Focus Annotation of Task-based Data: Establishing the Quality of Crowd Annotation

Kordula De Kuthy **Ramon Ziai** **Detmar Meurers**
Collaborative Research Center 833
University of Tübingen
{kdk,rziai,dm}@sfs.uni-tuebingen.de

Abstract

We explore the annotation of information structure in German and compare the quality of expert annotation with crowd-sourced annotation taking into account the cost of reaching crowd consensus.

Concretely, we discuss a crowd-sourcing effort annotating *focus* in a task-based corpus of German containing reading comprehension questions and answers. Against the backdrop of a gold standard reference resulting from adjudicated expert annotation, we evaluate a crowd sourcing experiment using majority voting to determine a baseline performance. To refine the crowd-sourcing setup, we introduce the Consensus Cost as a measure of agreement within the crowd. We investigate the usefulness of Consensus Cost as a measure of crowd annotation quality both intrinsically, in relation to the expert gold standard, and extrinsically, by integrating focus annotation information into a system performing Short Answer Assessment taking into account the Consensus Cost.

We find that low Consensus Cost in crowd sourcing indicates high quality, though high cost does not necessarily indicate low accuracy but increased variability. Overall, taking Consensus Cost into account improves both intrinsic and extrinsic evaluation measures.

1 Introduction

This paper addresses the question of how to explore and evaluate the annotation of information structural concepts to support the analysis of authentic data. 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).

Theoretical linguists have discussed the notion of focus for decades (cf., e.g., Jackendoff 1972; Stechow 1981; Rooth 1992; Schwarzschild 1999; Büring 2007). Following the work of Rooth (1992), one of the widely used definitions of focus is that "Focus indicates the presence of alternatives that are relevant for the linguistic expressions" (cf. Krifka 2007). Which part of an utterance is in the focus thus depends on the context of the utterance, as illustrated by the question-answers pairs in examples (1) and (2).

(1) A: *What did John show Mary?*
B: *John showed Mary* $[\![$*the PICTures*$]\!]_F$.

(2) A: *Who did John show the pictures?*
B: *John showed* $[\![$*MARy*$]\!]_F$ *the pictures.*

Since focus is signalled by prosodic prominence in an intonation language like English, the answers also show different prominence patterns, as indicated by the pitch accents on *picture* in (1) and *Mary* in (2).

The linguistic discussions of focus phenomena generally are based on few example sentences, without an apparent exploration of substantial amounts of authentic data. Only few attempts at systematically identifying focus in authentic data have been made (Ritz et al., 2008; Calhoun et al., 2010). They generally ran into significant problems trying to reach good inter-annotator agreement, as they tried to identify focus in newspaper text or other data types where no explicit questions are available, making the task of determining the question under discussion, and thus reliably annotating focus, particularly difficult.

Proceedings of LAW X – The 10th Linguistic Annotation Workshop, pages 110–119,
Berlin, Germany, August 11, 2016.

More 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 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 resulting from integration of focus information into a Short Answer Assessment system.

However, manual focus annotation by experts is time consuming, both for annotator training and the annotation itself. Additionally, in computational linguistics it has been argued (Riezler, 2014) that annotation of theoretical linguistic notions by experts should be complemented by external grounding, either in the form of extrinsic evaluation, as reported above, or by using crowdsourcing: 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 explore the use of crowdsourcing – 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. 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.

Having established the general feasibility of non-expert focus annotation, we refine the crowdsourcing approach by taking into account the variability within the set of crowd judgements. The approach is based on the idea that sentences with little variation in the annotation provided by the crowd are more reliably annotated, i.e., are of a higher quality. We spell out a measure of crowd diversity, Consensus Cost, and investigate its usefulness both intrinsically, by relating it to the expert-based gold-standard, and extrinsically, by integrating cost-based focus annotation data in a Short Answer Assessment system.

2 Data

We base our work on the CREG corpus (Ott et al., 2012), a freely available 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. Each answer was rated by two annotators with respect to whether it is a correct (appropriate) answer or not. 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 (equal number of correct and incorrect answers), a minimum answer length of four tokens, and a language course level at the intermediate level or above.

(3) provides an example of a question-answer pair from the CREG corpus.

(3) Q: *Welches Thema wurde am 4. November nicht diskutiert?*
which topic was on the 4^{th} November not discussed
'Which topic was not discussed on Nov. 4th?'

