Crowdsourcing Discourse Relation Annotations by a Two-Step Connective Insertion Task

Frances Yung[†], **Merel C.J. Scholman**[†] and **Vera Demberg**[†,‡]
[†]Dept. of Language Science and Technology
[‡]Dept. of Mathematics and Computer Science, Saarland University
Saarland Informatic Campus, 66123 Saarbrücken, Germany
[frances|m.c.j.scholman|vera]@coli.uni-saarland.de

Abstract

The perspective of being able to crowd-source coherence relations bears the promise of acquiring annotations for new texts quickly, which could then increase the size and variety of discourse-annotated corpora. It would also open the avenue to answering new research questions: Collecting annotations from a larger number of individuals per instance would allow to investigate the distribution of inferred relations, and to study individual differences in coherence relation interpretation.

However, annotating coherence relations with untrained workers is not trivial. We here propose a novel two-step annotation procedure, which extends an earlier method by Scholman and Demberg (2017a). In our approach, coherence relation labels are inferred from connectives that workers insert into the text.

We show that the proposed method leads to replicable coherence annotations, and analyse the agreement between the obtained relation labels and annotations from PDTB and RST-DT on the same texts.

1 Introduction

Implicit coherence relations are connections between text segments that are not overtly marked. Annotating implicit coherence relations using crowd-sourcing is methodologically challenging, because assigning coherence relation labels as used in popular discourse frameworks like the Penn Discourse Treebank style (PDTB, Prasad et al., 2008, 2018) or the Rhetorical Structure Theory (RST, Mann and Thompson, 1988; Carlson et al., 2003) requires linguistic knowledge and substantial training. It is thus not possible to obtain high quality annotations of coherence relation labels from untrained crowd workers (Kawahara et al., 2014; Kishimoto et al., 2018).

A more promising method for obtaining discourse annotations through crowd-sourcing is to ask workers to insert discourse connectives (Rohde et al., 2016; Scholman and Demberg, 2017a). However, this method so far has only been used in settings where it was sufficient to give workers a small set of connectives to choose from, and not in broad-coverage coherence relation annotation. For example, Rohde et al. (2016) focused on identifying cases where several coherence relations may hold between two segments. They provided participants with relations that were already marked with a discourse adverbial, and asked them to additionally insert a conjunction out of a list of six highly frequent connectives (*and, but, so, because, before, or*).

Highly frequent connectives are often ambiguous, for instance, the insertion of *but* does not allow us to infer whether the relation is a contrast or a concession relation. When we want to do fine-grained relation annotation, providing only general connectives is thus not sufficient. Scholman and Demberg (2017a) addressed this problem by restricting the types of relations that could occur in their experiment. They selected six types of coherence relations from the overlapping part of the PDTB2.0 and RST-DT corpora, and re-annotated them using crowd-sourced annotators. Workers in this study could choose from a list of connectives which distinguish unambiguously between the six relation types of interest. For example, instead of the connective *but*, they provided a choice between *nevertheless* and *by contrast*.

However, for annotating text more generally, we need to provide connectives that can capture all types of relations, and on top of that make sure that the insertions can help us to disambiguate between coherence relations. This poses the problem that the list of connectives that participants should choose from would be come unwieldily large – it's unlikely that participants would be very capable of choosing one connective to insert from a list of 50

Proceedings of the 13th Linguistic Annotation Workshop, pages 16–25
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

connectives.

In this work, we therefore propose a new annotation procedure which builds on the method of Scholman and Demberg (2017a). Our contributions in this paper consist of:

- a novel two-step procedure for eliciting discourse connective insertions from naïve workers;

- a demonstration that the generalized method is comparable in reliability of annotations to the original more restricted crowd-sourcing method proposed by (Scholman and Demberg, 2017a);

- a "connective bank" consisting of 800 entries including traditional connectives as well as variations of connectives and alternative lexicalizations;

- an analysis comparing the obtained coherence labels to labels from professionally annotated discourse treebanks. Our analysis shows that crowd-sourcing captures a mixture of characteristics from PDTB 3.0 and RST-DT annotations.

The data collected in this study, including the crowdsourced annotations of 447 implicit discourse relations and a *connective bank* of 800 connective phrases, is freely available for the community.[1]

2 Background

Crowd-sourcing is an increasing popular alternative to professional annotation of linguistic materials because of time efficiency. However, classification of discourse relations is not a trivial task. This is especially true for implicit relations, where explicit connectives are missing. Detailed guidelines and extensive training are used in traditional annotation by experts.

