Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task

Merel C.J. Scholman
Language Science and Technology
Saarland University
Saarbrücken, Germany

Vera Demberg
Computer Science
Saarland University
Saarbrücken, Germany

{m.c.j.scholman,vera}@coli.uni-saarland.de

Abstract

Traditional discourse annotation tasks are considered costly and time-consuming, and the reliability and validity of these tasks is in question. In this paper, we investigate whether crowdsourcing can be used to obtain reliable discourse relation annotations. We also examine the influence of context on the reliability of the data. The results of the crowdsourced connective insertion task showed that the majority of the inserted connectives converged with the original label. Further, the distribution of inserted connectives revealed that multiple senses can often be inferred for a single relation. Regarding the presence of context, the results show no significant difference in distributions of insertions between conditions overall. However, a by-item comparison revealed several characteristics of segments that determine whether the presence of context makes a difference in annotations. The findings discussed in this paper can be taken as preliminary evidence that crowdsourcing can be used as a valuable method to obtain insights into the sense(s) of relations.

1 Introduction

In order to study discourse coherence, researchers need large amounts of discourse-annotated data, and these data need to be reliable and valid. However, manually coding coherence relations is a difficult task that is prone to individual variation (Spooren and Degand, 2010). Because the task requires a large amount of time and resources, researchers try to find a balance between obtaining reliable data and sparing resources. This has led to the standard practice of using two trained, expert annotators to code data.

Not only is this procedure time-consuming and therefore costly, it also raises questions regarding the reliability and validity of the data. When using trained, expert annotators, they may agree because they share implicit knowledge and know the purpose of the research well, rather than because they are carefully following instructions (Artstein and Poesio, 2008; Riezler, 2014). Krippendorff (2004) therefore notes that the more annotators participate in the process and the less expert they are, the more likely they can ensure the reliability of the data.

In this paper, we investigate how useful crowdsourcing can be in obtaining discourse annotations. We present an experiment in which subjects were asked to insert ("drag and drop") a connecting phrase from a pre-defined list between the two segments of coherence relations. By employing non-trained, non-expert (also referred as naïve) subjects to code the data, large amounts of data can be coded in a short period of time, and it is ensured that the obtained annotations are independent and do not rely on implicit expert knowledge. Instead, the task allows us to tap into the naïve subjects' interpretations directly.

However, crowdsourcing has rarely been used to obtain discourse relation annotations. This could be due to the nature of crowdsourcing: Typically, crowdsourced tasks are small and intuitive tasks. Under these conditions, crowdsourced annotators – unlike expert annotators or in-lab naïve annotators – cannot be asked to code according to a specific framework because this would require them to study manuals. Therefore, rather than asking for relation labels, we ask them to insert a connective from a predefined list. In order to ensure that these connectives are not ambiguous (Asr and Demberg, 2013), we chose connectives based

Proceedings of the 11th Linguistic Annotation Workshop, pages 24–33,
Valencia, Spain, April 3, 2017.

on a classification of connective substitutability by Knott and Dale (1994). We investigate how reliable the obtained annotations are by comparing them to expert annotations from two existing corpora.

Moreover, we examine the effect of the design of the task on the reliability of the data. Researchers agree that discourse relations should be supplied with linguistic context in order to be annotated reliably but there are no clear guidelines for how much context is needed. The current contribution experimentally examines the influence of context on the interpretation of a discourse relation, with a specific focus on whether there is an interaction between characteristics of the segment and the presence of context.

The contributions of this paper include the following:

- We evaluate a new crowdsourcing method to elicit discourse interpretations and obtain discourse annotations, showing that such a task has the potential to function as a reliable alternative to traditional annotation methods.
- The distributions of inserted connectives per item reveal that, often, annotators converged on two or three dominant interpretations, rather than one single interpretation. We also found that this distribution is replicable with high reliability. This is evidence that relations can have multiple senses.
- We show that the presence of context led to higher annotator agreement when (i) the first segment of a relation refers to an entity or event in the context, or introduces important background information; (ii) the first segment consists of a deranked subordinate clause attaching to the context; or (iii) the context sentence following the relation expands on the second argument of the relation. This knowledge can be used in the design of discourse relation annotation tasks.

