Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2148–2156
Marseille, 20-25 June 2022

Design Choices in Crowdsourcing Discourse Relation Annotations: The Effect of Worker Selection and Training

Merel C.J. Scholman[a], Valentina Pyatkin[b], Frances Yung[a],
Ido Dagan[b], Reut Tsarfaty[b], Vera Demberg[a]
[a]Saarland University, [b]Bar Ilan University
Saarbrücken, Germany; Ramat Gan, Israel
{m.c.j.scholman,frances,vera}@coli.uni-saarland.de
{pyatkiv,reut.tsarfaty}@biu.ac.il; dagan@cs.biu.ac.il

Abstract

Obtaining linguistic annotation from novice crowdworkers is far from trivial. A case in point is the annotation of discourse relations, which is a complicated task. Recent methods have obtained promising results by extracting relation labels from either discourse connectives (DCs) or question-answer (QA) pairs that participants provide. The current contribution studies the effect of worker selection and training on the agreement on implicit relation labels between workers and gold labels, for both the DC and the QA method. In Study 1, workers were not specifically selected or trained, and the results show that there is much room for improvement. Study 2 shows that a combination of selection and training does lead to improved results, but the method is cost- and time-intensive. Study 3 shows that a selection-only approach is a viable alternative; it results in annotations of comparable quality compared to annotations from trained participants. The results generalized over both the DC and QA method and therefore indicate that a selection-only approach could also be effective for other crowdsourced discourse annotation tasks.

Keywords: discourse annotations, crowdsourcing, training, participant selection

1. Introduction

Obtaining linguistic annotations from novice crowdworkers is far from trivial. A case in point is the annotation of discourse relations, which is a complicated task. Discourse relations (DRs) are logical relations that readers infer between segments of a text, such as *Cause* or *Contrast*. Classifying these relations into categories relies on annotators' interpretation of a text, which makes it a particularly difficult task, as reflected in relatively low inter-annotator agreement (Spooren and Degand, 2010).

An additional complicating factor is the variety of relation types that can be distinguished: The linguistic formalism used in the current study, Penn Discourse Treebank 3.0 (PDTB) (Webber et al., 2019), distinguishes 36 unique types. A thorough understanding of these distinctions requires expert knowledge. DR annotations are therefore traditionally obtained from expert, trained annotators. However, employing such annotators is time-consuming and costly, and so researchers have looked at other ways to obtain reliable annotations, such as crowdsourcing.

Crowdsourced studies typically consist of easier and more intuitive tasks, which can be performed by laymen. Crowdsourcing allows for fast and cost-effective collection of labelled data, but because the tasks need to be intuitive, crowdworkers cannot be asked to code according to a specific linguistic formalism such as PDTB. Further, with such complex tasks, one would need to make sure that the crowdworkers can reach an acceptable level of proficiency.

Annotation tasks are often made suitable for crowdsourcing by considering two dimensions: worker selection and training procedures, and task design. The current study explores this first dimension of obtaining reliable crowd annotations. Annotation tasks for other types of linguistic phenomena have shown that training and selecting crowdworkers are effective strategies to improve the reliability of the data (Roit et al., 2020; Nangia et al., 2021; Parrish et al., 2021). We extend this line of research to discourse relation classification, a task that is known to be difficult. We evaluate the effect of training and selection on agreement using crowdsourced annotation tasks, and compare agreement between workers that did not pass through a selection stage (Study 1), workers that passed through a selection stage and were then trained (Study 2), and workers that were selected but not trained (Study 3).

The second dimension of obtaining reliable crowd annotations for discourse relations – task design – has been addressed in previous, independent efforts. Yung et al. (2019) and Pyatkin et al. (2020) developed different tasks geared to produce the same output: PDTB3-labels for discourse relations. Their methods differ in how they aimed to make the annotation task easier and more intuitive for the crowdworkers. We here collect annotations on various texts using both methods, which allows us to determine whether training effects are generalizable or dependent on the specific crowdsourcing method used. These results will also function as input for the next phase of this project, where we will collect data on a larger scale to be able to compare the effects of the methods on the obtained annotations.

The contributions of this paper are as follows:

- We draw attention to the trade-off in resources and reliability of crowdsourced annotations obtained through a training-and-selection procedure and a selection-only procedure, across two independent crowdsourced discourse annotation methods.
- We show that training crowdworkers improves annotation quality, while at the same time proves to be time- and cost-intensive, and therefore not suitable for certain efforts.
- We propose a selection-only procedure with an implicit learning component, that allows us to obtain annotations of comparable quality to annotations obtained from trained participants, using less resources.

Our results suggest the selection-only approach to be a viable and cost-effective alternative to the training-and-selection approach which dominates NLP nowadays.

