Focus Annotation of Task-based Data: A Comparison of Expert and Crowd-Sourced Annotation in a Reading Comprehension Corpus

Kordula De Kuthy, Ramon Ziai, Detmar Meurers
Sonderforschungsbereich 833
Eberhard Karls Universität Tübingen
{kdk,rziai,dm}@sfs.uni-tuebingen.de

Abstract

While the formal pragmatic concepts in information structure, such as the *focus* of an utterance, are precisely defined in theoretical linguistics and potentially very useful in conceptual and practical terms, it has turned out to be difficult to reliably annotate such notions in corpus data (Ritz et al., 2008; Calhoun et al., 2010). We present a large-scale focus annotation effort designed to overcome this problem. Our annotation study is based on the tasked-based corpus CREG (Ott et al., 2012), which consists of answers to explicitly given reading comprehension questions. We compare focus annotation by trained annotators with a crowd-sourcing setup making use of untrained native speakers. Given the task context and an annotation process incrementally making the question form and answer type explicit, the trained annotators reach substantial agreement for focus annotation. Interestingly, the crowd-sourcing setup also supports high-quality annotation – for specific subtypes of data. Finally, we turn to the question whether the relevance of focus annotation can be extrinsically evaluated. We show that automatic short-answer assessment significantly improves for focus annotated data. The focus annotated CREG corpus is freely available and constitutes the largest such resource for German.

Keywords: information structure, focus, crowd sourcing, learner corpora, short-answer assessment

1. Introduction

The information structure of a sentence is receiving significant interest in linguistics as the attention has shifted from individual sentences to the question how the information is packaged in sentences analyzed in context. Complementing the theoretical interest, identifying information structural concepts has also been shown to be relevant in practical computational linguistic tasks such as Short Answer Assessment: While some approaches have integrated aspects of *givenness* (Bailey and Meurers, 2008; Mohler et al., 2011), more recent work (Meurers et al., 2011; Hahn and Meurers, 2012) has argued for relying on *focus* as discussed in formal pragmatics (e.g., Krifka, 2007, p. 18).

In this paper, we present a comprehensive focus annotation study based on the tasked-based corpus CREG (Ott et al., 2012), consisting of answers to explicitly given reading comprehension questions. We compare focus annotation by trained annotators with a crowd-sourcing setup making use of untrained native speakers. As a result of these annotation efforts, with this paper we provide both a substantial new corpus resource with gold standard focus annotation and conceptual insights into the nature of focus annotation in different types of contexts.

While theoretical linguists have discussed the notion of focus for decades (cf., e.g., Jackendoff, 1972; Stechow, 1981; Rooth, 1992; Schwarzschild, 1999; Büring, 2007), only few attempts at systematically identifying focus in authentic data have been made (e.g., Ritz et al., 2008; Calhoun et al., 2010) . These approaches generally were only rewarded with limited success, as they have tried to identify focus in newspaper text or other data types where no explicit questions are available, making the task of determining the Question under Discussion (QUD, Roberts, 2012), and thus reliably annotating focus, particularly difficult. Yet, many of the natural tasks and authentic data in which focus annotation would be relevant actually do contain explicit task and context information of relevance to determining focus. Building on the work presented in Ziai and Meurers (2014), we show that reliable focus annotation in authentic data is feasible, even for somewhat ill-formed learner language, if one has access to explicit questions and explicitly takes them into account in an incremental annotation scheme. We demonstrate the effectiveness of the approach by reporting both substantial inter-annotator agreement and a substantial extrinsic improvement in automatically evaluating the meaning of answers if focus/background information is integrated into the system.

Since manual focus annotation by experts is very time-consuming for large data sets, both for annotator training and the annotation itself, a second component of our work on annotating authentic data explores the use of crowd-sourcing for focus annotation. Crowd-sourcing as a way of collecting linguistically annotated data has been shown to work well for a number of tasks (cf., e.g., Finin et al., 2010; Tetreault et al., 2010; Zaidan and Callison-Burch, 2011)). We investigate how systematically the untrained crowd can identify a meaning-based linguistic notion like focus in authentic data and which characteristics of the data and context lead to consistent annotation results.

