Learning what the crowd can do: A case study on focus annotation

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

I. Introduction

This paper addresses the question of how to explore and advance the conceptualization and applicability of information structural notions to support the analysis of authentic data. With this we aim at further establishing where advances in linguistic modeling also result in quantifiable gains in real-life tasks. Can, for example, computational linguistic applications be improved by integrating information structural notions? One of the necessary prerequisites for answering this question are large enough sets of data which are annotated with the relevant information structural concepts. The main problem here is that notions like focus are often discussed in theoretic literature by means of example sentences but rarely analyzed in substantial amounts of authentic data. Theoretical linguists have discussed the notion of focus for decades (cf., e.g., Jackendoff, 1972; Stechow, 1981; Rooth, 1992; Schwarzschild, 1999; Büring, 2007), while only few attempts at systematically identifying focus in authentic data have been made (e.g., Ritz et al., 2008; Calhoun et al., 2010). Most of these approaches were not rewarded with much 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, and thus reliably annotating focus, particularly difficult.

Recently, Ziai and Meurers (2014) showed that reliable focus annotation 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. They 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. However, manual focus annotation by experts is still time-consuming, both for annotator training and the annotation itself. Additionally, in computational linguistics it has been claimed (Riezler, 2014) that annotation of theoretical linguistic notions by experts is problematic and should be complemented by external grounding, either in the form of extrinsic evaluation as mentioned above, or by using crowd-sourcing: by formulating the annotation task in such a way that non-experts can understand it and carry it out, one ensures that the task does not depend on implicit knowledge shared only by a team of experts.

In this paper, we thus explore the use of crowd-sourcing, which has been shown to work well for a number of linguistic tasks (see e.g. Finin et al., 2010; Tetreault et al., 2010; Zaidan and Callison-Burch, 2011), for focus annotation. In doing so, we on the one hand contribute to theoretical linguistics by examining how systematically the untrained crowd can identify a meaning-based linguistic notion like focus and which characteristics of the data and context lead to consistent annotation results. Among other attributes, we therefore investigate how different types of questions impact the systematic identification of focus. On the other hand, we contribute to quantitative research on discourse phenomena by demonstrating a way of obtaining large amounts of focus-annotated data in a fast and cost-effective way.

II. Data

The first data set we used for our crowd-sourcing experiments is a small collection of Q/A pairs obtained using the Questionnaire for Information Structure (QUIS, Skopeteas et al., 2006), a systematic way of eliciting data for the analysis of Information Structure in a controlled fashion. We used the 40 German Q/A pairs from (Ritz et al., 2008), which were obtained by asking subjects questions about simple pictures. The data was originally recorded in spoken form and then transcribed, which occasionally led to disfluency and repetition. An example from these QUIS data is shown in (1).

(1) Q: *Was schlägt die Frau?*
what beats the woman
A: *Die Frau schlägt einen Baum.*
the woman beats a tree

The other, much larger data set we used in our crowd-sourcing experiments consists of 1032 answers from the Corpus of Reading Comprehension Exercises in German (CREG-1032, Meurers et al., 2011), which contains answers of US learners of German to reading comprehension questions and the target answers formulated by teachers. Every learner answer is rated by two annotators with respect to whether it answers the question or not (correctness), and CREG-1032 contains equal proportions of each. (2) is an example of a Q/A pair from CREG.

(2) Q: *Welches Thema wurde am 4. November nicht diskutiert?*
which topic was on the 4^{th} November not discussed
A: *Die deutsche Einheit stand nicht auf der Agenda.*
the German unity stood not on the agenda

Table I sums up the two data sets in terms of number of words, number of answers and language characteristics.