Kawahara et al. (2014) presented a first attempt to crowd-source discourse relation annotation. The workers first decided whether text spans were connected by a relation, and then assigned one out of seven sense labels in case a relation was identified. The proposal is appealing in terms of time efficiency, but the quality is questionable because evaluation was not carried out. Kishimoto et al. (2018) later re-annotated a portion of the relations by trained annotators, and found that the quality of the annotation from crowd-sourcing was not satisfactory. They argued that the naïve workers did not completely understand the definition of relation senses and the task was too demanding.

Following the success of analyzing multiple coherence interpretation based on connective insertions by crowd workers (Rohde et al., 2016), Scholman and Demberg (2017a) proposed to use a connective insertion task as a more intuitive alternative to the annotation of coherence relation labels, when working with untrained annotators.

In their experiment, workers are asked to "drag-and-drop" one out of eight unambiguous connectives into the blank between two text spans to express the discourse relation holding between them.[2]

Scholman and Demberg (2017a) evaluated the annotation method by re-annotating a portion of the WSJ text for which professional coherence relation annotations (PDTB, RST-DT) are also available. The majority of the crowd-sourced labels converged with the label of PDTB, showing that the method is reliable, at least in this simplified setting where the set of possible discourse relations is limited and given.

Furthermore, replicability and robustness of the crowd-sourced annotation was demonstrated by replicating the crowd-sourced annotation on the same coherence relations without providing the participants with extra contexts. The resulting connective distributions of the two experiments closely agreed with each other, showing that the annotation is replicable even when contexts are absent.

However, the method used by Scholman and Demberg (2017a) also presents some shortcomings: firstly, it doesn't easily scale up to distinguish between the full set of coherence relations that can occur in a text, and secondly, prompting workers to choose among a set of given connectives might affect their interpretation of the coherence relation[3]. For example, workers might have refrained from inserting an unambiguous but rather heavy-handed connective like "as an illustration" if the text doesn't sound "natural" after

[1] https://git.sfb1102.uni-saarland.de/francesyung/2-step-crowdsourced-discourse-annotation

[2] The connectives are *because, as a result, in addition, even though, nevertheless, by contrast, as an illustration* and *more specifically*.

[3] Although workers were also allowed to type other phrases, such manual inputs were rare.

inserting the connective.

We here propose a two-step design which allows the workers to mark each relation by a free insertion step followed by a customized disambiguation step.

3 Method

3.1 Annotation task design

In the first step, workers are shown a short text passage containing a blank between two text segments. They are asked to type in a connective that they think best expresses the relation between the textual arguments. They are also given the option to type *nothing* if they think no phrase possibly fits between the segments.

We expected that freely inserted connectives chosen by workers might often be ambiguous, such that we would not be able to infer a specific coherence relation label from these free insertions. We therefore include a second step, where participants are presented with a list of at most 10 connectives that disambiguate the connective phrase they chose to insert in the first step. The selection of the connectives is determined dynamically from their choice in the first step. They are then asked to drag and drop the phrase that best expresses the relation holding between the text segments. They can choose the *none of these* option if they think none of the given options fit.

For example, the worker had typed *however* in the first step, and this connective can mark ARG1-AS-DENIER, ARG2-AS-DENIER, and CONTRAST, the connectives *even though*, *despite this* and *on the contrary* will be given as a choice to the worker in the second step. If the first free insertion is already an unambiguous connective, the second step is skipped, and the worker proceeds to the next task.

In order to allow us to determine what connectives should be shown in the second step, we constructed a connective bank containing a collection of connective phrases and their (multiple) senses. We also created a list of unambiguous connective phrases for each of the coherence relations that we distinguish.

In some cases, the insertion in the first step did not match any of the entries in our connective bank (see Section 3.2). This might happen because of typos, insertions that are not actually connective phrases, or which are new connective phrases that are not yet contained in our connective bank. We observed during the development of our method that this happens particularly frequently in cases where none of the frequent connectives seem to fit the text well. We therefore created a list of ten connectives that typically fit such cases well. This default list is presented to workers when we do not recognize their insertion from the first step, or if they typed *nothing*. This list of default connectives includes *accordingly, actually, as you can see, essentially, evidently, in other words, in summary, on top of that, specifically,* and *to provide some background information*.