2 Background

In recent years, several researchers have set out to investigate whether naïve coders can also be employed to annotate data. Working with such annotators has the practical advantage that they are easier to come by, and it is therefore also easier to employ a larger number of annotators, which decreases the effect of annotator bias (Artstein and Poesio, 2005; Artstein and Poesio, 2008). Studies employing naïve annotators have found high agreement between these annotators and expert annotators for Natural Language tasks (e.g., Snow et al., 2008). Classifying coherence relations, however, is considered to be a different and especially difficult type of task due to the complex semantic interpretations of relations and the fact that textual coherence does not reside in the verbal material, but rather in the readers' mental representation (Spooren and Degand, 2010). Nevertheless, naïve annotators have recently also been employed successfully in coherence relation annotation tasks (Kawahara et al., 2014; Scholman et al., 2016) and connective insertion tasks (Rohde et al., 2015, 2016) similar to the one reported in this paper.

Rohde et al. (2016) showed that readers can infer an additional reading for a discourse relation connected by an adverbial. By obtaining many observations for a single fragment rather than only two, they were able to identify patterns of co-occurring relations; for example, readers can often infer an additional causal reading for a relation marked by *otherwise*. These results highlight a problem with double-coded data: Without a substantial number of observations, differences in annotations might be written off as annotator error or disagreement. In reality, there might be multiple interpretations for a relation, without there being a single correct interpretation. The connective insertion method used by Rohde et al. (2016) is therefore more sensitive to the possibility that relations can have multiple readings.

The current study uses a similar method as Rohde et al. (2016), but applies it to answer a different type of question. Rohde et al. (2016) investigated whether readers can infer an additional sense for a pair of sentences already marked by an adverbial. They did not have any expectations on whether there was a correct answer; rather, they set out to identify specific patters of connective insertions. In the current study, we investigate whether crowdsourcing can be used to obtain annotated data that is similar in quality to data annotated by experts. Crucially, we assume that there is a correct answer, namely the original label that was assigned by expert annotators. We therefore will compare the results from the current study to the original annotations in order to evaluate the usability of the connective insertion method for dis-

course annotation.

The design of the current study also differs from other connective-based annotation approaches such as Rohde et al. (2016) and the Penn Discourse Treebank (PDTB, Prasad et al., 2008) in that the connectives were selected to unambiguously mark a specific type of relation. Certain connectives are known to mark different types of relations, such as *but*, which can mark CONTRAST, CONCESSION and ALTERNATIVE relations. In the current study, we excluded such ambiguous connectives in order to be able to derive relation types from the insertions. For example, the connecting phrase AS AN ILLUSTRATION is taken to be a dominant marker for INSTANTIATION relations. The procedure for selecting phrases will be explained in Section 3.

Given the limited amount of research into using naïve subjects for discourse relation annotation, it is important to investigate how this task should be designed. One aspect of this design is the inclusion of context. The benefits of context are widely acknowledged in the field of discourse analysis. Context is necessary to ground the discourse being constructed (Cornish, 2009), and the interpretation of any sentence other than the first in a discourse is therefore constrained by the preceding context (Song, 2010). This preceding context has significant effects on essential parts of discourse annotation, such as determining the rhetorical role each sentence plays in the discourse, and the temporal relations between the events described (Lascarides et al., 1992; Spooren and Degand, 2010). The knowledge of context is therefore assumed to be a requirement for discourse analysis.

Although researchers agree that relations should be supplied with linguistic context in order to be annotated reliably, there are no clear guidelines for how much context is needed. As a result, studies have diverged in their methodology. For some annotation experiments, coders annotate the entire text (e.g., Rehbein et al., 2016; Zufferey et al., 2012). In these cases, they automatically take the context of the relation at hand into account when they annotate a text linearly. By contrast, in experiments where the entire text does not have to be annotated, or the task is split into smaller tasks for crowdsourcing purposes, the relations (or connectives) are often presented with a certain amount of context preceding and following the segments under investigation (e.g., Hoek and Zufferey, 2015; Scholman et al., 2016).