A: *Die deutsche Einheit stand nicht auf der Agenda.*
the German unity stood not on the agenda
'The German unification was not on the agenda.'

2.1 Gold Standard Annotation

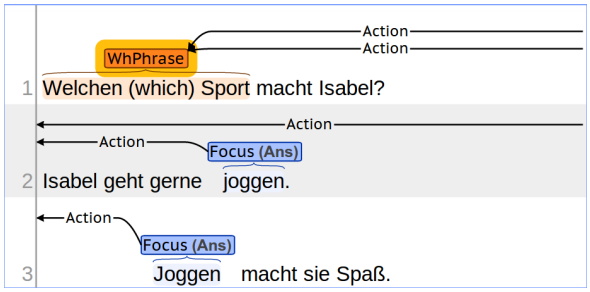
As a reference point for the evaluation of the focus annotation by crowd workers, we first obtained a gold-standard annotation using experts. We based this effort on the focus annotation scheme and annotation of the CREG-1032 data set provided in Ziai and Meurers (2014). We extended this by manually focus-annotating both target answers and student answers in the larger CREG-5K data set. The annotation was performed by two graduate research assistants in linguistics using the *brat*[1] rapid annotation tool directly at token level. An important characteristic of the 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 following three types of categories are distinguished:

- **Question Form** encodes the surface form of a question (e.g., `WhPhrase`, `Yes/No` or `Alternative`).

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

- **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. Examples include `Time/Date`, `Location`, `Entity`, `Action`, and `Reason`.

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`.



Q: '*Which sport* does Isabel do?'

TA: 'She likes to go $[\![\text{jogging}]\!]_F$.'

SA: '$[\![\text{Jogging}]\!]_F$ is fun for her.'

Figure 1: Brat annotation example

In the following we will only evaluate the agreement results for the category *Focus* of our annotation scheme. Ziai and Meurers (2014) annotated 1,255 answers (1,032 student answers and 223 target answers of CREG-1032) and reported 88.1% percentage agreement for focus in all answers, 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) of CREG-5K using two annotators and obtained a percentage agreement for focus annotation calculated over all answer tokens of 86.3%, with $\kappa = .70$, demonstrating the robustness of the annotation approach when applied to new data. Altogether, 4,177 answers (3,187 student answers and 990 target answers) of the CREG-5K corpus are manually annotated with focus. The overall percentage agreement for focus is 86.6% with a κ of 0.71.

To obtain the gold standard focus annotation of the combined corpus, the two annotation versions were merged into one focus annotation by a third expert, who determined the annotation in case the two annotators disagreed.

3 Crowd Annotation

3.1 Setup of the crowd-sourcing experiment

To study non-expert focus annotation, we implemented a crowd-sourcing task using the crowd-sourcing platform CrowdFlower[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 as transparently, and it has a built-in quality control mechanism ensuring that workers throughout the entire job maintain a certain level of accuracy on interspersed test items.

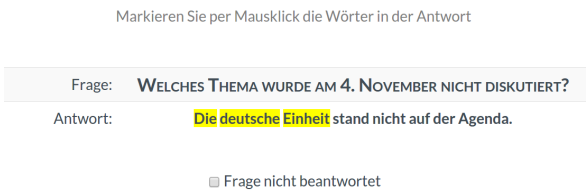
As data for our crowd-sourcing experiment, we used 5,597 question-answer pairs from the CREG-5K corpus and 100 manually constructed test question-answer pairs. The task of the crowd workers was to mark those words in an answer sentence that "contain the information asked for in the question". Workers were shown five question-answer pairs at a time. One of those five was from our set of hand-crafted test question-answer pairs. The workers were paid 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, for example, when the learner misunderstood the question. In the gold standard annotation described in section 2.1, the annotators had the option to mark such cases as "question ignored". Since we also wanted to provide the crowd workers with this option, we included a checkbox "Frage nicht beantwortet" ("*question not answered*"). When this option is selected, no word in the answer sentence can be marked as focus.

Figure 2 shows an example CrowdFlower task with the marked words in yellow. These marked words are the ones that we counted as focus. The English translation shown below was not part of the CrowdFlower task.

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.

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



Q: 'Which topic was not discussed on November 4th?'
A: '$[\![$The German unification$]\!]_F$ was not on the agenda.'

