

Focus Annotation in Reading Comprehension Data

Ramon Ziai **Detmar Meurers**
Sonderforschungsbereich 833
Eberhard Karls Universität Tübingen
{rzi, dm}@sfs.uni-tuebingen.de

Abstract

When characterizing the information structure of sentences, the so-called *focus* identifies the part of a sentence addressing the current question under discussion in the discourse. While this notion is precisely defined in formal semantics and potentially very useful in theoretical and practical terms, it has turned out to be difficult to reliably annotate focus in corpus data.

We present a new focus annotation effort designed to overcome this problem. On the one hand, it is based on a task-based corpus providing more explicit context. The annotation study is based on the CREG corpus (Ott et al., 2012), which consists of answers to explicitly given reading comprehension questions. On the other hand, we operationalize focus annotation as an incremental process including several substeps which provide guidance, such as explicit answer typing.

We evaluate the focus annotation both intrinsically by calculating agreement between annotators and extrinsically by showing that the focus information substantially improves the automatic meaning assessment of answers in the CoMiC system (Meurers et al., 2011).

1 Introduction

This paper discusses the interplay of linguistic and computational linguistic aspects in the analysis of focus as a core notion of information structure. Empirically, our work focuses on analyzing the responses to reading comprehension questions. In computational linguistics, automatic meaning assessment determining whether a response appropriately answers a given question about a given text has developed into an active field of research. Short Answer Assessment recently was also highlighted by the *Joint Student Response Analysis and Textual Entailment Challenge* (Dzikovska et al., 2013). Some research in this domain has pointed out the relevance of identifying which parts of a response are given by the question (Bailey and Meurers, 2008; Mohler et al., 2011), with recent work pointing out that the relevant notion here is that of *focus* as discussed in formal pragmatics (Meurers et al., 2011; Hahn and Meurers, 2012).

Figure 1 provides an example of answer comparison for meaning assessment, where the focus (marked by square brackets) can effectively be used to zoom in on the information that is relevant for comparing a target answer (TA) with a student answer (SA) given a question (Q).

Q: *Wohin* ging Schorlemmer nach der Demo?
"Where did Schorlemmer go after the rally?"

TA: Er ist glücklich **[[nach Hause]]_F** gelaufen.
He is happily to home walked

SA: Schorlemmer ging **[[heim]]_F**.
Schorlemmer went home

Figure 1: Answer comparison with the help of focus

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Page numbers and proceedings footer are added by the organisers. Licence details: <http://creativecommons.org/licenses/by/4.0/>

To support this line of research, one needs to be able to identify the focus in a response. As a first step, we have designed an annotation scheme and manual annotation process for identifying the focus in a corpus of reading comprehension responses. Focus here is understood in the sense of Krifka (2007) as indicating the presence of alternatives in the context and being a direct answer to the Question Under Discussion (QUD, Roberts 1996). This semantic view of focus is essentially language-independent.

Some attempts at systematically identifying focus in authentic data have been made in the past (Dipper et al., 2007; Calhoun et al., 2010). However, most approaches either capture a notion of focus more closely related to particular language features, such as the Topic-Focus Articulation and its relation to the word order in Czech (Buráňová et al., 2000), or the approaches were not rewarded with much success (Ritz et al., 2008). The latter have tried to identify focus in newspaper text or other data types where no explicit questions are available, making the task of determining the QUD, and thus reliably annotating focus, very hard. In contrast, in the research presented here, we work with responses to explicitly given questions that are asked about an explicitly given text. Thus, we can make use of the characteristics of the questions and text to obtain reliable focus annotation for the responses.

Theoretical linguists have discussed the notion of focus for decades, cf., e.g., Jackendoff (1972), Stechow (1981), Rooth (1992), Schwarzschild (1999) and Büring (2007). However, for insights and arguments from theoretical work to be applicable in computational linguistics, they need to be linked to thorough empirical work – an area where some work remains to be done (cf., e.g., De Kuthy and Meurers, 2012), with some recent research making significant headway (Riester and Baumann, 2013). As it stands, computational linguists have not yet been able to fully profit from the theoretical debate on focus. An important reason complementing the one just mentioned is the fact that the context in which the text to be analyzed is produced has rarely been explicitly taken into account and encoded. Yet, many of the natural tasks in which focus annotation would be relevant actually do contain explicit task and context information of relevance to determining focus. To move things forward, this paper builds on the availability and relevance of task-based language data and presents an annotation study of focus on authentic reading comprehension data. As a second component of our proposal, we operationalize the focus annotation in terms of several incremental steps, such as explicit answer typing, which provide relevant information guiding the focus annotation as such.