Knowing how much context is minimally needed to be able to reliably annotate data will save resources; after all, the less context annotators have to read, the less time they need to spend on the task. The goal of the current experiment is therefore to test the reliability of crowdsourced discourse annotations compared to original corpus annotations, as well as the effect of context on the reliability of the task.

3 Method

Participants were asked to insert connectives into coherence relations. The items were divided into several batches. Each batch contained items with context or without context, but these two types were not mixed.

3.1 Participants

167 native English speakers completed one or more batches of this experiment. They were recruited via Prolific Academic and reimbursed for their participation (2 GBP per batch with context; 1.5 GBP per batch without context). Their education level ranged between an undergraduate degree and a doctorate degree.

3.2 Materials

The experimental passages consisted of 192 implicit and 42 explicit relations from Wall Street Journal texts. These relations are part of both the Penn Discourse Treebank (PDTB, Prasad et al., 2008) and the Rhetorical Structure Theory Discourse Treebank (RST-DT, Carlson et al., 2003), and therefore carry labels that were assigned by the respective expert annotators at the time of the creation of the corpora. The following types of relations were included: 24 CAUSE, 24 CONJUNCTION (additive), 36 CONCESSION, 36 CONTRAST, 54 INSTANTIATION and 60 SPECIFICATION relations. For the first four relation types, the PDTB and RST-DT annotators were in agreement on the label. The latter two types were chosen to accommodate a related experiment, and for most of these, the PDTB and RST-DT annotators were not in agreement. Lower agreement on these relations is therefore also expected in the current experiment.

The 234 items were divided into 12 batches, with 2 CAUSE, 2 CONJUNCTION, 3 CONCESSION, 3 CONTRAST, 4 or 5 INSTANTIATION and

5 SPECIFICATION items per batch. Order of presentation of the items per batch was randomized to prevent order effects. Subjects were allowed to complete more than one batch, but saw every item only once. Average completion time per batch was 16 minutes with context and 12 minutes without context. Due to presentation errors in one CONJUNCTION, two CAUSE, and two CONCESSION items, the final dataset for analysis consists of 229 items.

Connecting phrases – Subjects were presented with a list of connectives and asked to insert the connective that best expresses the relation holding between the textual spans. The connectives were chosen to distinguish between different relation types as unambiguously as possible, based on an investigation on connective substitutability by Knott and Dale (1994). The list of connecting phrases consisted of: *because*, *as a result*, *in addition*, *even though*, *nevertheless*, *by contrast*, *as an illustration* and *more specifically*.

3.3 Procedure

The experiment was hosted on LingoTurk (Pusse et al., 2016). Participants were presented with a box with predefined connectives followed by the text passage. In the context condition, the passage consisted of black and grey sentences. The black sentences were the two arguments of the coherence relation, and the grey sentences functioned as context (two sentences preceding and one following the relation). Subjects were instructed to choose the connecting phrase that best reflected the meaning between the black text elements, but to take the grey text into account. In the no-context condition, the grey sentences were not presented or mentioned.

Punctuation markers following the first argument of the relation were replaced by a double slash (//, cf. Rohde et al., 2015) to avoid participants from being influenced by the original punctuation markers.

In between the two arguments of the coherence relation was a box. Participants were instructed to "drag and drop" the connecting phrase that "best reflected the meaning of the connection between the arguments" (cf. Rohde et al., 2015) into this green box. Participants could also choose two connecting phrases using the option "add another connective". Moreover, they could manually insert a connecting phrase by clicking "none of these". Participants were allowed to complete more than one batch, but they were never able to complete the same batch in both conditions.

4 Results

Prior to analysis, 5 participants from the context condition and 4 participants from the no-context condition were removed from the analysis because they had very short completion times (<10 minutes for 20 passages of 5 sentences each; <5 minutes for 20 passages of 2 sentences each) and showed high disagreement compared to other participants. The following analyses do not take the responses of these participants into consideration. In total, each list was completed by 12 to 14 participants.