3.2 Connective bank

Based on existing discourse resources, we constructed a bank of discourse connecting phrases and manually annotated the possible senses of each phrase. The set of labels is adapted from the sense hierarchy of PDTB3; it is shown in Table 2[4].

We tested the coverage of the connective bank in a number of pretests with a separate group of crowd workers, using materials from PDTB, as well as transcripts of TED talks, in order to capture the possible connectives used by the naïve workers. The free insertions collected from the pretests were manually classified as to whether they are connective phrases. The identified connectives are furthermore labelled with discourse senses and added to the connective bank.

The final version of our connective bank contains 800 entries, which include typical discourse connectives (e.g. *because*), variation of connectives (e.g. *largely because*), combination of connectives (e.g. *and because*) and ”alternative lexicalization” (e.g. *the reason is that*).[5] The bank can be expanded with the new free insertions collected after each round of annotation.

The list given in Step 2 contains connectives that mark the relation senses that we want to distinguish as unambiguously as possible. We determined these connectives with the help of Knott (1996)'s connective hierarchy. The complete list is shown in Table 2.

[4] We cover each Level-3 sense in PDTB 3.0, except the 4 speech-act relations, because the speech-act relations are rare and cannot be distinguished with their non-speech-act versions by means of the inserted connective. In addition, we included two extra relations: PRESENTATIONAL and BACKGROUND

[5] We also find a lot of frequent typos among the insertions in the first step, such as “becuase”. These typos are also stored as variants in the connective bank, but are not counted towards the 800 entries.

3.3 Aggregation of annotation

From each worker, we thus typically collect one freely inserted label and one forced choice label. In order to determine the coherence relation label, we retrieve the potential relation senses of both the freely inserted and the forced choice connectives from the connective bank, and calculate the intersection of the relation senses they can mark. The exact algorithm is shown in the Appendix.

Each worker assigns either a single or multiple senses to a relation. If the intersection set contains one sense, the relation is resolved to a single unambiguous sense. If the worker chooses an ambiguous phrase in the first step and "none of these" in the second step, then the relation is annotated with the multiple senses of the ambiguous phrase.[6]

It can however happen that participants type a phrase we do not know (and cannot interpret, e.g. because it is not a connective), or choose to insert *nothing* in the first step, and then choose *none of these* in the second step. In these cases, which are rare (3% of the annotation), we remove the data from further analysis.

The multiple annotations collected from multiple workers for each item are aggregated to a sense distribution per item. If a worker assigned more than one sense to the item, the count is equally split among the multiple senses.

We conducted two annotation experiments to evaluate the methodology and reliability of the proposed method.

4 Experiment 1

The objective of this experiment is to confirm the proposed task design and compare it with the forced-choice design proposed by Scholman and Demberg (2017a).

4.1 Materials

Experiment 1 used the same set of items as in Scholman and Demberg (2017a). These are 234 items of six types of explicit and implicit relations chosen from the PDTB[7], which are also annotated in RST-DT.

In the PDTB, each of these items consists of two consecutive text segments connected by a discourse connective, which is either present in the original text (explicit relation) or inserted by the PDTB annotators (implicit relation).
An example of each is shown below.

1. *Some automotive programs have been delayed,* **while** they haven't been canceled. [wsj_0628: explicit relation= ARG1-AS-DENIER]

2. *The explosions began when a seal blew out.* **As a result,** dozens of workers were injured. [wsj_1320: implicit relation= RESULT]

In the experiment, workers see the text segments and are asked to insert a connective phrase.

For the CAUSE, CONJUNCTION, CONCESSION and CONTRAST relations that are contained in this experiment, both PDTB and RST-DT annotations agreed with one another. The INSTANTIATION and LEVEL-OF-DETAIL items were however selected such that RST-DT annotations do not always agree with PDTB annotations (see Scholman and Demberg (2017a) for more details). Therefore, these two types of relations are expected to be more ambiguous. The number of instances per relation is given in the subgraph titles in Figure 2. The items are divided into 12 sense-balanced batches.

Following the experimental design in (Scholman and Demberg, 2017a), we conducted two versions of this experiment – one with context and the other without, where context is defined by the window of two sentences before and one sentence after the text spans linked by the coherence relation.

4.2 Procedure

Each set of items was divided into 12 batches, and each batch of 17-20 questions was annotated by 12 workers.

In total, 380 workers were recruited and each of them completed one or more batches, but never the same batch in two conditions. Workers who inserted less than three different phrases in step one, or selected "none of these" in step two in more than 60% of their responses were screened and their annotations were examined and, if necessary, replaced by annotations of newly recruited workers.