	QUIS	CREG-1032
# of answers	40	1032
# of words	288	12253
words/answer	7.2	11.9
source form	spoken	written
produced by	native speakers	learners

TABLE I. STATISTICS FOR QUIS AND CREG-1032

III. GOLD STANDARD ANNOTATION

In order to have a reference point for the evaluation of the focus annotation by crowd workers, we needed to obtain an expert annotation of reasonable quality. We used the incremental annotation scheme from Ziai and Meurers (2014), where three types of categories are distinguished (exemplified in Fig. 1):

- **Question Form** encodes the surface form of a question.
- **Focus** marks the focused words or phrases in an answer.
- **Answer Type** expresses the semantic category of the focus in relation to the question form.

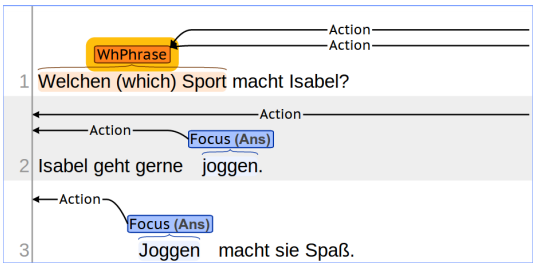

Fig. 1. Example for Question Form (WhPhrase), Focus and Answer Type (Action).

We built upon the already existing annotation experiment reported in Ziai and Meurers (2014) and applied the scheme to all of CREG-1032 using two annotators. Percentage agreement for focus in this data was 88.1%, with $\kappa = 0.75$, calculated over all answer tokens. The QUIS data were also annotated in the same fashion, with a percentage agreement of 93.9% and $\kappa = 0.87$.

IV. CROWD ANNOTATION OF QUIS DATA

A. Setup of the crowd-sourcing experiment

As a first step in our enterprise of testing whether non-experts can provide reliable focus annotation in a specific data set, we ran a crowd-sourcing experiment with the QUIS data set described in section II. We used the crowd-sourcing platform CrowdFlower[1] 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. In addition to the 40 Q/A pairs from the QUIS data we therefore designed 24 test Q/A pairs for which we defined the correct focus annotations in the answer sentences as the gold standard. The setup of our CrowdFlower experiment included the collection of 10 focus annotations per answer sentence, for which workers were paid 3.3 cents per

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

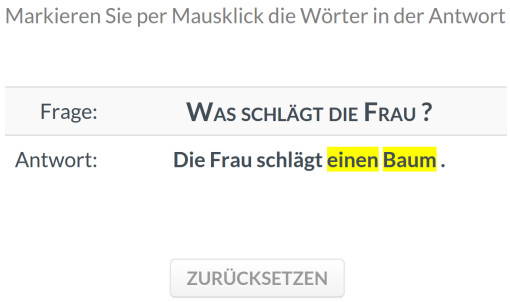

Fig. 2. Example annotation task from the QUIS data set

annotated sentence. In the instructions shown to the workers they were told that their task was to identify those words in an answer sentence that "contain the information asked for in the question". The instructions also included some examples illustrating that sometimes only one word provides the requested information, and sometimes larger chunks of the sentence do so. The actual task of the workers then was to click on those words that they wanted to mark as providing the requested information. Figure 2 shows the example from (1) as a CrowdFlower task where the marked words are highlighted in yellow. In our final evaluation, these marked words were the ones that we counted as being annotated for focus.

Workers were shown 3 Q/A pairs at a time out of which one was always from our set of hand-crafted test Q/A pairs. The workers had to maintain a minimum accuracy of 66% on these test cases throughout the entire experiment. If a worker's accuracy fell below 66% during the experiment, none of the focus annotations of that worker were collected for our end result. Altogether 607 annotated sentences were collected within 7 hours.

B. Comparing the crowd to experts

In evaluating the results of our first focus annotation experiment we wanted to find out how the annotations produced by the crowd workers compare to the gold-standard for the QUIS data described in section III. We therefore calculated all possibilities of combining 1…10 workers into one "virtual" annotator using majority voting on individual word judgments. Ties in voting were resolved by random assignment. The procedure is similar to the approach described by Snow et al. (2008).

In measuring agreement between crowd workers and the gold annotation on the word level, we opted for percentage agreement (PA) 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 not the case with crowd workers, 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.