2. Related work

2.1. Annotating discourse relations

Expert-level quality for various linguistic annotation tasks can be achieved by aggregating the annotation of as few as four turkers (Snow et al., 2008). However, obtaining high-quality discourse annotations through crowdsourcing appears to be more challenging, due to the complexity of the task and the fact that relations can often convey multiple interpretations. Indeed, Kishimoto et al. (2018) investigated the quality of a crowdsourced discourse relation dataset (Kawahara et al., 2014) and found multiple issues with the crowdsourced annotations, especially for implicit relations. Yung et al. (2019) and Pyatkin et al. (2020), however, both report promising results on their datasets using similar tasks but with different designs.

2.1.1. Discourse Connective method

The two-step discourse connective (DC) method was proposed by Yung et al. (2019). In the first step, participants typed a connective they thought best expressed the relation between two textual arguments. They were also given the option to type *nothing* if they thought no phrase could possibly fit between the segments. In the second step, participants were provided with a new list of connectives that could disambiguate the connective they inserted in the first step. The selection of this list was determined dynamically from their choice in the first step through automatic mapping from our connective bank. When the insertion in the first step did not match any of the entries in the connective bank, or when the participant typed *nothing*, participants were presented with a default list of twelve connectives that can express a variety of relations.

In a series of crowdsourced studies, Yung et al. (2019) showed that the method can be successfully used to reproduce the original PDTB and RST-DT labels for implicit relations, and that the obtained annotations are robust and replicable. Moreover, the method captured the ambiguity of relations by providing a distribution of relation senses, which more accurately represented the range of interpretations that workers inferred.

2.1.2. Question-Answer method

Another proposed method for annotating discourse relations through natural language is by representing them as question and answer (QA) pairs (Pyatkin et al., 2020). Crowdworkers were shown a sentence with eventive words marked in bold, and were then instructed to formulate multiple questions relating such propositions. The questions are formed by choosing a question prefix from a list of 17 prefixes, such as *Despite what* or *What is similar to*, and by then copying parts of the sentence containing one of the propositions, to complete the question body. The answer to a given question is then again made up from parts of the sentence containing the other proposition. Each question prefix represents a soft mapping to a sense in PDTB 3.0 (Webber et al., 2019).

The resulting QADiscourse dataset contains more than 16,600 QA pairs. The relations captured in QADiscourse are intra-sentential and could be both explicit or implicit. The current work on the other hand aims to look at crowdsourcing methods for inter-sentential, implicit discourse relations. For this purpose, we extended the QADiscourse annotation approach to capture relations between two sentences, instead of between two propositions. Additionally, we added more questions prefixes and refined the mapping to more exhaustively map to PDTB. Since the relations hold between two sentences, annotators can now simply mark which of the sentences constitutes the answer and which is part of the question, facilitating the annotation procedure.

2.2. Worker training and selection

Various lines of research have studied how to improve the quality of annotations obtained through crowdsourcing. Promising solutions include controlled annotation protocols, annotation curricula, and worker retention. These will be discussed in turn below.

Controlled crowdsourcing annotation protocols consist of annotator screening and training. Typically, workers first perform a preliminary, crowd-wide round (here referred to as a *recruitment task*). Workers exhibiting acceptable performance (threshold set by the researchers) are then invited to participate in training or qualification rounds, after which the best-performing workers continue to participate in a series of tasks. Such annotation protocols have been successfully applied in other areas in NLP, such as QA-SRL (Roit et al., 2020) and question-answering (Nangia et al., 2021), as well as in the QA method (Pyatkin et al., 2020).

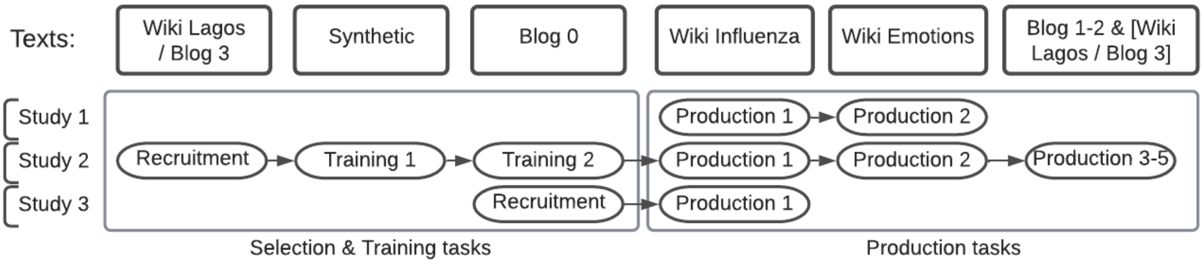

Figure 1: Texts used in Study 1, 2 and 3 and the goal of the task (recruitment, training or production) per study.