Overall, the paper tries to accomplish two goals, which are also reflected in the way the annotation is evaluated: i) to present an effective focus annotation scheme and to evaluate how consistently it can be applied, and ii) to explore the possible impact of focus annotation on Short Answer Assessment. Establishing a focus annotation scheme for question-response pairs from authentic reading comprehension data involves sharpening and linking the concepts and tests from theoretical linguistic with the wide range of properties realized in the authentic reading comprehension data. The work thus stands to contribute both to an empirical evaluation and enrichment of the linguistic concepts as well as to the development of automatic focus annotation approaches in computational linguistics.

The paper is organized as follows: Section 2 presents the corpus data on which we base the annotation effort and the annotation process. Section 3 introduces the scheme we developed for annotating the reading comprehension data. Section 4 then launches into both intrinsic and extrinsic evaluation of the manual annotation, before section 5 concludes the paper.

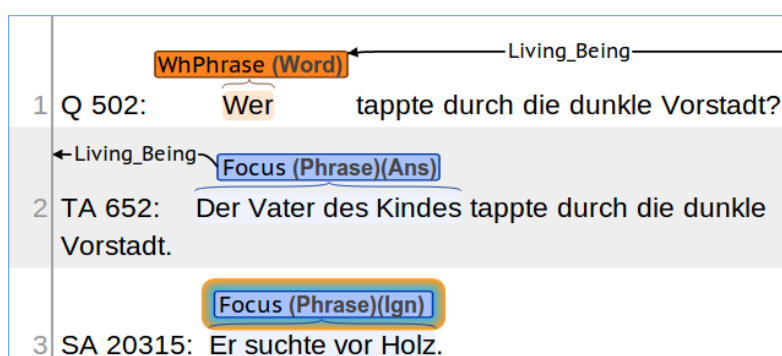
2 Data and Annotation Setup

We base our work on the CREG corpus (Ott et al., 2012), a task-based corpus consisting of answers to reading comprehension questions written by learners of German at the university level. The overall corpus includes 164 reading texts, 1,517 reading comprehension questions, 2,057 target answers provided by the teachers, and 36,335 learner answers. We use the CREG-1032 data subset (Meurers et al., 2011) for the present annotation work in order to enable comparison to previously published results on that data set (Meurers et al., 2011; Hahn and Meurers, 2012; Horbach et al., 2013). The CREG-1032 data set consists of two sub-corpora, which correspond to the sites they were collected at, Kansas University (KU) and Ohio State University (OSU). For the present work, we limited ourselves to the OSU portion of the data because it contains longer answers and more answers per question.

The OSU subset consists of 422 student answers to 60 questions, for which 87 target answers are available. The student answers were produced by 175 intermediate learners of German in the US, who on average wrote about 15 tokens per answer. All student answers were rated by two annotators with respect to whether they answer the question or not. The subset is balanced, i.e. it contains the same number of correct and incorrect answers, and both annotators agreed on the meaning assessment.

To obtain a gold-standard focus annotation for this data set, we set out to manually annotate both target answers and student answers with focus. We also annotated the question forms in the question. The annotation was performed by two graduate research assistants in linguistics using the *brat*¹ rapid annotation tool 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, we drew a random sample of 100 questions, target answers and student answers from each sub-corpus of CREG and trained our two annotators on them. During this piloting process, the first author met with the annotators to discuss difficult cases and decide how the scheme would accommodate them.

Figure 2 shows a sample screenshot of the *brat* tool. The question asks for a person, namely the one ‘wandering through the dark outskirts’. The target response provides an answer with an appropriate focus. The student response instead appears to answer a question about the reason for this person’s action, such as ‘Why did he wander through the dark outskirts?’.



Q: ‘Who wandered through the dark outskirts?’
 TA: ‘The child’s father wandered through the dark outskirts.’
 SA: ‘He searched for wood.’

Figure 2: Example with a *who*-question and a different QUD for the student answer

3 Annotation Scheme

In this section, we introduce the annotation scheme we developed. 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, sec. 3), 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 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 1 gives an overview and examples of this dimension.

Focus is used to mark the focused words or phrases in an answer. We do not distinguish between contrastive and new information focus, as this is not relevant for assessing an answer. Multiple foci can be encoded and in fact do occur in the data.