We can now compare the annotation produced by the crowd workers to the annotation by our expert annotators. In particular, we will look at (a) whether the crowd workers converge on the expert annotation and (b) whether the performance of the annotators differs by factors like answer type or question form.

C. Results

As a first step in evaluating our crowd annotation study, we look at the resulting agreement numbers with respect to question form. We classified the questions occurring in the QUIS data set into three types: *wh*-questions, *or*-questions and *yes-no* questions. The resulting per-token PAs between the "virtual" annotator and one gold annotator are shown in Fig. 3. In this and all following figures, the PAs shown for a given number of workers are averages of all worker combinations with that number.

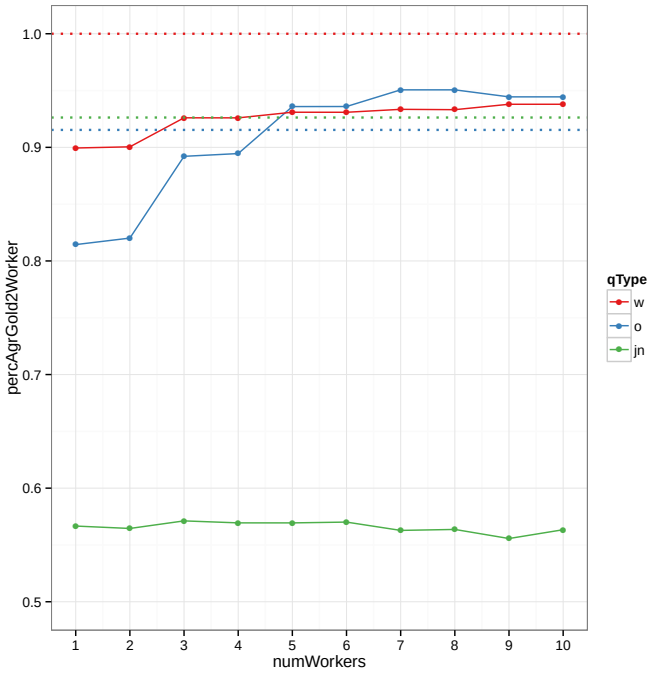

Fig. 3. Percentage agreement between gold annotation and crowd worker

The dotted lines always show the PAs for the two gold annotators. The performance of the crowd workers compared to the gold annotation clearly differs depending on the question type: The PA of the "virtual" crowd worker and one gold annotator for *yes/no*-questions was always far below that between the two gold annotators (under 60%), while the PA for *wh*-questions is high independent of the number of workers taken into account (above 90%). For *or*-questions, the PA between five crowd workers or more and one gold annotator is even higher than that between the two gold annotators (95% vs. 92% for the gold annotators).

V. Crowd annotation of CREG-1032 data

A. Setup

The results of our first annotation experiment showed that crowd workers can provide focus annotations that reach a level close to expert annotations for certain types of data. As a next step we tested whether similar results can also be obtained for more diverse data as they occur in the CREG-1032 corpus. We used the almost identical setup as in our first experiment with one addition: Since CREG-1032 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 III 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 "question not answered". If this option is selected, no word in the answer sentence can be marked as focus. The instructions were modified accordingly to explain this option. Figure 4 shows the example from (2) as a CrowdFlower task with the marked words in yellow and the added checkbox.

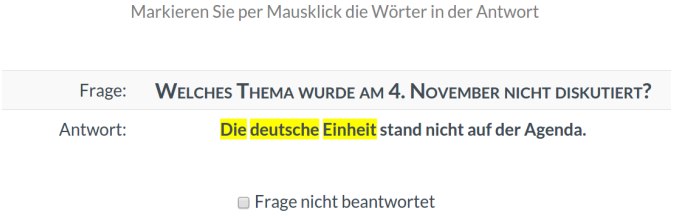

Fig. 4. Example annotation task from the CREG-1032 data set

Our data set consisted of 1087 Q/A pairs from the CREG-1032 corpus and 34 manually constructed test Q/A pairs. The discrepancy from the number 1032 stems from two changes we made in preparing the CREG data for crowd annotation: 1) all questions containing multiple sub-questions were removed and 2) for the remaining questions, we used all available target answers, not only the ones for the original 1032 answers. Similar to the previous experiment we collected 10 focus annotations per answer sentence and crowd workers had to maintain an accuracy of 60% on the test Q/A pairs. Altogether we collected 10907 annotated sentences within 16 hours.

