of conjuncts of fields
KOORD	field for coordinating particles
LK	left sentence bracket (Linke (Satz-)Klammer)
MF	middle field (Mittelfeld)
MFE	middle field between VCE and VC
NF	final field (Nachfeld)
PARORD	field for non-coordinating particles
VC	verb complex (Verbkomplex)
VCE	verb complex with the split finite verb of *Ersatzinfinitiv* constructions
VF	initial field (Vorfeld)
FKONJ	conjunct consisting of more than one field
	Root Node Labels
DM	discourse marker
P-SIMPX	paratactic construction of simplex clauses
R-SIMPX	relative clause
SIMPX	simplex clause