Figure 2: Example CrowdFlower annotation task

3.2 Evaluation

To evaluate the quality of our crowd focus annotation, we wanted to find out how the annotations produced by the crowd workers compare to the gold standard expert annotation described in section 2.1. We therefore chose to calculate all possibilities of combining one through eleven workers into one "virtual" annotator using majority voting on individual word judgments. Ties in voting are resolved by random assignment. The procedure is similar to the approach described by Snow et al. (2008). We did not employ any bias correction or other types of weighting schemes, as discussed, e.g., by Qing et al. (2014), but plan to do so in future research.

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

3.2.1 Overall agreement of crowd with gold

We performed the evaluation on the CREG-5K data subset for which we obtained both expert and crowd annotations. Figure 3 shows the observed per-token percentage agreement reached by the crowd workers compared to the gold standard annotation.

As reference, the dotted lines show the percentage agreement between the two expert annotators. We see that the quality improves from 74.9% for one worker to 79.8% for eleven workers[3]. Given

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

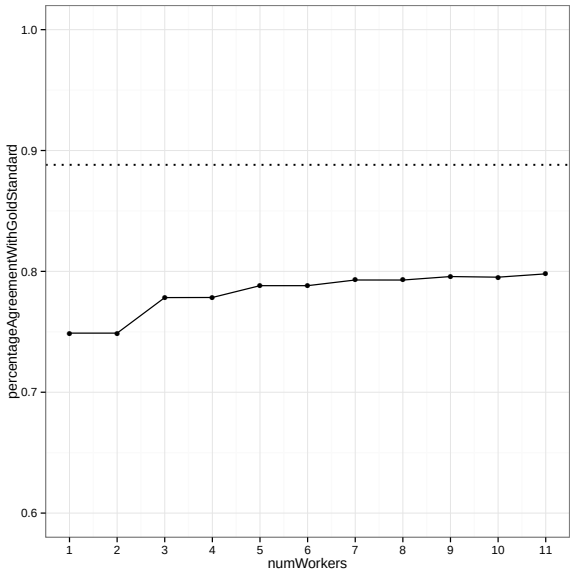


Figure 3: Agreement of crowd with gold standard

that this is below the agreement of 88.8% reached by the expert annotators for this data set, we next investigated which cases the crowd can handle, and which ones turn out to be difficult for the non-experts.

3.2.2 Evaluation for different question forms

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

We carried out the comparison for the specific question form subtypes distinguishing surface forms of *wh*-questions as annotated in CREG (Meurers et al., 2011). Figure 4 shows how the different question form subtypes impact the agreement between the crowd and the gold-standard focus annotation.

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

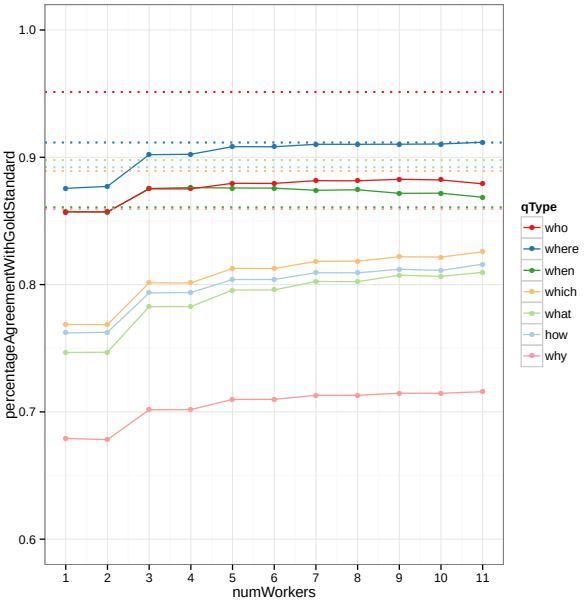


Figure 4: Agreement by question form

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

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

To test whether more explicit guidelines could also help the crowd annotators to be more systematic in their focus annotation, we conducted a small additional crowdsourcing annotation study with a smaller data set only containing answers to *why* and *what*-questions. While the general set up was the same as described in section 3.1, we provided the crowd workers with more examples illustrating focus in different kind of answers. The result was only a small improvement in agreement between crowd and gold standard annotation, with answers to *what*-questions 1% higher than before, and 2% higher for *why*-questions. Even more explicit guidelines thus do not seem to help the non-experts to handle answers occurring with *why*-questions when annotating focus.