2. Data

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 American 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. The CREG-5K subset used for the present annotation study is an extended version of CREG-1032 (Meurers et al., 2011), selected using the same criteria after the overall, four year corpus collection effort was completed. The criteria include balancedness (same number of correct and incorrect answers), a minimum answer length of four tokens, and a language course level at the intermediate level or above. Both CREG-1032 and CREG-5K are characterized in terms of summary statistics in Table 1.

	CREG-1032	CREG-5K
Reading Texts	31	96
avg. Token #	318.33	974.68
Questions	177	877
avg. Token #	10.82	11.82
Q's per text	5.71	9.14
Target Answers	223	966
avg. Token #	13.24	15.89
Student Answers	**1032**	**5138**
avg. Token #	11.91	11.75
SA's per question	5.83	5.86
Question form subtype distribution		
what	25.4% (45)	26.3% (231)
which	20.3% (36)	16.0% (141)
multiple	11.3% (20)	16.0% (141)
why	13.6% (24)	14.4% (126)
how	16.4% (29)	14.4% (126)
who	6.8% (12)	4.7% (41)
where	1.7% (3)	3.6% (32)
when	1.7% (3)	2.1% (18)
yes/no	0.6% (1)	0.8% (7)
alternative	2.3% (4)	0.8% (7)
unknown	0.0% (0)	0.8% (7)

Table 1: Data set characteristics

In addition to information on the number of texts, questions and answers, along with their average length, we included the distribution of *question form subtypes*, a surface-based classification of questions mainly guided by their question word (e.g. 'what' or 'why'). These types were automatically determined using a regular-expression-based approach, and manually post-corrected, following our work in Meurers et al. (2011).

In the upper part of Table 1, one can see that CREG-1032 and CREG-5K are very similar in terms of how long the answers are and how many answers there are per question. Also, the length of the questions does not differ much between the two data sets.

A clear difference, however, seems to lie in the nature of the reading texts upon which both questions and answers are based: the reading texts of CREG-5K are more than three times longer (974.68 tokens vs. 318.33 tokens) on average than those of CREG-1032, and there are significantly more questions per text (9.14 vs. 5.71). This difference suggests a higher complexity in CREG-5K with respect to how much information requested by questions is encoded in the reading texts, an interesting characteristic of reading comprehension tasks that we plan to investigate further in the near future.

3. Expert Annotation

3.1. Annotation Scheme

We used the focus annotation scheme we previously developed in Ziai and Meurers (2014). To obtain an expert gold-standard focus annotation for the CREG-5K data set, we set out to manually annotate both target answers and student answers with focus. The annotation was performed by two graduate research assistants in linguistics using the *brat*[1] rapid annotation tool directly at 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 second author met with the annotators to discuss difficult cases and decide how the scheme would accommodate them.

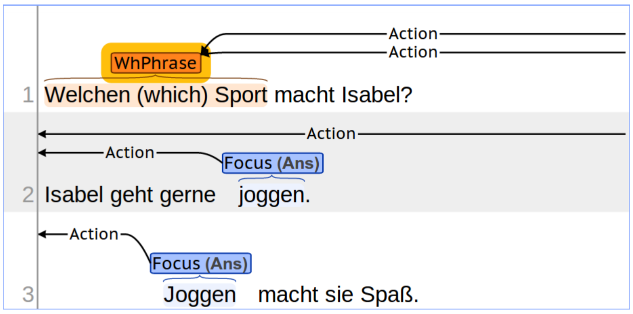
An important characteristic of our annotation scheme is that it is applied incrementally: annotators first look at the surface question form, then determine the set of alternatives (Krifka, 2007, sec. 3), and finally mark instances of the alternative set in answers.