B. Results

In the evaluation of the focus annotation of the CREG-1032 data we wanted to find out how the annotations produced by the crowd workers compare to the gold-standard for the CREG-1032 data described in section III. We again compared one "virtual annotator" resulting from all possible combinations of 1…10 workers to one gold annotator measuring per-token percentage agreement.

In trying to identify patterns that show which kinds of data can be annotated with focus most consistently by crowd workers we investigated characteristics that are specific to learner data as contained in the CREG-1032 data set. As mentioned in section II, one potentially interesting distinction of the learner answers is their correctness: all student answers in the CREG-1032 are rated with respect to whether they answer the question or not, and the corpus is balanced, i.e. it contains the same number of correct and incorrect answers.

The per-token PAs distinguished by the two types of correctness occurring in the CREG-1032 data are shown in Fig. 5. The two dotted lines show, that the PA between the two gold annotators is much higher for focus annotation in correct answers (95%) than in incorrect answers (84%). An interesting pattern can be seen for the PA between the crowd workers and one gold annotator: the agreement is much lower for incorrect answers than for correct answers when only comparing one

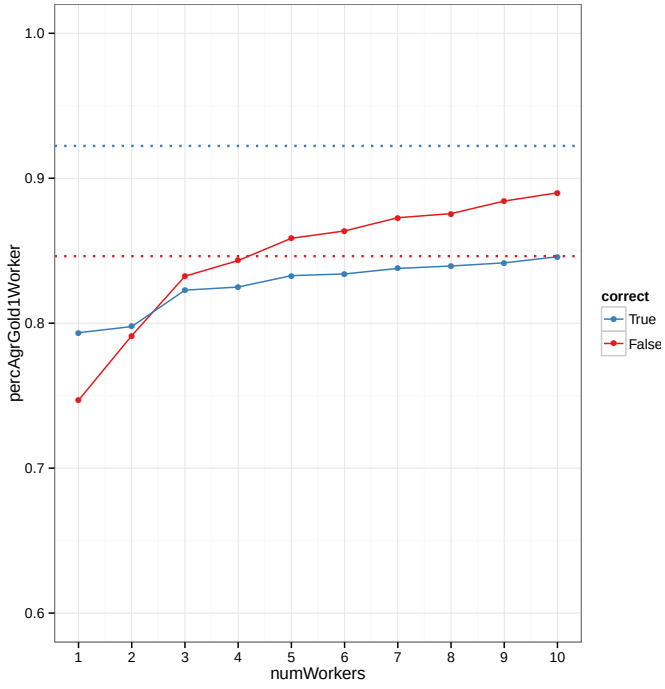

Fig. 5. Percentage agreement for correct/incorrect answers

worker at a time with the gold annotator. For four or more workers taken together, however, the PA is even higher than that of the two gold annotators.