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

3.2.3 Qualitative discussion

To gain a better understanding of why the annotation agreement differs so widely with respect to question types for the crowd annotators, we take a closer look at the variation in the linguistic material that apparently impacts focus annotation. We discuss a typical example for a *who*-question (4) and a *why*-question (5) together with a sample of given answers from the CREG-5K data set as the two most extreme cases with respect to the observed annotation agreement.

In the case of the different answers to the *who*-question shown in (4), we can see that the variation both in meaning and form is very limited:

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

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

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

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

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

Syntactically, the focused part of the answers shown in ⟦...⟧$_F$ is expressed as a nominal phrase. Contentwise, the same type of entity (a person) is expressed by semantically related words. The rest of the sentence shows no variation at all. The only inconsistency in annotation by the crowd occurred with NPs such as *Die drei Männer* in answer A3 in (4), where some of the crowd annotated the entire NP as the focus, while the rest of the crowd annotators only marked *drei Männer* as the focus, leaving out the definite article.

In the case of the various answers to the *why*-question shown in (5), multiple ways of answering the same questions can be observed, both syntactically and semantically.

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

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

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

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

A4: *Das Haus der Kamelie ist so interessant,*
the house of the camellia is so interesting
⟦*weil es ist ein fahrbares Haus.*⟧$_F$
because it is a mobile house

A5: *Der Kamelie ist interessant* ⟦*wegen des*
the camellia is interesting because of the
Computers.⟧$_F$
computer

Syntactically, the focused part of the answer is either expressed as the entire sentence as in A1 and A3 in (5), the subordinate clause starting with *weil* (because) as in A2 and A4 in (5), or as a PP introduced by *wegen* (because of) as in A5. Semantically, all four answers present a different propositional content. The relation between the question and potential answers thus is not particularly obvious or direct. Establishing the relation between question and answer – as needed to identify the focus of the answer – thus requires more effort by the annotator. This leads to less consistent results in the annotation for the crowd. For example, parts of the crowd annotators did not interpret the sentence A3 in (5) as an answer to the *why*-question in (5) at all and consequently did not mark any words in that sentence as focus, while the rest of the crowd annotators marked the entire clause as the focus.

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

4 Predicting when the crowd is reliable

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

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

4.1 Calculating the cost of crowd consensus

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

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

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

Figure 5 exemplifies the calculation of the Consensus Cost for the actual eleven crowd annotations from the crowdsourcing experiment for the short example answer *Die/the drei/three Männer/men war/was an/at der/the Tür/door* from our CREG data.

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

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

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

Figure 5: Calculating the Consensus Cost

of 0. The resulting Consensus Cost for the focus annotation for this sentence according to our formula is 0.11.

Since not all crowd workers perform equally well, it would in principle make sense to incorporate their individual reliability. As a first step towards this idea, we are excluding all workers from annotation who fail to reach a particular accuracy threshold (0.66) on the test questions.

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

4.2 Consensus Cost and Annotation Quality

In order to determine whether Consensus Cost can function as a proxy for annotation quality, let us compare it to the agreement of the crowd workers with the gold standard expert annotation we discussed in section 3.2.

To explore the relation between Consensus Cost and quality of the annotation of an answer, we divided the possible values (0.0 to 1.0) of Consensus Cost into four ranges, using 0.25, 0.5 and 0.75 as boundaries. Figure 6 shows the boxplots for each of the four groups of answers by Consensus Cost, with the percentage agreement with the gold standard shown on the y-axis. The width of the box plots indicates the number of instances represented, whereas the height represents the distribution of agreement values.

For answers annotated with low Consensus Cost (< 0.5), the quality of annotation is generally high, with agreement with the gold standard between 0.7 and 1.0. The majority of data points fall into this interval. Interestingly, answers annotated

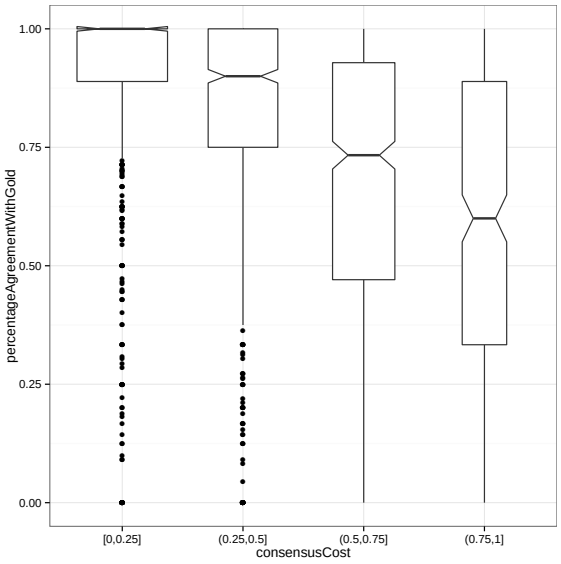


Figure 6: Consensus Cost and Annotation Quality

with higher Consensus Cost values, in the intervals (0.5,0.75] and (0.75,1], show a more heterogeneous picture. While their median agreement is much lower, they also show a more varied distribution, including some high quality annotations.