The nature and design of the controlled annotation protocols tends to vary. Parrish et al. (2021), for example, use linguists to iteratively propose constraints to workers. Their results show that dynamically adding linguistically motivated constraints results in a more challenging and diverse evaluation dataset. However, a direct communication channel between workers and linguists was not beneficial for data quality. Similarly, for multiple-choice QA, expert feedback and qualifications were shown to be effective, while crowdsourced feedback and asking workers to write justifications were less helpful (Nangia et al., 2021). Rechkemmer and Yin (2020) also studied training effects for crowdsourced tasks, and found that workers who are motivated to learn benefit more from training. Hirth et al. (2013) developed a cost model to identify the main cost factors of different quality checking approaches and how the quality of workers influences the weight of different cost factors. They found that using better and well-trained workers can save cost, even if they are slightly more expensive than other workers.

Annotation curricula have also been used successfully to implicitly train annotators (Lee et al., 2021; Tauchmann et al., 2020). The approach is based on the notion that annotation can be mentally taxing to inexperienced coders: they need to familiarize themselves with the task. To alleviate this, annotation curricula gradually introduce workers to the task by ordering items according to a learning curriculum, with easier examples presented before more difficult ones. This results in a significant reduction of annotation time, without harming annotation quality. Other studies have also found that cognitive factors in the design of the crowdsourcing task, such as visual salience and working memory load, can affect workers' performance (Finnerty et al., 2013; Alagarai Sampath et al., 2014).

Finally, while often not explicitly discussed, many crowdsourced annotation efforts aim to optimize the cost/quality trade-off by investing in worker retainment, to ensure crowdworkers participate throughout the whole data collection cycle. This is often done by promising a bonus for completing certain milestones. Parrish et al. (2021), for example, pay a bonus of 10% of the base pay after completion of 10, 50 and 100 tasks per round and a $20 bonus after the final round. This indicates that worker retention is a prevalent (and often costly) problem. We discuss the problem of crowdworker dropout in this paper and how it affects the viability of training crowdworkers.

Study	N	F	Age range	Avg age	Location
1	20	11	20-46	32	19 UK, 1 CA
2: DC	10	6	21-64	37	8 UK, 2 CA
2: QA	18	14	18-67	31	8 US, 10 UK
3: DC	30	25	21-67	38	28 UK, 1 IE, 1 US
3: QA	30	26	23-70	41	26 UK, 4 CA

Table 1: Descriptive statistics of workers per study; *N*: total number; *F*: number of women.

3. Method

This section presents details relevant to the methodology of all three studies. Study-specific details on methodology, such as the training procedure of Study 2, are elaborated on in the individual Study sections.

Data The items were inter-sentential implicit relations, taken from Wikipedia and the Blog Authorship Corpus (Schler et al., 2006). Figure 1 presents an overview of all texts used throughout the paper. Texts contained on average 20 items.

In order to evaluate the impact of selection and training, we created gold labels for the items. This allows us to compare worker performance across the studies with a set of reference labels. Two expert annotators provided PDTB sense labels for all items.

For all agreement scores reported in the paper, we calculated Cohen's kappa (Cohen, 1960) using the `kappa2` function from the `irr` package in R. In cases where the gold label consists of more than one sense, we calculated agreement in terms of a soft match, whereby any intersection between the labels of two measures is considered agreement. For the gold label annotations, the two annotators showed 83% agreement on a level three distinction; Cohen's κ=.78.[1]

[1] As a general guideline, Spooren and Degand (2010) consider a Cohen's kappa of .7 to signal good agreement for DR annotation. IAA on implicit relations is known to be lower than on explicits (see, e.g., Kishimoto et al. (2018; Hoek et al. (2021)); however, most annotation efforts do not report

Task	DC				QA			
	Group	κ	Agree gold-maj	Agree w/maj	Group	κ	Agree gold-maj	Agree w/maj
Influenza	1	.27	.45	.51	2	.18	.18	.41
Emotions	2	.20	.28	.52	1	.09	.17	.31

Table 2: Agreement statistics for the two texts included in Study 1. κ: Cohen's kappa agreement between the gold label and majority label per item; *Agree gold-maj*: proportion agreement between the gold label and majority label; *Agree w/maj*: proportion agreement of all insertions with majority label.

Implicit relations can often signal more than one relation sense, and so disagreements do not necessarily reflect incorrect labels (Aroyo and Welty, 2013). This is why a third annotator adjudicated the two annotations to create the reference label, possibly consisting of multiple senses. In total, 61% of data received a single reference label, 22% of data was labeled with two senses, 12% with three senses and 5% with four senses.