As with any discourse annotation task, some variation in the distribution of insertions can be expected. We are therefore interested in larger shifts in the distribution of insertions. To evaluate these distributions, we report percentages of agreement (cf. De Kuthy et al., 2016). Typically, annotation tasks are evaluated using Cohen's or Fleiss' Kappa (Cohen, 1960; Fleiss, 1971). However, Kappa is not suitable for the current task because it assumes that all coders annotate all fragments.

Participants were given the option of inserting two connecting phrases if they thought that both phrases reflected the meaning of the relation. 3.4% of all answers consisted of two connecting phrases. For most items that received a double insertion, only one answer consisted of a double insertion. The data on multiple insertions therefore does not allow us to draw any strong conclusions. This will be elaborated on in the discussion.

2% of all insertions were manual answers. There was no clear pattern in these manual answers: Only a few items received manual answers, and these items usually received at most two manual answers. An additional 1% of the data consisted of 'blank insertions': Subjects used the 'manual answer' option to not insert anything. As with the manual answers, there was no clear pattern. We aggregate the class 'manual answer' and 'no answer' for our analyses.

We also aggregated frequencies of the connectives that fell into the same class: *because* and *as a result* were aggregated as causal connectives, and *even though* and *nevertheless* were aggregated as concessive connectives.

In the next section, we first show evidence that

the method is reliable. We then turn to the reliability of the no-context condition in comparison to the context condition to be able to determine whether the presence of context led to higher agreement on the sense(s) of items. Finally, we look at the entropy per item and per condition.

4.1 Overall reliability

The results showed that the method is successful: The connectives inserted by the participants are consistent with the original annotation. This is shown in Figure 1a, with the bars reflecting the inserted connective per original class and condition. Figure 1b shows this distribution in more detail by displaying the percentage of inserted connectives per item for the context condition. The distribution for the no-context condition is not included since it is almost identical to the distribution of the context condition. Every stacked bar on the x-axis represents an item; the colours on the bars represent the inserted connective.

These visualizations reveal several trends. First, for CAUSE and CONCESSION relations, the insertions often converge with the original label. 78% of the inserted connectives in items with a causal original label were causal connectives, and 67% of the inserted connectives in concessive items were concessive connectives. For both classes, the second most frequent category of inserted connectives was the other class: For CAUSE, the second most frequent category was CONCESSION (10%), and for CONCESSION, the second most frequent category was CAUSE (15%). On closer inspection of the items, we find that the disagreement between crowdsourced annotations and original annotations can be traced back to difficulties with specific items, and not to unreliability of the workers: The main cause for the confusion of causal and concessive relations can be attributed to the lack of context and/or background knowledge, especially for items with economic topics. For these topics, it can be very hard to judge whether a situation mentioned in one segment is a consequence of the other segment, or a denied expectation.

The second pattern that Figure 1a reveals concerns the classes CONJUNCTION and CONTRAST. The distribution of inserted connectives for these classes look similar: The expected marker is used most often (40% and 44%, respectively), with the corresponding causal relation as the second most frequent inserted connective type (27% causal insertions and 32% concessive insertions, respectively). A closer look at the annotations for items in these classes reveals that this is due to genuine ambiguity of the relation. For relations originally annotated as additive, we find that oftentimes a causal relation can also be inferred. The same explanation holds for CONTRAST relations: Relations from this class that often receive concessive insertions are characterized by the reference to contrasting expectations. Some confusion between these relations is expected, as it is known that concessive and contrastive relations are relatively difficult to distinguish even for trained annotators (see, for example, Robaldo and Miltsakaki, 2014; Zufferey and Degand, 2013).