In sum, we can conclude that there is a clear association between Consensus Cost and annotation quality. A low Consensus Cost can serve as a proxy for high annotation quality. The relationship is not a simple linear one, though, so that some annotations with high Consensus Cost may also be of high quality.

4.3 Consensus Costs by Question Type

When we evaluated the quality of the crowd focus annotation in relation to the gold-standard expert annotation in section 3.2, we found that the crowd annotations fall into three groups with respect to question types: Answers to the *who*, *when* and *where* questions showed a high percentage agreement with the expert annotation, answers to *which*, *what* and *how* questions had a much lower percentage agreement and answers to *why* questions were the most difficult ones for the crowd and had the lowest agreement numbers. The data by question type thus makes an interesting test case for Consensus Cost as a proxy for annotation quality. If sentences with a low consensus cost provide annotation of higher quality, we should be able to find a similar division of the annotation in terms of question types as as in comparison with the expert annotation.

Figure 7 shows the consensus cost of our crowd annotation plotted according to question types.

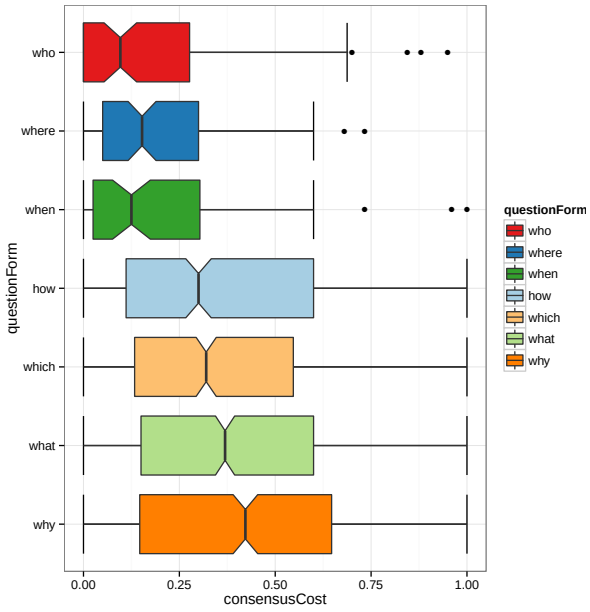


Figure 7: Consensus Cost per Question Type

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

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

4.4 Extrinsic evaluation

To externally establish the relevance and quality of the crowd focus annotation, we extrinsically evaluated the expert gold standard annotation in an independent task, Short Answer Assessment, specifically 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.

Our goal here is twofold: on the one hand, we want to find out whether the previously introduced Consensus Cost measure is helpful in determining the quality of focus annotation as measured by its impact on Short Answer Assessment. On the other hand, it is interesting to determine whether the state of the art in automatic answer assessment can be advanced by integrating non-expert annotation of focus (as a step towards automatic focus annotation developed using the crowd-annotated data).

To cleanly separate the data used for testing the Answer Assessment system CoMiC from the data used for training CoMiC, we randomly sampled approximately 20% of the CREG-5K data set and set it aside as the final test set. The remaining 80% was used as training set.

In exploring the impact of different Consensus Costs, we used the same four cutoffs as before: 0.25, 0.5, 0.75 and the maximum value 1.0. For each cutoff, we picked the answers with crowd focus annotations satisfying the cutoff constraint in training and test set, and ran CoMiC on the resulting data excerpt, aligning only words in student and target answer that are focused. For the rest of the data, which did not meet the Consensus Cost criterion or for which no focus annotation was available, we used the standard version of CoMiC that only aligns words not previously mentioned in the question. We then calculated a weighted average (by number of test instances) of both system accuracies in order to arrive at an overall system result for the respective Consensus Cost value. The results are displayed in Table 1.