The task was implemented by LingoTurk (Pusse et al., 2016) and the workers were recruited through Prolific.[8] They were awarded with 2.2

[6] Scholman and Demberg (2017a) allowed insertion of multiple connectives, but they found that workers seldom do so, possibly due to increased workload.

[7] We used the same items but the updated sense labels from PDTB3.

[8] https://prolific.ac

British pounds on average for each batch of annotation.

4.3 Results

We first analyzed the free and forced insertions collected in each step of the two-step approach, and then compared the annotations with those of Scholman and Demberg (2017a).

The results showed that the proposed two-step free-choice annotation method successfully scaled the connective insertion task to a procedure for crowd-sourcing discourse annotation.

4.3.1 Connective insertion in Steps 1 and 2

First we tested whether the proposed method worked as it was intended. On one hand, if workers mostly inserted an unambiguous connective in the first step, the second step would not be necessary. On the other hand, if the workers often inserted ambiguous connectives in the first step but failed to choose any connectives in the second step, the 2-step operation failed in labeling the relation with a precise sense.

The experiment results demonstrated that the proposed method is flexible and useful. Table 1 shows the proportion of connectives inserted by the workers in each step of the experiment.

<table>
<tr><td rowspan="3">Step 1</td><td colspan="6">free insertion</td></tr>
<tr><td>unamb.</td><td colspan="2">ambiguous</td><td>unknown</td><td colspan="2">nothing</td></tr>
<tr><td>23%</td><td colspan="2">64%</td><td>9%</td><td colspan="2">4%</td></tr>
<tr><td rowspan="3">Step 2</td><td>skip</td><td colspan="2">customized</td><td colspan="3">default</td></tr>
<tr><td>unamb.</td><td>unamb.</td><td>amb.</td><td>unamb.</td><td>amb.</td><td>none</td></tr>
<tr><td>23%</td><td>58%</td><td>6%</td><td>6%</td><td>4%</td><td>3%</td></tr>
</table>

Table 1: Proportions of insertion normalized per step. The proportion of the unambiguous connective in Step 1 is carried over to Step 2.

In the first step, workers freely typed a connectives between the two text segments. Most (87%) of the connectives were identified in our connective bank, and the majority (64%) of them were ambiguous.

Table 2 lists the most common connective phrases the workers typed in Step 1. Naïve workers tended to insert common connectives that are usually ambiguous, such as *and, as* and *but*. The unambiguous connecting phrases, such as *simultaneously*, are uncommon expressions that people do not intuitively produce.

relation sense to be labelled	most common free insertion in Step 1	connective for disambiguation in Step 2
CAUSE		
reason	because	for the reason that
result	and	as a result
negative result*	-	that's why it is impossible that
reason-belief	because	considering that
result-belief	so	so I think
CONCESSION		
arg1-as-denier	but	even though
arg2-as-denier	however	despite this,
CONTRAST		
contrast	however	on the contrary
CONJUNCTION		
conjunction	and	in addition in conjunction with this
INSTANTIATION		
arg1-as-instance*	-	this example illustrates that
arg2-as-instan.	for example	as an example
LEVEL-OF-DETAIL		
arg1-as-detail	actually	in general
arg2-as-detail	specifically	in more detail, specifically
OTHERS		
synchronous	as	simultaneously
precedence	and	afterwards
succession	previously	previously
arg1-as-cond.	in this case	in this case
arg2-as-cond.	where	if
arg1-as-neg.cond.*	-	if not
arg2-as-neg.cond.*	-	unless
arg1-as-goal	through	for that purpose
arg2-as-goal	in order to	in order to
arg1-as-manner	by doing so	by doing so
arg2-as-manner	by	by means of
arg1-as-subst	-	rather than, instead of
arg2-as-subst	but	instead
disjunction*	-	and/or
equivalence	*nothing*	in other words, that is to say
arg1-as-except.*	-	other than that
arg2-as-except.	but	except
similarity	as	in a similar manner
background	*nothing*	to provide some background information
presentational	*nothing*	as you can see

Table 2: The list of 33 discourse relations to be annotated by the two-step connective insertion task and the most common phrase workers typed in Step 1 alongside the unambiguous connective defined in the connective bank for the identification of relation in Step 2. Relations marked by * (6 in total) are defined but never annotated by the workers. BACKGROUND and PRESENTATIONAL are two additional senses that are not from the PDTB3 taxonomy.