Participants Participants were recruited via Prolific. Table 1 shows descriptive statistics for the participants that took part in any of the three studies. The participants used in every study are unique (i.e., did not participate in the other studies). In Study 2, for the DC method, 40 participants took part in the recruitment task, and 30 of these were selected to take part in the training. 18 participants completed training, of which 10 took part in all studies. For QA, 48 participants took part in the recruitment task in Study 2. Of these, 18 successfully completed training. These 18 annotators were then invited to participate in production rounds, where each task collected annotations from 10 annotators.

Label extraction For both the DC and QA method, PDTB3 labels were extracted from the crowdsourced annotations; see Yung et al. (2019) and Pyatkin et al. (2020) for additional details.[2]

4. Study 1: No training - baseline

In this study we establish a baseline for annotation agreement by employing 2 groups of crowdworkers. One group annotated text 1 (Influenza) with the DC method and text 2 (Emotions) with the QA method, the other annotated text Emotions with DC and Influenza with QA. The study ended with a survey on task enjoyment. This survey consisted of five questions on a Likert scale of 1 to 5, with 1 meaning strongly disagree and 5 meaning strongly agree. The study lasted approx. 65 minutes and participants were reimbursed £8, following Prolific's recommended rate.

Results Table 2 presents the agreement statistics for the two texts. It shows there is much room for improvement in terms of agreement between the majority and gold labels; the kappa scores are far below the desired threshold of 0.7. There was no pattern distinguishable in participant accuracy between the methods.

The closing survey results indicated that participants thought both tasks were difficult (DC: average of 2.7 on a scale of 1-5, with 5 being "very difficult"; QA: 3.1) and that more instructions were required (DC: 2.6; QA: 2.1). Participants also felt additional training might help them perform better (DC: 2.25, QA: 2.45). Some participants also provided feedback that the level of detail in the instructions was difficult to follow, and that it took time to get used to thinking about language in this manner. This indicates that more training, at a slower pace, might make the task easier for crowd workers. Finally, 3 out of 20 DC participants and 8 out of 20 QA participants indicated that they would not like to take part in further such studies.

Conclusion Study 1 The results indicate that there is much room for improvement. These results show a discrepancy with the original results of both methods, which reported results of higher quality (DC: a precision and recall score of .44; QA: F1=83). It should be noted that the current study made several alterations to the previous procedures. Yung et al. (2019) used selected classes of original PDTB items as data, and participants passed a quality check. Pyatkin et al. (2020)'s data included Wikipedia and WikiNews explicit and implicit intra-sentential relations, and participants received training. These factors can explain the discrepancy, and emphasize the need for further evaluation of the DC and QA methods, as is the goal of the current project.

5. Study 2: Selection-and-training

In Study 2, we evaluate the effect of a selection-and-training procedure on agreement using both the DC and QA method. Participants first completed a recruitment task (either a Wikipedia text or a blog post). Workers who scored over 30% agreement with the gold labels were invited to take part in two training sessions. Both methods used the same items for these training sessions, but had different guidelines to explain the respective tasks.[3] These guidelines were created in conjunction and explained which general types of relations

IAA on implicits separately.

[2] We cover each Level-3 sense in PDTB 3.0, except the belief and speech-act relations; these cannot be distinguished reliably by means of the inserted connective or QA-pair.

[3] DC: https://github.com/merelscholman/DC-annotation/blob/main/DCguidelinesFull.pdf; QA: https://github.com/ValentinaPy/QADiscourse/blob/master/CS/newQAguidelines.pdf

Task	Text	DC			QA		
		κ	Agree gold-maj	Agree w/maj	κ	Agree gold-maj	Agree w/maj
Recruit. a	Wiki: Lagos	.62	.75	.37	.42	.58	.44
Recruit. b	Blog	.55	.6	.51	.54	.63	.58
Training 1	Synthetic	1	1	.76	.84	.85	.54
Training 2	Blog 0	.92	.94	.59	.84	.89	.52
Prod. 1	Wiki: influenza	.61	.73	.58	.47	.64	.48
Prod. 2	Wiki: emotions	.64	.72	.49	.35	.44	.46
Prod. 3	Blog 1	.79	.83	.63	.69	.74	.52
Prod. 4	Blog 2	.72	.73	.49	.52	.59	.47
Prod. 5a	Wiki: Lagos	.56	.67	.77	.51	.67	.65
Prod. 5b	Blog	.58	.67	.76	.67	.75	.71

Table 3: Kappa and accuracy agreement between gold and majority label (out of 10) and average agreement between workers and the majority per task, for annotations obtained in Study 2.

exist and which connectives or QA pairs can be used to express these relations.

The first training session contained 20 synthetic items, constructed by the authors according to a learning curriculum. These items were pretested to ensure that they clearly conveyed the intended relation. The second training contained 20 instances from a blog post. In both training sessions, participants received immediate feedback: If their answer did not match the expected answer, they were shown an explanation of what was expected. We opted for such immediate feedback instead of writing personalized feedback to each worker, as has been done in controlled crowdsourcing (Roit et al., 2020), to make the process more scalable.