Cost	Focus		Given		Avg
$\leq$	train/test	%	train/test	%	%
base	–		4136/1001	81.5	81.5
0.25	1009/252	88.1	3127/749	80.4	82.3
0.5	2019/489	84.5	2117/512	80.7	82.5
0.75	3087/747	84.5	1049/254	79.5	**83.2**
1.0	3638/882	82.7	498/119	76.5	81.9

Table 1: Results on the “unseen answers” test set

The ‘train/test’ column shows the number of training and test instances each system was run on, and the ‘%’ column shows the classification accuracy achieved. The ‘base’ row gives the baseline resulting from using CoMiC as-is, without any focus information.

Looking at the results for the focus partition of the data, one can see that accuracy drops when taking into account focus annotation with higher Consensus Cost, even though thereby in principle

more training data is becoming available.

For the 'Given' column, when data with higher Consensus Cost is used for the 'Focus' version of the system and thereby less data is available for training the 'Given' system, accuracy of the latter decreases.

Overall, a Consensus Cost cutoff of 0.75 gives the optimal trade-off between both system variants, yielding 83.2% classification accuracy.

Test with answers to unseen questions In a second experiment, we also compiled a question-based train/test split, meaning that for approximately 20% of randomly picked questions in CREG-5K, all answers were held out as the test set. This is a much harder benchmark since the system in the test has to classify answers to previously unseen questions, providing some indication of the system's ability to learn something general rather than about specific question-answer pairs. The remainder of the testing procedure was the same as described above, yielding the results detailed in Table 2.

Cost ≤	Focus train/test	%	Given train/test	%	Avg %
base	–		4016/1121	78.8	78.8
0.25	970/291	81.4	3046/830	78.2	79.0
0.5	1938/570	80.4	2078/551	78.2	79.3
0.75	2973/861	81.6	1043/260	76.9	**80.6**
1.0	3515/1005	79.6	501/116	78.4	79.5

Table 2: Results on the "unseen questions" test set

The accuracies are generally lower due to the harder test scenario. Moreover, the clear trends observed above with regard to training and test size do not seem to apply as clearly here, likely again owing to the 'unseen questions' scenario. Given the many different types of potential questions and the relatively small number of different questions the system sees during training, it is more important for which questions the system has seen answers, than how many. However, despite the differences to the previous experiment, the optimal result is again achieved with a Consensus Cost of 0.75, supporting the conclusion that Consensus Cost supports a systematic characterization of annotation quality.

5 Conclusion

We described a crowd-sourcing experiment for the annotation of focus, establishing its success both intrinsically by comparing it to a gold-standard expert annotation, and extrinsically by using the resulting annotations successfully in an independent CL task, Short Answer Assessment.

In order to distinguish between high and low quality crowd annotations, we define the measure of Consensus Cost, which essentially is the number of minority votes for each markable. We show that low values of Consensus Cost indicate high annotation quality and that training data selection based on Consensus Cost is beneficial in the Short Answer Assessment task.

In the future, we plan to extend our assessment of annotation quality beyond simple Consensus Cost cut-offs to a supervised machine-learning approach that can also take other characteristics of the authentic data (e.g., the question type) into account. The relationship between Consensus Cost and annotation quality is not simply linear and the additional information could help determine which of the more variable-quality data with high Consensus Cost is of high quality.

References

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.

Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 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*. Association for Computational Linguistics, Stroudsburg, PA, USA, CSLDAMT '10, pages 80–88.

Simone Hantke, Erik Marchi, and Björn Schuller. 2016. Introducing the weighted trustability evaluator for crowdsourcing exemplified by speaker likability classification. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene

Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '09, pages 27–35.

Ray Jackendoff. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA.

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

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*. ACL, Edinburgh, pages 1–9.

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*, Benjamins, Amsterdam, Hamburg Studies in Multilingualism (HSM), pages 47–69.

Ciyang Qing, Ulle Endriss, Raquel Fernandez, and Justin Kruger. 2014. Empirical analysis of aggregation methods for collective annotation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pages 1533–1542.

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*. Marrakech, Morocco, pages 2137–2142.

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.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '08, pages 254–263.

Arnim von Stechow. 1981. Topic, focus, and local relevance. In Wolfgang Klein and W. Levelt, editors, *Crossing the Boundaries in Linguistics*, Reidel, Dordrecht, pages 95–130.

Joel Tetreault, Elena Filatova, and Martin Chodorow. 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.

Omar F. Zaidan and Chris Callison-Burch. 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*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pages 1220–1229.

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