However, people were still able to use these uncommon expressions when they were prompted to do so in the second step. The majority of the ambiguous connectives in the first step were disambiguated to a single sense in the second step. For example, *however* was readily distinguished between the ARG2-AS-DENIER and CONTRAST senses; and *and* was disambiguated between PRECEDENCE, RESULT and CONJUNCTION.

A manual check of the responses inserted as free text revealed that 9% of the insertions in this first step were not actually connectives. This is not surprising, given that untrained workers may not know the concept of discourse connectives and could insert non-connective phrases depending on context, such as *unfortunately*, or *they think*. Also, workers preferred not to insert any phrases in 4% of the instances. This is also expected because some discourse relations, e.g. CONJUNCTIONS, are often implicit.

Nonetheless, workers were able to choose a connective from the default options suggested to them for most of the unknown/nothing cases. This shows that our default list of connectives successfully helped the untrained workers to express discourse relations that were not obvious to them.

Overall, the two-step approach resolved the workers' insertions to a single label in 87% of the cases and 27 types of sense labels were collected (See Table 2). This is encouraging because untrained workers would not have been able to carry out such fine-grained classification in one step.

4.3.2 Comparison between forced and free insertions

Next, we compared the methodology of the proposed two-step free-choice task with the one-step forced-choice task of Scholman and Demberg (2017a). We wanted to see if workers' identification of the discourse relation was biased to the set of options available to them and whether contexts were necessary for workers to infer the relations.

The overall distributions of the annotated senses under different annotation conditions are shown in Figure 1.

It can be seen that the relative distribution of the senses was maintained across different approaches, suggesting that the 2-step setup successfully replicates the results obtained from the force-choice method. However, the distributions were statistically different across the two methods because 12% of the annotated sense did not belong to the 6 original classes of relations. This is expected because the workers were free to assign any relations instead of from a predefined list.

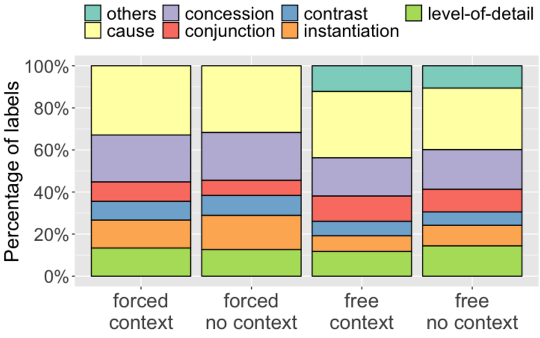

Figure 1: Label distribution per annotation condition of the S&D set

Another finding was that the distributions between the *no context* and *context* conditions were similar. Pearson's χ^2 tests showed a significant difference in the distribution of senses between the two conditions for the original CAUSE ($p = .0478$) and LEVEL-OF-DETAIL ($p = .0159$) items but no significant difference for the other items (CONCESSION: $p = .991$, CONJUNCTION: $p = .258$, CONTRAST: $p = .975$, INSTANTIATION: $p = .232$).

This result partially replicates the finding in Scholman and Demberg (2017a) that contexts offer limited help in this set of items.

4.3.3 Comparison with reference annotation

To assess the quality of the annotations collected by the proposed method, we compared the collected labels with the original expert label per item.

We selected the majority label of each item based on the aggregated distribution for comparison. If an item had more than one majority label, one of them was selected randomly.[9]

Figure 2 shows the distribution of the crowdsourced labels, grouped by their original PDTB label. Only the results under the *context* conditions are shown because the results under the *without context* condition are similar. It can be seen that the distribution mostly replicated the distribution obtained in Scholman and Demberg (2017a),

[9] We also tried aggregation by an annotation model (Dawid and Skene, 1979; Passonneau and Carpenter, 2014), but the predicted labels were mostly the same as the majority label.