Upon completion of training, participants annotated two Wikipedia texts and two blog posts, as well as the Wikipedia or blog post that they had not seen as recruitment task. This allowed us to compare performance on the task by trained versus untrained annotators.

Participants were reimbursed £1.88 per production task, as well as £2.50 for the recruitment task, £5.60 for training 1, £1.88 for training 2, and a bonus of £2 for completing the studies. Workers were originally rewarded with the bonus upon completion of training, but participants trained later in the data collection stage were rewarded after completing all production tasks, due to a high dropout rate after receiving the bonus. In other words, participants received £9.98 and some also £2 bonus before providing any usable output in production tasks.

Results: Agreement per task Table 3 presents the agreement between the gold and majority labels per task. We see that agreement per task varies, but in general, it comes much closer to the desired level of κ=0.7 compared to Study 1 and is therefore of more acceptable quality.

What stands out is that the agreement between the gold and majority labels is high for both methods for training 1, which contained synthetic items. This indicates that workers were able to use the correct connectives or QA pairs to express the intended relations. Hence, the task and methods are feasible. Disagreements in other texts are therefore likely due to the ambiguity of natural language, not due to an inability to use connectives or QA pairs to express relations.

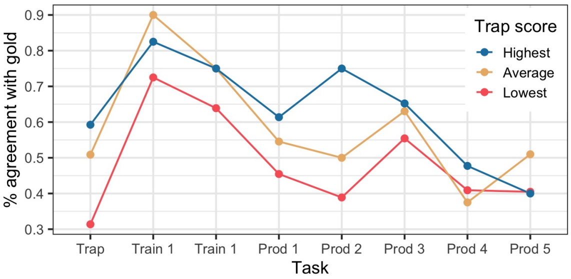

Figure 2: Workers' agreement (%) with the gold per task; workers divided into groups of highest-, average- and lowest-performing workers based on performance in the recruitment task.

Looking at individual performance, we see quite large ranges of agreement with the gold label across the tasks; for some participants, up to 80% of their annotations agree with the gold label, whereas for others, as few as 20% of their annotations agree with the gold label. If this range is consistent within participants (that is, if workers who perform poorly on the recruitment task also consistently show poorer performance on other tasks), this would suggest that more stringent selection criteria could benefit the quality of the obtained annotations. To evaluate whether we can identify consistently high- or low-performing annotators, we refer to Figure 2, which shows the performance of three groups of participants across tasks, based on their performance on the recruitment task. This visualization only contains data from the DC task, because not all workers from the QA stream participated in all tasks. Figure 2 shows that participants who score lowest in the recruitment task, also tend to perform poorer on the other tasks. This speaks to the importance of

Stage	DC			QA		
	κ	Agree gold-maj	Agree w/maj	κ	Agree gold-maj	Agree w/maj
Recruit	.61	.67	.5	.53	.61	.51
Training	.97	.97	.72	.85	.84	.53
Production	.7	.74	.57	.56	.62	.55

Table 4: Agreement statistics per pipeline stage in Study 2.

worker selection. Figure 2 also shows that all three groups show lower agreement with the gold in Production 4 and 5. This could be due to the difficulty of these particular texts, to the effect of training "wearing off", or to an issue with the gold labels.

Results: Effect of training Table 4 presents the agreement on texts before, during and after training (i.e. collapsing the annotations of training and production texts). All agreement metrics are higher after training than before training, which indicates that training is effective. Note also that they are highest during training, which is in part due to the nature of training 1: synthetic relations for which the senses are easier to infer than for relations found in natural language.

These agreement statistics are based on annotations of different texts, from different domains. However, there is overlap between the recruitment tasks and production task 5: half of the workers annotated the Wikipedia text as recruitment and blog text as production task 5, and vice versa. This allows us to compare the accuracy on these texts. Table 3 shows that the agreement between the gold and majority labels indeed improves after training for all texts for the QA task, and for recruitment b (but not a) for the DC task. This further emphasizes the positive effect of training on agreement scores, in particular for the QA method.

Additionally, we can compare the agreement on the overlapping texts between Study 1 and Study 2: Influenza and Emotions. Comparing Table 2 to 3 shows that there is a clear boost in performance between the untrained group in Study 1 and the trained group in Study 2. Note that the trained group is not only trained; they also passed through a pre-selection stage.

Conclusion Study 2 The results indicate that the selection-and-training method leads to increased performance for both methods. The effect of training therefore appears to be generalizable across discourse annotation methods. However, the major drawback of this procedure is that it is not scalable with certain budget constraints. We originally set out to train one set of 10 annotators per method, who were expected to take part in all five production tasks. This was meant to minimize investment in training: given that training is costly, a project's cost efficiency is maximized when as many workers are trained as observations are needed. This proved to be infeasible.