¹<http://brat.nlplab.org>

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 1: Question Forms in the annotation scheme

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. We operationalize this by testing whether a given part of the utterance is needed to distinguish between alternatives in the QUD. 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 in the QUD. For instance, consider the example in (1), where the focus is made explicit by the square brackets.

- (1) Where does Heike live?
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 (2). Here the set of alternatives is more constrained and hence “in” is not focused.

- (2) In what city does Heike live?
She lives in [Berlin]_F.

Other criteria we defined to guide focus annotation include the following:

- Coordination: If several foci are coordinated, each should be marked separately.
- Givenness: Avoid marking given material except where needed to distinguish between alternatives.
- 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 (discussed below).
- Focus never crosses sentence boundaries.
- Focus does not apply to sub-lexical units, such as syllables.
- Punctuation at focus boundaries is to be excluded.

In addition to marking focus, we annotate the relation between the explicitly given question and the Question Under Discussion 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 (3), the QUD answered can be a subquestion of the explicit question, which we encode as *question narrowed down*. When it addresses a more general QUD, as in (4), the response is annotated as *question generalized*.

- (3) What did the pop stars wear?
The female pop stars wore caftans.
- (4) Would you like a Coke or a Sprite?
I’d like a beer.

Finally, we also mark complete failures of question answer congruence 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.

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, leading to potentially 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).

Category	Description	Example (translated)
<code>Time_Date</code>	time/date expression, usually incl. preposition	The movie starts <i>at 5:50</i>
<code>Living_Being</code>	individual, animal or plant	<i>The father of the child</i> padded through the dark outskirts.
<code>Thing</code>	concrete object which is not alive	For the Spaniards <i>toilet and stove</i> are more important than the internet.
<code>Abstract_Entity</code>	entity that is not concrete	The applicant needs a <i>completed vocational training as a cook</i> .
<code>Report</code>	reported incident or statement	The speaker says " <i>We ask all youths to have their passports ready.</i> "
<code>Reason</code>	reason or cause for a statement	The maintenance of a raised garden bed is easier <i>because one does not need to stoop</i> .
<code>Location</code>	place or relative location	She is from <i>Berlin</i> .
<code>Action</code>	activity or happening.	In the vegetable garden one needs to <i>hoe and water</i> .
<code>Property</code>	attribute of something	Reputation and money are <i>important</i> for Til.
<code>Yes.No</code>	polar answer, including whole statement if not elliptic	<i>The mermaid does not marry the prince.</i>
<code>Manner</code>	way in which something is done	The word is used <i>ironically</i> in this story.
<code>Quantity/Duration</code>	countable amount of something	The company seeks <i>75</i> employees.
<code>State</code>	state something is in, or result of some action	If he works hard now, <i>he won't have to work in the future</i> .

Table 2: Answer Types with examples

4 Evaluation

The approach is evaluated in two ways. First, the consistency with which the focus annotation scheme was applied is evaluated in section 4.1 by calculating inter-annotator agreement. In section 4.2 we then explore the effect of focus annotation on Short Answer Assessment. For both evaluations, we provide a qualitative discussion of characteristic examples.

4.1 Intrinsic Evaluation

4.1.1 Quantitative Results

Having carried out the manual annotation experiment, the question arises how to compare and calculate agreement of spans of tokens in focus annotation. While comparing individual spans and calculating some kind of overlap measure is certainly possible, it is hard to interpret the meaning of such numbers. We therefore decided to make as few assumptions as possible and treat each token as a markable for which the annotator needs to make a decision. On that basis, we then follow standard evaluation procedures in calculating percentage agreement and Cohen's Kappa (Artstein and Poesio, 2009).

Table 3 summarizes the agreement results. 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.

Type of distinction	Type of answers	# tokens	%	κ
Binary (focus/background)	Student	6329	82.8	.65
	Target	6983	84.9	.69
Detailed (13 Answer Types + background)	Student	5198	72.6	.61
	Target	6839	76.5	.67

Table 3: Inter-annotator agreement on student and target answers

The results show that all numbers are in the area of substantial agreement ($\kappa > .6$). This is a noticeably 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). 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 84.0% and $\kappa = .67$ for student answers, and 87.4% and $\kappa = .74$ for target answers. A few longer answers which are harder to annotate thus noticeably affected the agreement results of Table 3 negatively.