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). Three types of categories are distinguished:

- **Question Form** encodes the surface form of a question (e.g. `WhPhrase`, `Yes/No` or `Alternative`). Question forms do not encode any semantics, but merely act as an explicit marker of the surface question form. Table 4 at the end of the paper lists all question forms and examples for them.
- **Focus** marks 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.
- **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. Examples include `Time/Date`, `Location`, `Entity`, `Action`, and `Reason`. Table 5 at the end of the paper provides a comprehensive list of Answer Types along with example answers.

Figure 1 shows a brat screen shot with an example including a `WhPhrase` Question Form and two answers, a target answer (TA) and a student answer (SA), containing a word selected as focus with Answer Type `Action`.

[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 1: Brat annotation example

To help determine the *extent* (i.e., the word span) of the focus in answers – an aspect reported to be particularly challenging by Ritz et al. (2008) and Calhoun et al. (2010) – annotators were instructed to apply a word substitution test: If the respective word is exchanged for a different one with the same POS, does the meaning change? For example in (1) and (2), substituting *in* by, e.g., *near* makes a meaning difference for (1) but not for (2).

(1) Where does Heike live?
She lives ⟦in Berlin.⟧$_F$

(2) In what city does Heike live?
She lives in ⟦Berlin.⟧$_F$

In example (1), “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”. In the same answer to a slightly different question in (2), the set of alternatives is more constrained and hence “in” is not focused.

As expected, the focus of an answer depends on its question context. The substitution test makes explicit the link between the meaning of the question and the words in the answer in a way that makes it possible to distinguish even relatively subtle differences as in the examples above.

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*.

(3) What did the pop stars wear?
The female pop stars wore caftans.

When it addresses a more general QUD, as in (4), the response is annotated as *question generalized*.

(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.

3.2. Evaluation

3.2.1. Inter-annotator agreement

The effort described in this paper builds on two annotation pilot studies (Ziai and Meurers, 2014; De Kuthy et al., 2015), where 1,255 answers (1,032 student answers and 223 target answers of CREG-1032) were annotated. Percentage agreement for focus in all answers reached 88.1%, with $\kappa = 0.75$, calculated over all answer tokens. We applied the approach to another 2,922 answers (2,155 student answers and 767 target answers, hereafter CREG-2155) of CREG-5K using two annotators and obtained a percentage agreement for focus annotation calculated over all answer tokens of 86.3%, with $\kappa = .70$. Altogether, 4,177 answers (3,187 student answers and 990 target answers) of the CREG-5K corpus are manually annotated with focus – we will refer to this corpus as CREG-ExpertFocus. The overall percentage agreement for focus in the CREG-ExpertFocus corpus reached 86.6% with κ of 0.71. Table 2 summarizes the agreement results for the three CREG data sets.

	% Agreement	κ
CREG-1032	88.1%	0.75
CREG-2155	86.3%	0.70
CREG-ExpertFocus	86.6%	0.71

Table 2: Inter-annotator agreement for focus

To obtain an expert gold standard focus annotation of the CREG-ExpertFocus corpus, the two annotation versions were subsequently merged into one focus annotation by a third person acting as judge 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.

3.2.2. Extrinsic evaluation

To independently establish the relevance and quality of the focus annotation, we extrinsically evaluated the expert gold standard annotation in an independent task, the automatic assessment of answers to reading comprehension questions. For this purpose, we employed the CoMiC system (Meurers et al., 2011), which assesses student answers by analyzing the quantity and quality of alignment links it finds between the student and the target answer.

The standard system employs a simplified notion of givenness, only aligning tokens which are not found in the question, and extracts several numeric features based on the number and kind of alignments found between non-given answer parts. For the present evaluation, we augmented this approach by adding a focus version of each feature, calculated on the basis of focused tokens instead of non-given ones. The system accuracy in leave-one-out testing is detailed in Table 3 for the three focus annotated CREG subsets.

	Standard	With Focus
CREG-1032	85.9%	88.6%
CREG-2155	82.1%	85.1%
CREG-ExpertFocus	83.2%	85.6%

Table 3: Answer classification accuracy with CoMiC

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 2.. Nevertheless, the improvement provided by focus annotation is stable across all different data sets.