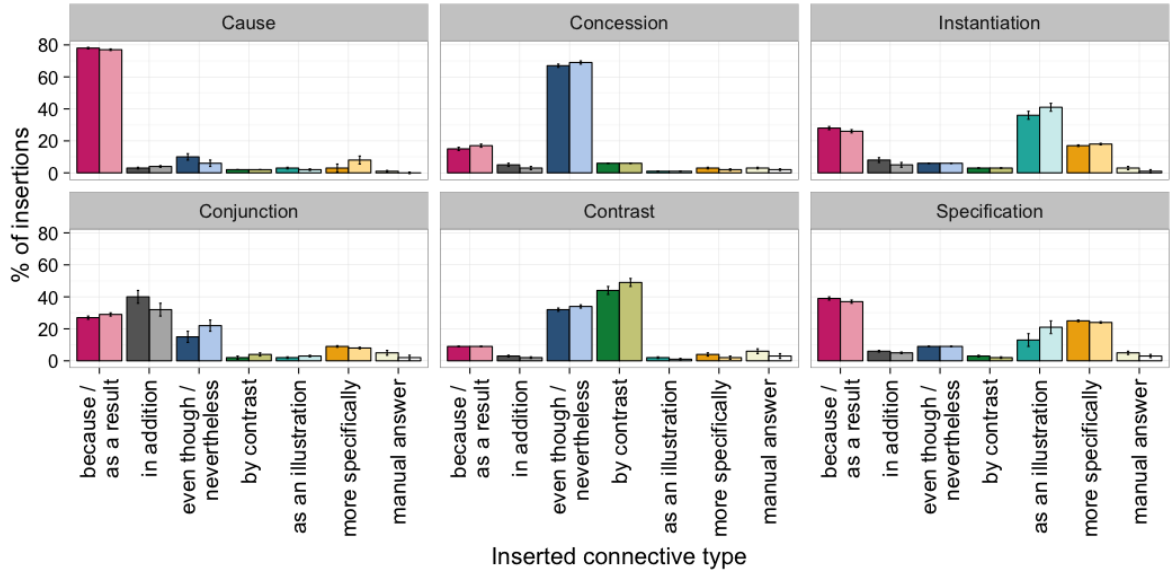
Finally, looking at INSTANTIATION and SPECIFICATION relations, we can see that there is more variety in terms of which connective participants inserted. This was expected, as these relations were chosen because original PDTB and RST annotators did not agree on them.

Looking at the no-context condition in Figure 1a, we find a near-perfect replication of the insertions in the context condition. This is further evidence for the reliability of the task. On average, the difference between the conditions on agreement with the original label differed only by 3.7%. Fisher exact tests showed no significant difference in the distribution of responses between conditions for any of the original classes (*Cause*: $p = .61$; *Conjunction*: $p = .62$; *Concession*: $p = .98$; *Contrast*: $p = .88$; *Instantiation*: $p = .93$; *Specification*: $p = .85$).

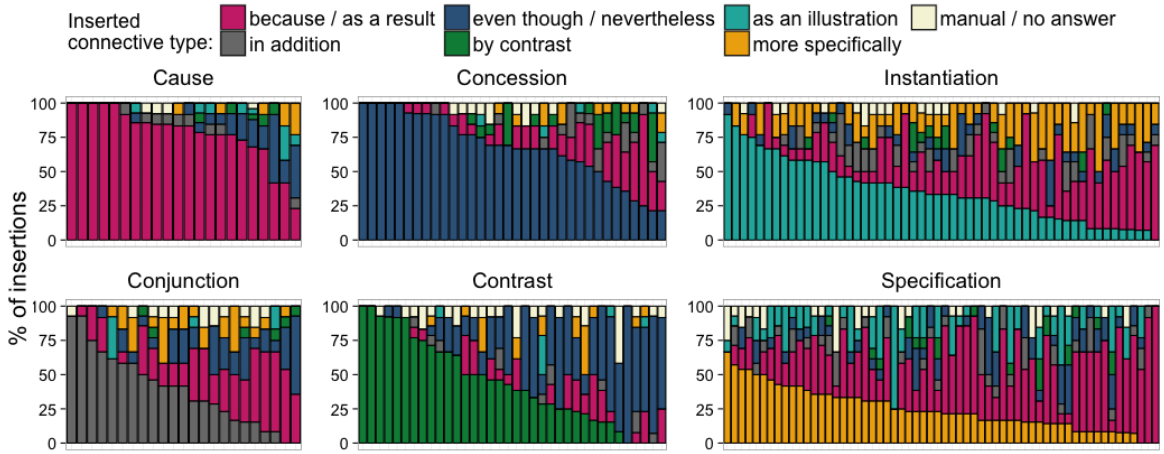
Another notable pattern, shown in Figure 1b, is that items often did not receive only one type of inserted connective; rather, they received multiple types of insertions. For INSTANTIATION and SPECIFICATION items, for example, participants often converged on two senses: Both the originally annotated sense, as well as a causal reading. This indicates that multiple interpretations are possible for a single relation.