4.1.2 Examples

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

- Q: Warum nennt der Autor Hamburg das “Tor zur Welt der Wissenschaft”?
‘Why does the author call Hamburg the “gate to the world of science”?’
- SA: \llbracket Hamburg hat viel renommierte Universitäten \rrbracket_F (annotator 1)
Hamburg hat \llbracket viel renommierte Universitäten \rrbracket_F (annotator 2)
‘Hamburg has many renowned universities’

Figure 3: Disagreement involving given material

Whereas annotator 1 marks the whole answer on the grounds that the focus is of Answer Type `Reason` and needs to include the whole proposition, annotator 2 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:

- Q: Wofür ist der Aufsichtsrat verantwortlich?
‘What is the supervisory board responsible for?’
- SA: Der Aufsichtsrat ist für \llbracket die Bestellung \rrbracket_F verantwortlich. (annotator 1)
Der Aufsichtsrat ist \llbracket für die Bestellung \rrbracket_F verantwortlich. (annotator 2)
‘The supervisory board is responsible for the appointment.’

Figure 4: Disagreement on a preposition

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.

4.2 Extrinsic Evaluation

It has been pointed out that evaluating an expert annotation of a theoretical linguistic notion only intrinsically is problematic because there is no non-theoretical grounding involved (Riezler, 2014). Therefore, besides calculating agreement measures, we also evaluated the resulting annotation in a larger computational task, the automatic meaning assessment of answers to reading comprehension questions.

We used the CoMiC system (Comparing Meaning in Context, Meurers et al., 2011) as a testbed for our experiment. CoMiC is an alignment-based system operating in three stages:

1. Annotating linguistic units (words, chunks and dependencies) in student and target answer on various levels of abstraction
2. Finding alignments of linguistic units between student and target answer based on annotation
3. Classifying the student answer based on number and type of alignments, using a supervised machine learning setup with 13 features in total

In stage 2, CoMiC integrates a simplistic approach to givenness, excluding all words from alignment that are mentioned in the question. 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. We only used the foci where students did not ignore the question according to the annotators.

For the present evaluation, 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. All three settings were tried out for annotator 1 and 2.

4.2.1 Quantitative Results

Table 4 summarizes the quantitative results. It shows that focus beats the basic givenness baseline of 84.6% on its own, pushing the classification accuracy to 86.7% for annotator 1 and 87.2% for annotator 2.

	Annotator 1	Annotator 2
Basic givenness only	84.6	
Focus only	86.7	87.2
Focus + givenness	90.3	89.3

Table 4: Answer classification accuracy with the CoMiC system

While this is an encouraging result already, the combination of basic givenness and focus performs substantially better, reaching 90.3% accuracy for annotator 1 and 89.3% for annotator 2.

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.

Interestingly, the focus annotation of annotator 2 on its own performed better than that of annotator 1, but worse when combined with basic givenness. We suspect that annotator 2's understanding of focus relied more on the concept of givenness than annotator 1's, causing the combination of the two to be less informative than for annotator 1.

4.2.2 Alignment Example

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:

- 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 \llbracket eingehaust \rrbracket_F
 ‘Twelve of the 47 wells are encased 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’

Figure 5: No alignments because the student answer ignores the question

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, CoMiC wrongly classifies this answer as correct, but with the added focus information, it has the means to judge this answer adequately.

5 Conclusion and Outlook

We presented a focus annotation study based on reading comprehension data, which we view as a contribution to the general goal of analyzing and annotating focus. Motivated by the limited success of approaches trying to tackle focus annotation from a general conceptual level, we aim to proceed from the concrete task to the more general setting. This allows us to separate a) identifying the QUD and b) determining the location and extent of the focus in the language material, where a) is informed and greatly simplified by the explicit question.

Using this approach in combination with semantically motivated annotation guidelines, we showed that focus annotation can be carried out systematically with Kappa values in the range of .61 to .69, depending on the well-formedness of the language and the number of classes distinguished.

With respect to the practical goal of improving automatic assessment of short student answers, we showed that information structural distinctions are relevant and able to quantitatively improve the results, as demonstrated by an increase from 84.6% to 90.3% accuracy in a binary classification task on a balanced data set.

While the manual annotation showcases the relevance and impact of focus annotation, we see the design of an automatic focus/background classification system on the basis of our annotated data as the logical next step. As such a system cannot perform the kind of introspective language analysis our human annotators employed, we will have to approximate focus through surface criteria such as word order, syntactic categories and focus sensitive particles. It remains to be seen how much of the potential benefit of focus annotation can be reached by automatic focus annotation using machine learning.

Finally, in order to obtain more human-annotated data, we are planning to turn focus annotation of answers to questions into a feasible crowd-sourcing task.