For every task, workers were asked to take part within three days after the invitation. To give workers enough opportunity to return, this deadline was usually extended to one week, after which the next task would start. Even then, a proportion of the (trained) participants would drop out after every task and not return to new tasks. This required us to again return to the beginning of the pipeline and train new participants.

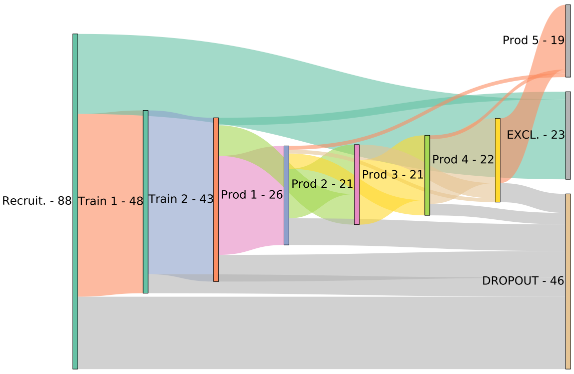

Figure 3: Illustration of Study 2's pipeline for both methods combined: Counts of how many workers dropped out, lost their qualification due to poor performance (excl.), or completed all tasks.

Figure 3 illustrates the recruitment task, to training, to production pipeline, and shows how many participants lost their qualification due to unsatisfactory performance or dropped out at various stages. Especially costly were dropouts during or after training, forcing us to recruit and retrain an additional set of workers. Training investment was therefore misspent on participants that did not provide (enough) useable output on production tasks. Moreover, data collection was slowed, due to the need to recruit and train additional workers. These facets are crucial to the cost and reliability trade-off, as will be discussed in Section 7. For these reasons, the selection-and-training method might not be optimal for certain research efforts, given the available resources.

6. Study 3: Selection-only

In Study 3, we evaluate whether a selection-only procedure is a viable alternative to a selection-and-training procedure, depending on the cost/quality trade-off.

For this approach, we engaged a larger pool of participants with a recruitment task and created a subpool of participants that show potential to be good annotators, based on their performance on the recruit-

				DC				QA			
				n	κ	Agree	Agree	n	κ	Agree	Agree
Cycle	Text	Participant type	Cost			gold-maj	w/maj			gold-maj	w/maj
Study 2	Train 2	Trained part.	10.10	10	.92	.94	.59	18	.84	.85	.54
Study 3	Recruit	All recruit. task	0	30	.84	.89	.52	30	.77	.83	.38
Study 3	Recruit	Self-selection	0	24	.84	.89	.53	22	.58	.67	.38
Study 3	Recruit	Final selection	0	20	.85	.89	.55	9	.7	.61	.45
Study 1	Infl.	Untrained	0	10	.27	.45	.51	10	.18	.18	.41
Study 2	Infl.	Trained	11.98	10	.61	.73	.58	10	.47	.64	.48
Study 3	Infl.	All selected	1.88	19	.41	.68	.48	9	.28	.41	.47
Study 3	Infl.	Decent selected	1.88	16	.58	.77	.54	6	.45	.55	.58

Table 5: Agreement for the blogpost included in Study 2 and 3, and the Influenza text included in all three studies. *cost*: the GBP investment per participant before taking part in this task; *n*: number of participants per iteration.

ment task. The selection-only procedure is more cost-efficient because it doesn't require training investment, and it might therefore be attractive to create a larger pool of qualified workers compared to a smaller pool of qualified and trained workers. This will also be more time-effective, as every task can be completed faster when the annotator pool consists of more workers.

To maximize the value of the recruitment task, we used the second training session of Study 2, in which workers received immediate feedback while they annotated. This means that there is an implicit training component in this selection procedure.

We took two further steps for our annotator selection procedure. First, we only selected people who were studying or had completed a university degree. This was done because the text-analytic component of the task might be more intuitive to people who have studied at a university, where meta-linguistic knowledge is required. Second, we included a self-selection component to be able to only retain motivated annotators: upon completing the task, workers were asked if they would like to participate in more tasks similar to these.

Results: Recruitment task Table 5 displays the results of the recruitment task.[4] The agreement between the gold and majority label is high when including all participants, and when looking at the final selection only. Considering that the workers in the current task have less experience with the methods compared to Study 2's workers, the results show promise.