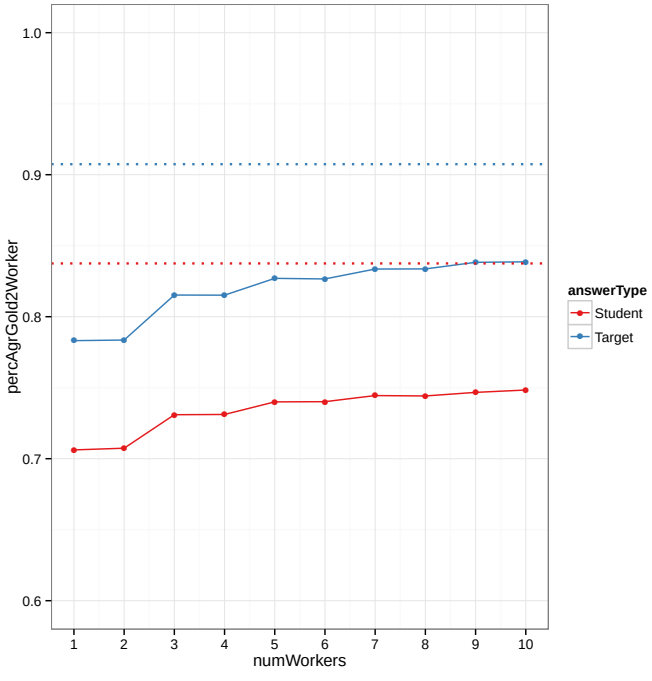

Fig. 6. Percentage agreement for student/target answer

Another interesting characteristic of the CREG data that can potentially make a difference in focus annotation is language well-formedness: Ziai and Meurers (2014) report that expert agreement for the more well-formed target answers is generally higher than for the often ill-formed student answers, so we investigated whether this is also the case for crowdsourcing annotation, as shown in Fig. 6. One can see that the trend observed in expert annotation with respect to the student/target distinction is also visible in crowd-sourcing: student answers are harder to annotate (PA at 74%) than target answers (PA at 83%), most likely due to their much higher potential for ill-formed language. The gap in consistency between crowd and expert annotation is also again visible in the fact that worker-gold agreement on target answers only reaches the level where gold annotators agree on student answers.

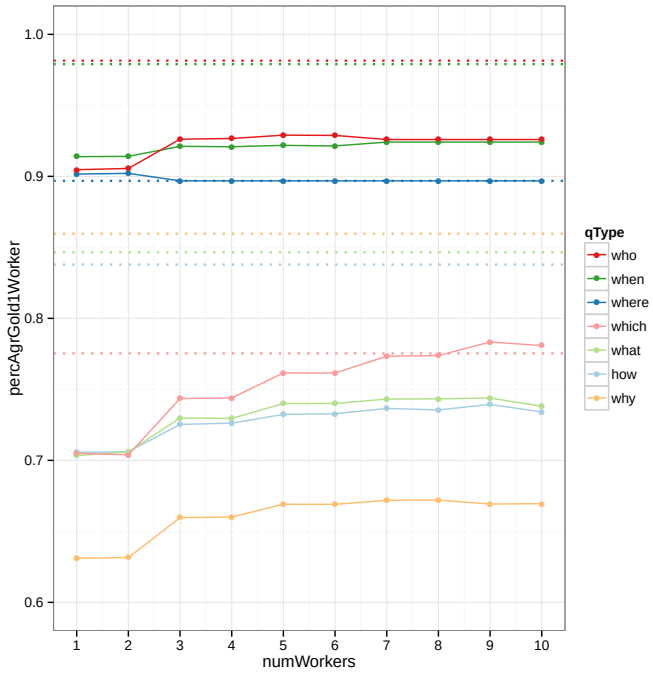

Fig. 7. Percentage agreement for question forms

Finally, in a similar manner to the results on the QUIS data set, we also investigated the impact of different question types on annotation agreement. Because CREG-1032 is a far larger data set and contains mostly *wh*-questions, we were able to distinguish several surface forms of *wh*-questions using the annotation done by Meurers et al. (2011). Fig. 7 shows the impact of this distinction on agreement. 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 PAs around or above 90%. The second group (*which*, *what* and *how*) are at 74–78%, 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 67%, 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 annotators 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%.

C. Discussion

To investigate why the annotation agreement differs so much with respect to question types for the crowd annotators, we now take a closer look at the variation in the linguistic material that can impact focus annotation. We discuss a typical example for a *who*-question (3) and a *why*-question (4) together with a sample of given answers from the CREG-1032 data set as the two most extreme cases with respect to the observed annotation agreement.

In the case of the answers to the *who*-question in (3), we can see that the variation both in meaning and form is very limited:

(3) 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 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.