Overall, the intrinsic evaluation shows that expert focus annotation is feasible given enough task context, and the extrinsic evaluation demonstrates the practical relevance of information-structural notions in computational linguistic applications.

4. Crowd Annotation

4.1. Annotation Set-up

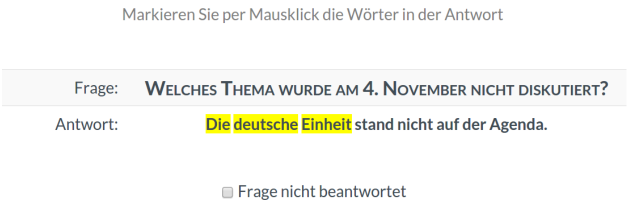
For our non-expert focus annotation study we implemented a crowd-sourcing task for the CREG-5K data set. We used the crowd-sourcing platform CrowdFlower[2] to collect focus annotations from crowd workers.

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 that easily) and it has a built-in quality control mechanism, which ensures that workers throughout the entire job maintain a certain level of accuracy.

As data for our crowd-sourcing experiment we used 5597 question-answer pairs from the CREG-5K corpus and 100 manually constructed test question-answer pairs. The crowd workers task 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 two cents 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, because, for example, the learner misunderstood the question. In the gold standard annotation described in section 3. the annotators had the option to mark these cases as “question ignored”. Since we also wanted to provide the crowd workers with this option we added a checkbox “Frage nicht beantwortet” (“*question not answered*”). If this option is selected, no word in the answer sentence can be marked as focus.

Figure 2 shows an example CrowdFlower task with the marked words in yellow. These marked words were the ones that we counted as being annotated for focus.



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: ‘$\llbracket$The German unification$\rrbracket_{\mathrm{F}}$ was not on the agenda.’

Figure 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.

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 expert annotation for the CREG data described in section 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 annotation on the word level, we opted for percentage agreement instead of Kappa or other measures that include a notion of expected agreement, for the following reasons: *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.

With the preliminaries out of the way, let us turn to comparing the percentage agreement between the two expert annotators reached for the annotation of certain types of data to the percentage agreement reached by the crowd workers compared to the gold standard annotation.

5. Comparing the Expert and the Crowd Annotation

To identify patterns that show which kinds of data can be annotated with focus most consistently by crowd workers compared to the experts, we investigated both characteristics of the answer and of the question context.

5.1. Comparison by Answer Correctness

In terms of characteristics of the answer, the CREG corpus contains teacher judgements marking for each answer whether it correctly answers the question or not. The CREG-5K corpus used as basis of our annotation experiments is balanced, i.e., it contains the same number of correct and incorrect answers. We therefore have a clean setup

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

for studying whether the correctness of the response impacts the agreement of the crowd with the expert.
Figure 3 shows the observed per-token percentage agreement for the responses correctly answering the question and for those not answering the question.