For both methods, a comparable number of workers indicated that they were not interested in participating in similar tasks, but the two methods show a greater difference when it comes to the gold agreement exclusion criterion: 4 DC participants did not meet this criterion, compared to 13 QA participants. In other words, many workers in the group of QA annotators performed worse than in the group of DC annotators, even though the performance in Study 2 was high for both methods on the same text. This could be due to two factors. First, there could be variability in the quality of crowdworkers that are recruited. When sampling a group of workers, it is natural that the quality might vary between them. A selection procedure is meant to target exactly this factor. A second possibility is that there is variability in the required training of the methods, whereby the DC method is easier for untrained workers than the QA method, but this difference disappears when workers receive QA training. This would explain why the difference in agreement between DC and QA is greater in Study 3 than in Study 2, and would suggest that the effect of training is larger in QA.

Finally, the results indicate that participants improve throughout the task. When looking at the performance on the first half of the text versus the second half, we see an improvement of 7% in the DC method (57% agreement with the gold in the first half to 64% in the second half) and 3% in the QA method (43% to 46%). This indicates that the feedback component is effective.

Results: Production task Of the 20 invited DC participants, 19 took part in the task within 48 hours. For QA, all 9 invited participants took part. Table 5 shows the agreement statistics for this text per study. When looking at results of all selected participants, we see that kappa agreement between the gold and majority labels does drop compared to Study 2. However, the results also show a large range of agreement with the gold of the selected participants: 16%-78%. In particular, 3 DC participants and 3 QA participants showed poor performance (<40%). When removing these from the analysis, performance rises significantly to meet that of the trained participants in Study 2. This indicates that a continuous quality monitoring procedure is of importance to obtain reliable data in a large-scale project. Ideally, gold labels are used as reference to determine worker quality. However, when gold labels are not available, researchers can also take workers' agreement with the majority as a quality measure; these scores tend to follow the same pattern.

Conclusion Study 3 The selection method appears to be an attractive alternative to the selection-and-training method: with more stringent selection criteria and con-

[4] The DC scores in the recruitment task change only slightly because the majority label didn't change when excluding the 6 and 4 workers, respectively, out of 30 workers.

tinuous quality monitoring it is possible to obtain annotations of comparable quality to those from trained participants.

7. Resource Comparison

We conclude that all three attempted worker selection techniques for obtaining discourse relations entail some sort of trade-off between resources spent and quality of the annotations. The quickest and cheapest approach is that of Study 1, where annotators from a large pool can directly work on a given production task. For discourse relation annotations, this generally results in the lowest-quality data comparatively, as seen in Table 5. It took approximately 6 hours to get 10 annotations per instance for the *Influenza* text in Study 1. To obtain 10 annotations per instance for a production task in Study 2, on the other hand, it takes approximately 205 hours.[5] That is the sum of the average time it took us to get a batch of annotators to complete the recruitment task, the two training tasks and finally one production task. This is not taking into account the workers who drop out during the process.

Figure 3 shows that for each task in the pipeline a fraction of annotators drop out: After the recruitment task, 67 annotators were invited to participate in training, of which 28% dropped out. During training or after completing training, an additional 13% dropped out. This resulted in £106 spent on workers who did not complete a single production task. We also experienced difficulties with worker retention during the production tasks, which required us to continuously re-train new workers, halting production collection timelines.

Study 3 evaluated a third method, which in terms of resources spent lies somewhere between Study 1 and 2. The three steps before production in Study 2 were reduced to a single step, with stronger qualification criteria. This resulted in approx. 2% of the original (Study 2) time cost. For a pool of 10 trained vs. 10 selected workers, £160.48 versus £36 have to be spent, respectively; a 77% decrease. This is taking into account dropout: According to results in Study 2, one would need to train approx. 18 good workers to end up with 10 trained workers that provide output (with the risk that they might drop out at later stages in production). For Study 3, assuming that approx. 53% of workers pass the selection (for DC it is 75% and for QA it is 30%), 19 workers need to pass through recruitment (à £1.88 each) in order to select 10 well-performing workers. In terms of performance, the untrained, selected workers show comparable performance to the trained workers.

We suggest the following considerations in choosing a crowdsourcing protocol for the annotation of discourse relations. Given the time, cost and performance trade-offs, selection-only could function as a viable alternative to selection-and-training, where one of the biggest drawbacks is the dropout rate. Another point of consideration with respect to that is the crowdsourcing platform used. Prolific is typically meant for smaller experimental studies and so participants are relatively flexible in when and whether they participate. Other platforms, such as Appen or Upwork, boast a pool of workers who can be hired to work on longer-term projects, and are therefore less likely to drop out during the pipeline. However, this comes with the trade-off of a higher cost for data collection.

Lastly, the results from the selection-only influenza task indicate that participants who passed an initial selection procedure can show poor performance when participating in more tasks. Such a procedure might therefore work best when continuing to monitor quality. We recommend to incentivize participants throughout the data collection cycle. This can be done by offering bonuses, reminding them that accuracy will be checked, and removing participants from the subpool after they fail certain intermediate quality checks. A portion of the budget would therefore need to be reserved for quality management.