In the case of the answers to the *why*-question in (4), multiple ways of answering the same questions can be observed, both syntactically and semantically.

(4) 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 und Beschattung.*⟧$_F$
a air computer regulates temperature ventilation humidity and shading
A2: *Das Haus der Kamelie ist so interessant,* ⟦*weil es 230 Jahre alt und 8,90 m hohe ist.*⟧$_F$
the house of the camellia is so interesting because it 230 years old and 8.90 m high is
A3: ⟦*In der warmen Jahreszeit wird das Haus neben die Kamelie gerollt.*⟧$_F$
in the warm season is the house next to the camellia rolled
A4: *Das Haus der Kamelie ist so interessant,* ⟦*weil es ist ein fahrbares Haus.*⟧$_F$
the house of the camellia is so interesting because it is a mobile house
A5: *Der Kamelie ist interessant* ⟦*wegen des Computers.*⟧$_F$
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 (4), the subordinate clause starting with *weil* (because) as in A2 and A4 in (4), or as a PP introduced by *wegen* (because of) as in A5. Semantically, all four answers present a different propositional content.

Our hypothesis is that the greater variation in examples such as (4) 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 types.

It would therefore be interesting to study whether more explicit guidelines could also help the crowd annotators to be more systematic in their focus annotation.

VI. Conclusion

We set out to explore how broader samples of authentic data can be successfully annotated with information structural concepts, with a particular interest in how to reliably annotate focus. Since manual focus annotation by experts is very time-consuming, we conducted two crowd-sourcing experiments in order to explore whether the much more time-efficient focus annotation by a large, but untrained crowd of annotators provides comparable results to an expert annotation.

The results of our first annotation experiment with a relatively small data set (40 Q/A pairs from QUIS) showed that: a) majority voting on crowd worker judgments compared to an expert annotation can reach expert level for specific cases (e.g. *or*-questions), b) even individual crowd workers can reliably identify focus for simple *wh*-cases and c) the crowd cannot handle *yes-no*-questions (or requires better instructions). The results of our second annotations experiment with a much larger data set (1087 Q/A-pairs from CREG-1032) showed similar results with respect to different question types (distinguished by the surface form of the question word): a) the PA between crowd workers and an expert reaches expert level agreement for specific cases (*who*-, *when*-, and *where*-questions), b) the agreement can reach expert level when the annotations of a larger number of crowd workers is taken into account, and c) the crowd cannot handle *why*-questions (which again probably require better instructions). Interesting patterns also emerged with respect to learner-data specific properties: correct student answers were annotated with a higher PA by the expert annotators than the incorrect ones, while four or more crowd workers compared to an expert reached a higher PA on the incorrect student answers than the two expert annotators. This is thus a case where focus annotation by a larger number of non-experts provides a more reliable result than the annotation by two experts.

Summing up, our study on crowd-sourcing focus annotation has shown that this type of non-expert annotation is promising a) for exploring the impact of different types of data and instructions on the annotation of authentic data and b) for the large-scale annotation of some types of data. Further research needs to explore how to improve the focus annotation for those types of data where the non-expert annotations did not reach the level of the expert annotations. This is a necessary step in order to obtain reliably annotated larger sets of data that can help to test whether computational linguistic applications can benefit from integrating information structural notions.

References

Büring, D. (2007). Intonation, semantics and information structure. In Ramchand, G. and Reiss, C., 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.

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.

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

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. ACL.

Riezler, S. (2014). On the problem of theoretical terms in empirical computational linguistics. *Computational Linguistics*, 40(1):235–245.

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.

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.

Skopeteas, S., Fiedler, I., Hellmuth, S., Schwarz, A., Stoel, R., Fanselow, G., Féry, C., and Krifka, M. (2006). *Questionnaire on information structure (QUIS): reference manual*, volume 4 of *Interdisciplinary studies on information structure (ISIS)*. Universitätsverlag Potsdam.

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 Klein, W. and Levelt, W., 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.