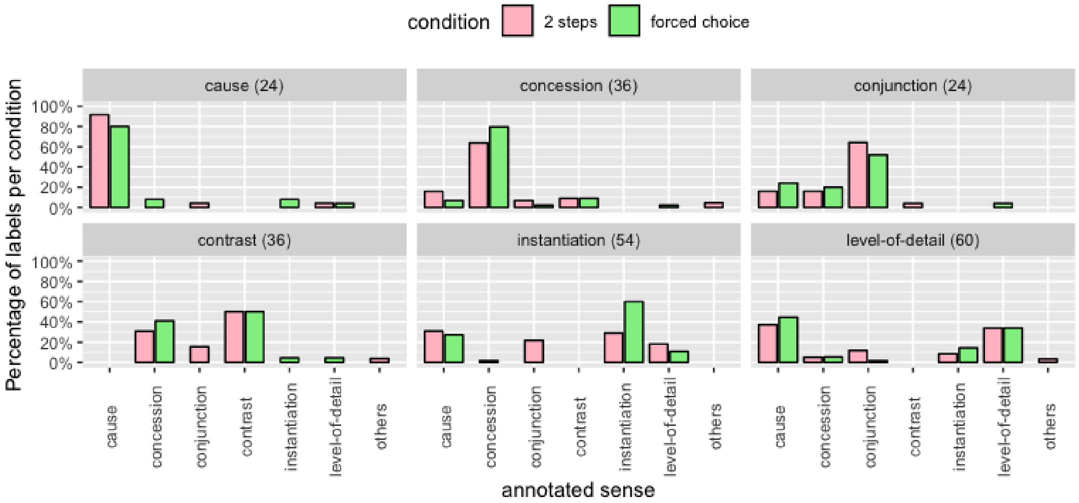

Figure 2: (Experiment 1 results) Distribution of majority sense of the items annotated by the **2 steps** approach in comparison with the **forced choice** approach under the *context* condition. Results are grouped by the original PDTB relation (titles of subgraphs). The item count of each group of relations are bracketed.

except for the INSTANTIATION items. For these items, workers tended to choose CONJUNCTION rather than INSTANTIATION in the two-step task comparing to the forced choice task.

It is known that INSTANTIATION relations have an additive function and thus often coexist with CONJUNCTIONS (Scholman and Demberg, 2017b). However, the labelling of CONJUNCTION could have been suppressed in the forced choice setting, because the single connective that was provided for CONJUNCTIONS was *in addition*, and this phrase may not fit in certain contexts.

Comparing with PDTB annotation, it can be observed that the distributions converged and diverged following the manipulation on the agreement between PDTB and RSTDT.

For example, the crowd-sourced labels converged on the CAUSE sense for the CAUSE items, which were selected if they had high cross-framework agreement. On the other hand, the crowd-sourced labels diverged to a number of senses for the LEVEL-OF-DETAIL items, which were selected if they had low cross-framework agreement.

In addition, CONTRAST items were often annotated as CONCESSION, which is not surprising because the two types of relations are easily confused even for expert annotators. In fact, the overall sense distribution of CONTRAST and CONCESSION reversed when the sense labels were updated from PDTB2 to PDTB3.

In sum, the results of Experiment 1 validated the flexibility and potential of the two-step design and showed that it can be used to obtain similarly reliable annotation as in the oracle forced-choice setting. We conducted another experiment to evaluate the performance of the approach in practical annotation.

5 Experiment 2

The items used in Experiment 1 were chosen such that RST-DT annotations for the same text spans were comparable to the PDTB annotations (for CONTRAST, CONCESSION, CAUSE AND CONJUNCTION). This means that the items were not entirely representative of a real-life annotation setting (i.e., the relations might have been easier to annotate). We therefore conducted another experiment using items that were selected without this constraint.

5.1 Materials

We selected a set of 215 items from the overlapping section of PDTB and RST-DT. We only chose relations where the argument spans were the same in PDTB and RST-DT and the second argument immediately follows the first argument. For comparability to the previous experiment, we restricted the selection to the same six sense classes. Items already tested in Experiment 1 were excluded. The distribution of relation labels in this new set provides a reference of the natural distribution of these six types of coherence relations. The items were randomly divided to 12 batches