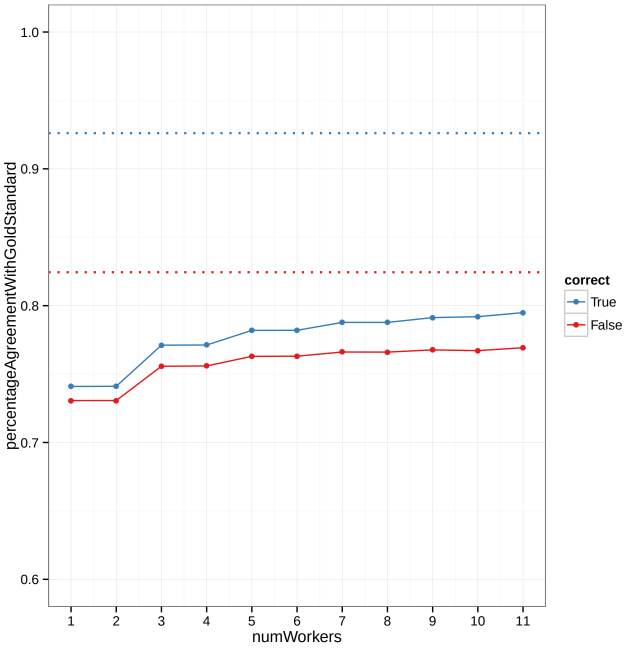


Figure 3: Percentage agreement for (in)correct answers

As reference, the two dotted lines show the percentage agreement between the two expert annotators. It shows that that the inter-expert agreement is significantly higher for focus annotation in correct answers (92%) than in incorrect answers (82%). This trend is also visible when we take a look at the agreement of the crowd-annotations with the expert gold standard, as shown by the solid lines. While the crowd annotated both, correct and incorrect answers, less consistently than the two experts, the percentage agreement between the expert gold standard and the crowd significantly improves the more crowd workers are taken into account. Interestingly, the difference between the percentage agreements for correct (79%) versus incorrect (77%) answers here is much smaller than the difference between the two expert annotators.

5.2. Comparison by Question Form Subtypes

In terms of a comparison taking characteristics of the context into account, we investigated the impact of different types of questions on annotation agreement. CREG-5K contains mostly *wh*-questions, so the general question forms distinguished by the annotation scheme introduced in section 3.1. is not specific enough. We therefore used the specific question form subtypes distinguishing the surface forms of *wh*-questions based on the annotation performed for Meurers et al. (2011). Figure 4 shows how the different question form subtypes impact the agreement for focus annotation.
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*).

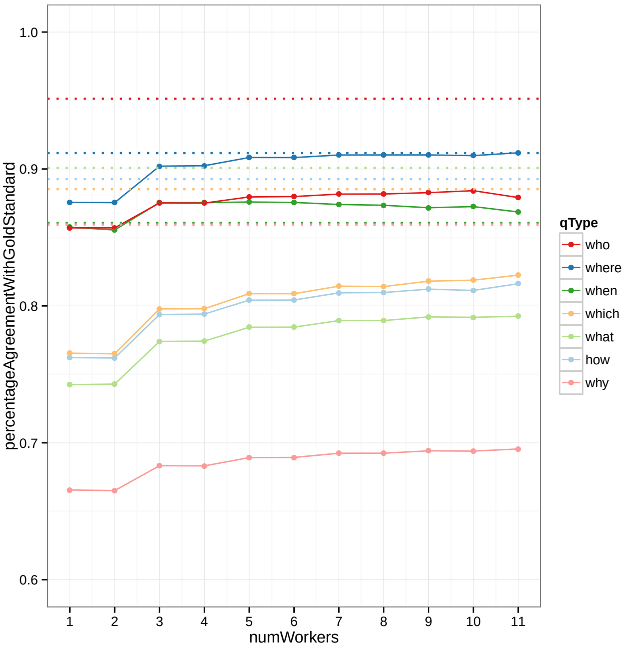


Figure 4: Percentage agreement for question form subtypes

The second group (*which*, *what* and *how*) are at 79–82% percentage agreement between crowd and gold standard annotation, which is likely due to their more ambiguous answer surface 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 69%, for which the variability in terms of answer realization is arguably the greatest, as reasons are typically realized as whole clauses instead of smaller phrasal units. 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%.
Summing up, the results of the crowd annotation study showed that a) majority voting on crowd worker judgments compared to the expert gold annotation can reach the expert level for specific cases (e.g., *where*-questions), and b) the percentage agreement improves the more crowd workers are taken into account.
With respect to the observed differences between the annotation quality of the answers to different question form subtypes, our hypothesis is that since certain examples, such as answers to *why*-questions, exhibit a much greater variation in terms of their linguistic material, this leads to less consistent results in the annotation, especially for the crowd. Since the expert annotators are trained with more explicit guidelines and are therefore possibly more aware of the variations that can occur for certain question types, this explains why the expert annotation agreement does not differ so much with respect to question form subtypes. It will therefore be interesting to study whether more explicit guidelines can also help the crowd annotators to be more systematic in their focus annotation.