8. Conclusion

Discourse relation annotation is a complex task that requires (meta-)linguistic awareness and knowledge from annotators. We here evaluated the effect of worker training and selection on the reliability of discourse relation annotations obtained using two separate crowdsourcing methods. The results show that training does lead to more reliable annotated data, but this comes at a high cost, both in terms of resources and time needed to train workers and collect the data. A selection-only approach might be more viable for certain projects in terms of resources. Individual projects will need to weigh all relevant factors – time, resources and reliability – in order to create the optimal design for their crowdsourced annotation task.

Crucially, we note that the current study provides further support to previous work showing that obtaining reliable annotations of discourse relations using crowdworkers is possible. Agreement on a set of synthetic items in Study 2 showed that workers are able to express the intended relation using discourse connective or QA pairs. This speaks to the feasibility of both methods. Any disagreement between the majority and gold labels in other texts can therefore be attributed to the difficulty of inferring discourse relations, which tends to be the case especially for implicit relations.

This work was the first step in a larger project, in which we investigate how design choices for discourse annotation tasks shape research results. In future work, we will collect and release more annotations for intersentential discourse relations from multiple genres using both methods. This will facilitate a more detailed comparison between the obtained annotations.

Acknowledgements

This research was funded in part by the German Research Foundation (DFG) as part of SFB 1102 “Information Density and Linguistic Encoding”, by grants

[5] These are estimates averaged between DC and QA.

from Intel Labs, Facebook, the Israel Science Foundation grant 1951/17 and by the European Research Council (ERC) under the Horizon 2020 research and innovation program, grant agreement No. 677352 (NLPRO), for which we are grateful.

9. Bibliographical References

Alagarai Sampath, H., Rajeshuni, R., and Indurkhya, B. (2014). Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3665–3674.

Aroyo, L. and Welty, C. (2013). Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013 ACM*, 2013(2013).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Finnerty, A., Kucherbaev, P., Tranquillini, S., and Convertino, G. (2013). Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, pages 1–4.

Hirth, M., Hoßfeld, T., and Tran-Gia, P. (2013). Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. *Mathematical and Computer Modelling*, 57(11-12):2918–2932.

Hoek, J., Scholman, M. C., and Sanders, T. J. (2021). Is there less annotator agreement when the discourse relation is underspecified? In *Integrating Perspectives on Discourse Annotation (DiscAnn)*, pages 1–6.

Kawahara, D., Machida, Y., Shibata, T., Kurohashi, S., Kobayashi, H., and Sassano, M. (2014). Rapid development of a corpus with discourse annotations using two-stage crowdsourcing. In *COLING*.

Kishimoto, Y., Sawada, S., Murawaki, Y., Kawahara, D., and Kurohashi, S. (2018). Improving crowdsourcing-based annotation of japanese discourse relations. In *LREC*.

Lee, J.-U., Klie, J.-C., and Gurevych, I. (2021). Annotation curricula to implicitly train non-expert annotators. *ArXiv*, abs/2106.02382.

Nangia, N., Sugawara, S., Trivedi, H., Warstadt, A., Vania, C., and Bowman, S. (2021). What ingredients make for an effective crowdsourcing protocol for difficult nlu data collection tasks? In *ACL/IJCNLP*.

Parrish, A., Huang, W., Agha, O., hwan Lee, S., Nangia, N., Warstadt, A., Aggarwal, K., Allaway, E., Linzen, T., and Bowman, S. R. (2021). Does putting a linguist in the loop improve nlu data collection? In *EMNLP*.

Pyatkin, V., Klein, A., Tsarfaty, R., and Dagan, I. (2020). QADiscourse-Discourse Relations as QA Pairs: Representation, crowdsourcing and baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819.

Rechkemmer, A. and Yin, M. (2020). Motivating novice crowd workers through goal setting: An investigation into the effects on complex crowdsourcing task training. In *AAAI 2020*.

Roit, P., Klein, A., Stepanov, D., Mamou, J., Michael, J., Stanovsky, G., Zettlemoyer, L., and Dagan, I. (2020). Controlled crowdsourcing for high-quality qa-srl annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013.

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

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)*, pages 254–263, Waikiki, HI.

Spooren, W. P. M. S. and Degand, L. (2010). Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, 6(2):241–266.

Tauchmann, C., Daxenberger, J., and Mieskes, M. (2020). The influence of input data complexity on crowdsourcing quality. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, pages 71–72.

Webber, B., Prasad, R., Lee, A., and Joshi, A. (2019). The Penn Discourse Treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.

Yung, F., Demberg, V., and Scholman, M. (2019). Crowdsourcing discourse relation annotations by a two-step connective insertion task. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 16–25.