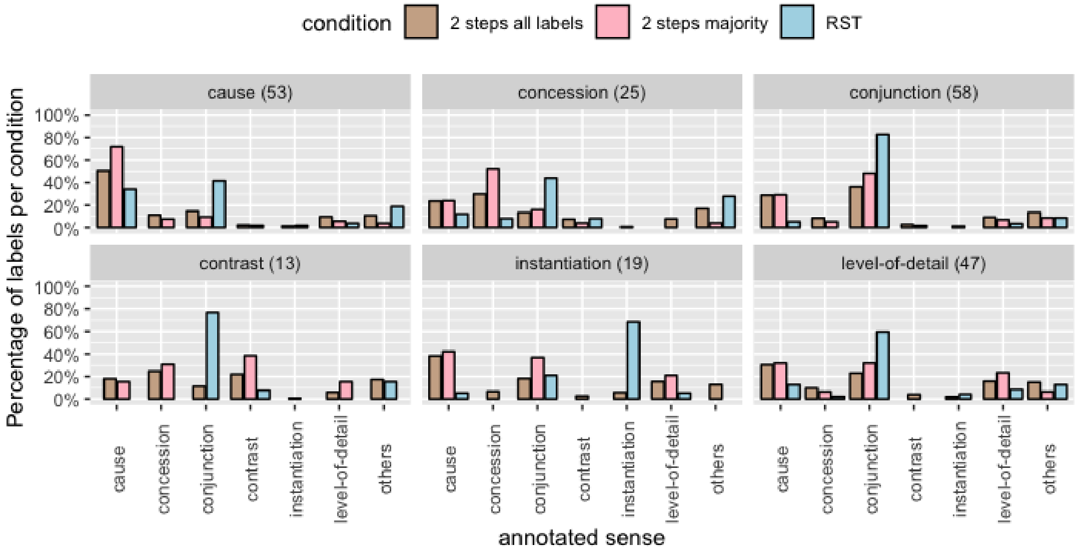

Figure 3: (Experiment 2 results) Distribution of all sense labels (**2 step all labels**) and the majority sense (**2 step majority**) of the items annotated by the *2 steps* approach under the *context* condition in comparison with annotation of RSTDT (**RST**). Results are grouped by the original PDTB relation (titles of subgraphs). The item count of each group of relations are bracketed.

(instead of being sense-balanced). This resembles a situation in which the proposed method is applied to annotate new items. The rest of the experimental set up was the same as in Experiment 1.

5.2 Results

Figure 3 shows the distribution of all the crowd-sourced labels as well as the majority labels collected for each group of relations as annotated in PDTB. Distribution of the RST-DT labels are also shown for comparison. The relation definitions of PDTB and RST-DT do not directly map with each other. In order to compare the annotations of both resources with the crowdsourced labels, we converted the RST labels to PDTB labels according to the Unifying Dimensions interlingua (Demberg et al., 2017).

The results showed that the distributions of the crowd-sourced labels overlapped with both PDTB and RST-DT annotations, except for INSTANTIATIONS (see discussion). The annotations of PDTB and RST-DT largely differ for this more representative selection.

Table 3 shows the agreement of the crowd-sourced labels with PDTB, compared with the agreement between the PDTB and RST-DT labels. It can be seen that the labels crowdsourced by the proposed method had higher overall agreement with PDTB comparing with RST-DT labels.

This experiment showed that expert annotation of discourse relations cannot be represented by a single label and the annotation crowdsourced by the two-step method captured the characteristics of both resources.

	2 steps		RST-DT	
PDTB3	Prec.	Recall	Prec.	Recall
cause	.44	.71	.58	.34
concession	.48	.52	.67	.06
conjunction	.47	.47	.39	.83
contrast	.63	.38	.33	.08
instantiation	.0	.0	.56	.47
level-of-detail	.46	.23	.44	.09
overall	.44	.44	.40	.40

Table 3: Agreement of the majority crowd-sourced and RST-DT labels with the PDTB3 labels and the label distribution of the random set.

6 Discussion

The results demonstrated that the multiple readings of discourse relations were reproduced across the two annotation conditions, even though there was not always agreement with professional annotations. While Scholman and Demberg (2017a) had already reported the reproduction of label distributions under the with and without context conditions, we found that the distributions are also reproduced when free insertion of connectives is allowed. This is stronger evidence that the limited labels collected by traditional annotations might not be sufficient to reflect the multiple reading of discourse relations, while a distribution of labels collected by multiple annotation is more informative.

However, we also identified potential problems: our naïve workers seem to have under-labelled INSTANTIATION relations, especially in Experiment 2. On top of the fact that INSTANTIATIONS are difficult in general, a closer look shows that these items mostly contain quotations, and it is difficult to distinguish whether the relation is between the previous argument and the content of the quote, or the fact that someone said something. This could be the source of confusion for the crowd workers, which deserves to be addressed more specifically in future research.