6. Conclusion

We presented and analyzed a comprehensive new corpus resource for researchers interested in information structure or, more generally, the analysis of language use in context. In practical terms, we thereby contribute to the general goal of providing substantial sets of richly annotated authentic data. The annotated corpus resource is made freely available to researchers under a standard Creative Commons by-nc-sa licence – see the project web site for more information: `http://purl.org/icall/comic`

On the conceptual side, we compared two annotation approaches. We showed that large-scale focus annotation can be carried out systematically by experts with high inter-annotator agreement given an incremental annotation setup and explicit authentic task contexts. But even untrained crowd workers can identify a meaning-based linguistic notion such as focus in authentic data successfully with agreement values that are close to the values reached for the expert focus annotation – though performance was shown to vary significantly for different context types.

Finally, we provided an extrinsic evaluation for the expert focus-annotated data. The increased performance in automatic short-answer assessment confirms that information structural notions from linguistics can lead to quantitative gains in independent computational tasks. In the near future, we want to evaluate whether this also holds for crowd-sourced focus annotation, for which we plan to integrate a measure of the agreement between crowd workers.

7. References

Bailey, S. and Meurers, D. (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In Joel Tetreault, et al., 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.

Büring, D. (2003). On d-trees, beans, and b-accents. *Linguistics and Philosophy*, 26(5):511–545.

Büring, D. (2007). Intonation, semantics and information structure. In Gillian Ramchand et al., editors, *The Oxford Handbook of Linguistic Interfaces*. Oxford University Press.

Calhoun, S., Carletta, J., Brenier, J., Mayo, N., Jurafsky, D., Steedman, M., and Beaver, D. (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.

De Kuthy, K., Ziai, R., and Meurers, D. (2015). Learning what the crowd can do: A case study on focus annotation. In *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics*.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., and Dredze, M. (2010). Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 80–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hahn, M. and Meurers, D. (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.

Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA.

Krifka, M. (2007). Basic notions of information structure. In Caroline Fery, et al., editors, *The notions of information structure*, volume 6 of *Interdisciplinary Studies on Information Structure (ISIS)*, pages 13–55. Universitätsverlag Potsdam, Potsdam.

Meurers, D., Ziai, R., Ott, N., and Kopp, J. (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, July. ACL.

Mohler, M., Bunescu, R., and Mihalcea, R. (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.

Ott, N., Ziai, R., and Meurers, D. (2012). Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context. In Thomas Schmidt et al., editors, *Multilingual Corpora and Multilingual Corpus Analysis*, Hamburg Studies in Multilingualism (HSM), pages 47–69. Benjamins, Amsterdam.

Qing, C., Endriss, U., Fernandez, R., and Kruger, J. (2014). Empirical analysis of aggregation methods for collective annotation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1533–1542, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Ritz, J., Dipper, S., and Götze, M. (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.

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69, December.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, 1(1):75–116.

Schwarzschild, R. (1999). GIVENness, AvoidF and other constraints on the placement of accent. *Natural Language Semantics*, 7(2):141–177.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (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, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stechow, A. v. (1981). Topic, focus, and local relevance. In Wolfgang Klein et al., editors, *Crossing the Boundaries in Linguistics*, pages 95–130. Reidel, Dordrecht.

Tetreault, J., Filatova, E., and Chodorow, M. (2010). Rethinking grammatical error annotation and evaluation with the amazon mechanical turk. In *NAACL-HLT: 2010 Proceedings of the 5th Workshop on Building Educational Applications (BEA-5)*. Association for Computational Linguistics.

Zaidan, O. F. and Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1220–1229, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ziai, R. and Meurers, D. (2014). Focus annotation in reading comprehension data. In *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII, 2014)*, pages 159–168, Dublin, Ireland. COLING, Association for Computational Linguistics.

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 4: Question Forms in the annotation scheme with examples

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 5: Answer Types in the annotation scheme with examples