Acknowledgements

We are grateful to Heike Cardoso and Stefanie Wolf for carrying out the manual annotation and providing valuable feedback. We also would like to thank Kordula De Kuthy, Verena Henrich, Niels Ott and the three anonymous reviewers for their helpful comments.

References

- Ron Artstein and Massimo Poesio. 2009. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):1–42.
- Stacey Bailey and Detmar Meurers. 2008. Diagnosing meaning errors in short answers to reading comprehension questions. In Joel Tetreault, Jill Burstein, and Rachele De Felice, editors, *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*, pages 107–115, Columbus, Ohio.
- Eva Buráňová, Eva Hajičová, and Petr Sgall. 2000. Tagging of very large corpora: topic-focus articulation. In *Proceedings of the 18th conference on Computational linguistics - Volume 1*, COLING '00, pages 139–144, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Daniel Büring. 2003. On d-trees, beans, and b-accent. *Linguistics and Philosophy*, 26(5):511–545.
- Daniel Büring. 2007. Intonation, semantics and information structure. In Gillian Ramchand and Charles Reiss, editors, *The Oxford Handbook of Linguistic Interfaces*. Oxford University Press.
- Sasha Calhoun, Jean Carletta, Jason Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and 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.
- Kordula De Kuthy and Detmar Meurers. 2012. Focus projection between theory and evidence. In Sam Featherston and Britta Stollerfoth, editors, *Empirical Approaches to Linguistic Theory – Studies in Meaning and Structure*, volume 111 of *Studies in Generative Grammar*, pages 207–240. De Gruyter.
- Stefanie Dipper, Michael Götze, and Stavros Skopeteas, editors. 2007. *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics and Information Structure*, volume 7 of *Interdisciplinary Studies on Information Structure*. Universitätsverlag Potsdam, Potsdam, Germany.
- Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA, June. Association for Computational Linguistics.
- Michael Hahn and 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*, pages 94–103, Montreal.
- Andrea Horbach, Alexis Palmer, and 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*, pages 286–295, Atlanta, Georgia, USA, June. Association for Computational Linguistics.
- Ray Jackendoff. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA.
- Manfred Krifka. 1992. A compositional semantics for multiple focus constructions. In Joachim Jacobs, editor, *Informationsstruktur und Grammatik*, pages 17–54. Westdeutscher Verlag, Opladen.
- Manfred Krifka. 2001. For a structured meaning account of questions and answers. In C. Fery and W. Sternefeld, editors, *Audiatur Vox Sapientia. A Festschrift for Arnim von Stechow*, volume 52 of *studia grammatica*, pages 287–319. Akademie Verlag, Berlin.
- Manfred Krifka. 2007. Basic notions of information structure. In Caroline Fery, Gisbert Fanselow, and Manfred Krifka, editors, *The notions of information structure*, volume 6 of *Interdisciplinary Studies on Information Structure (ISIS)*, pages 13–55. Universitätsverlag Potsdam, Potsdam.
- Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. 2011. 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*, pages 1–9, Edinburgh, Scotland, UK, July. Association for Computational Linguistics.
- Michael Mohler, Razvan Bunescu, and 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*, pages 752–762, Portland, Oregon, USA, June. Association for Computational Linguistics.

- Niels Ott, Ramon Ziai, and Detmar Meurers. 2012. Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context. In Thomas Schmidt and Kai Wörner, editors, *Multilingual Corpora and Multilingual Corpus Analysis*, Hamburg Studies in Multilingualism (HSM), pages 47–69. Benjamins, Amsterdam.
- Arndt Riestler and Stefan Baumann. 2013. Focus triggers and focus types from a corpus perspective. *Dialogue & Discourse*, 4(2):215–248.
- Stefan Riezler. 2014. On the problem of theoretical terms in empirical computational linguistics. *Computational Linguistics*, 40(1):235–245.
- Julia Ritz, Stefanie Dipper, and 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*, pages 2137–2142, Marrakech, Morocco.
- Craige Roberts. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. In Jae-Hak Yoon and Andreas Kathol, editors, *OSU Working Papers in Linguistics No. 49: Papers in Semantics*. The Ohio State University.
- Mats Rooth. 1992. A theory of focus interpretation. *Natural Language Semantics*, 1(1):75–116.
- Roger Schwarzschild. 1999. GIVENness, AvoidF and other constraints on the placement of accent. *Natural Language Semantics*, 7(2):141–177.
- Arnim von Stechow. 1981. Topic, focus, and local relevance. In Wolfgang Klein and W. Levelt, editors, *Crossing the Boundaries in Linguistics*, pages 95–130. Reidel, Dordrecht.