Another challenge is the causal preference bias (Sanders, 2005). Although we expected that over-interpretation would be reduced in the *free* insertion approach compared with *forced* selection from an available list, we observed an over-interpretation of CAUSE relations. CAUSE relations may be over-labelled because readers readily infer causality during text processing: Scholman (2019) shows that readers infer causal relations readily when not processing the text very deeply. Since the materials we used came from outdated news journal texts from the US, they were likely to be hard to understand for the workers who mostly come from the UK, and the causality bias could hence be particularly prominent in our study. A future study on a different text type would be informative in this respect.

In terms of methodology, we also plan to extend the method to make better use of the connectives provided during the free insertion step. For example, if a worker types *and* in the first step and chooses *so* in the second step, the current algorithm would simply combine the two insertions to a CAUSE relation by taking the intersection of senses. However, there is a chance that the forced choice was prompted by the given options, and that the inference of the relation was thus strengthened by the task. A more dynamic approach should take into account the pragmatic choice of *and* over other alternatives, in order to determine whether the worker inferred a causal relation in the first place.

Lastly, the current method assumes that all discourse relations can be made explicit – in our experiments, we only used items where a connective phrase originally existed or can possibly be inserted. However, it is not always possible to insert a connective. For example, there are no explicit markers for ENTITY RELATIONS. Furthermore, there is also the possibility that the two consecutive segments are *unrelated*. The current method has to be extended to identify these cases for practical annotations.

7 Conclusion

We propose a two-step procedure to convert the challenging task of fine-grained implicit discourse relation annotation to an intuitive task that naïve crowd workers can manage. The method can be directly applied to annotate coherence relations in other languages, and crowdsourcing is a time efficient alternative. On top of the discourse annotation, the methodology also allows creation of large connective banks in other languages.

The results from the current studies also indicate that the discourse relation annotations are more representative when they can be characterized by sense distributions. Automatic discourse relation classification is a bottleneck task, and resources annotated with sense distributions allow more informative evaluation by ranking.

We plan to carry out large scale annotation using the two-step approach to build discourse annotated resources in a variety of data.

Acknowledgments

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

References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Vera Demberg, Fatemeh Torabi Asr, and Merel C.J. Scholman. 2017. How compatible are our discourse annotations? Insights from mapping RST-

DT and PDTB annotations. *arXiv preprint arXiv:1704.08893*.

Daisuke Kawahara, Yuichiro Machida, Tomohide Shibata, Sadao Kurohashi, Hayato Kobayashi, and Manabu Sassano. 2014. Rapid development of a corpus with discourse annotations using two-stage crowdsourcing. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 269–278.

Yudai Kishimoto, Shinnosuke Sawada, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. 2018. Improving crowdsourcing-based annotation of japanese discourse relations. In *Proceedings of the 11th Language Resources and Evaluation Conference*.

Alastair Knott. 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. thesis, University of Edinburgh.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Rebecca J Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th Language Resources and Evaluation Conference*.

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. In *Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 87–97.

Florian Pusse, Asad Sayeed, and Vera Demberg. 2016. Lingoturk: managing crowdsourced tasks for psycholinguistics. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 57–61.

Hannah Rohde, Anna Dickinson, Nathan Schneider, Christopher NL Clark, Annie Louis, and Bonnie Webber. 2016. Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. In *Proceedings of the 10th Linguistic Annotation Workshop*, pages 49–58.

Ted Sanders. 2005. Coherence, causality and cognitive complexity in discourse. In *Proceedings of the First International Symposium on the exploration and modelling of meaning*, pages 105–114. University of Toulouse-le-Mirail Toulouse.

Merel C.J. Scholman. 2019. *Coherence relations in discourse and cognition: comparing approaches, annotations and interpretations*. Ph.D. thesis, University of Saarland.

Merel C.J. Scholman and Vera Demberg. 2017a. Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 24–33.

Merel C.J. Scholman and Vera Demberg. 2017b. Examples and specifications that prove a point: Identifying elaborative and argumentative discourse relations. *Dialogue & Discourse*, 8(2):56–83.

A Appendix: Algorithm for the combination of inserted connectives

```
for each insertion pair do
  if free insertion $\in$ connective bank then
    $R1 \leftarrow$ sense(s) of free insertion
  else
    manual check
    if free insertion is connective then
      added to connective bank
      manual sense annotation
      $R1 \leftarrow$ sense(s) of free insertion
    else
      $R1 \leftarrow \emptyset$
    end if
  end if
  if forced insertion = none of these then
    $R2 \leftarrow \emptyset$
  else
    $R2 \leftarrow$ sense(s) of forced insertion
  end if
  $S \leftarrow R1 \cap R2$
end for
```