Another way to analyse the data is to assign to each relation the label corresponding to the connective that was inserted most frequently by our participants (in Figure 1b, this corresponds to the largest bar per item). We can then calculate agreement between the dominant response per item and the original label. These results are reported in Table 1.

Table 1 shows that the dominant response con-



(a) Distributions (%) of inserted connectives per original class. For every type of insertion, darker colours represent the context condition and lighter colours represent the no-context condition.



(b) By-item distributions (%) for the context condition. Every bar represents a single item; the colours on the bars represent the inserted connective. Plots are arranged according to the amount of dominant insertions corresponding to the original label.

Figure 1: Distributions (%) of inserted connectives per original class.

Original class	Context	No context
CAUSE	91	95
CONJUNCTION	52	35
CONCESSION	85	79
CONTRAST	53	58
INSTANTIATION	54	46
SPECIFICATION	25	20

Table 1: Percentage agreement between the original label and the dominant response per condition.

verges with the original label often for CAUSE and CONCESSION relations and a majority of the time for CONJUNCTION (in the context condition), CONTRAST and INSTANTIATION relations (in the context condition). The dominant response for SPECIFICATION items hardly converges with their original classification. This is as expected, as PDTB and RST-DT annotators also showed little agreement on SPECIFICATION relations.

Looking at the effect of context, we see that agreement between the dominant response and the original label is slightly higher when context is present for four of six types of relations (CONCESSION, INSTANTIATION and SPECIFICATION relations). For CONJUNCTION relations, the agree-

ment is even 17% higher in the context condition compared to the no-context condition. These results suggest that presence of context does have an influence on the subjects' interpretations of the relations. In the next sections, we will look at the distribution of individual items in more detail.

4.2 Effect of context: Dominant response per item

For 9% of the items, the dominant response shifts from one category to another depending on the presence of context. Manual inspection of these items revealed several characteristics that they have in common. First, it was found that often the topic is introduced in the context, and the (lack of) knowledge of the topic influenced the subject's interpretation of the relation. This is illustrated using the following CONJUNCTION example:

(1) Quite the contrary – it results from years of work by members of the National Council on the Handicapped, all appointed by President Reagan. You depict the bill as something Democratic leaders "hoodwinked" the administration into endorsing.
Arg1: The opposite is true: It's the product of many meetings with administration officials, Senate staffers, advocates, and business and transportation officials //
Arg2: many congressmen are citing the compromise on the Americans With Disabilities Act of 1989 as a model for bipartisan deliberations.
Most National Council members are themselves disabled or are parents of children with disabilities. wsj_694

In Example 1, the context introduces the topic. The first argument (Arg1) then presents one argument for the claim that the bill results from years of hard work (as mentioned in the context), and the second argument (Arg2) is another argument for this claim. However, without the context, Arg2 can be taken as a result of Arg1. While this interpretation might be true, it does not seem to be the intended purpose of the relation. In the context condition, subjects interpreted the relation as a CONJUNCTION relation (58% of insertions were *in addition*. In the no-context condition, however, the dominant response was causal (58% of insertions), and the conjunctive *in addition* only accounted for 17% of all insertions.

Another common characteristic in items for which the presence or absence of context changes the dominant response, is that the context sentence following the relation expands on Arg2, thereby changing the probability distribution of that relation. This is common in INSTANTIATION and SPECIFICATION relations, where the second argument provides an example or specification of Arg1. Often, the sentence following Arg2 also provides an example or further specification, which emphasizes the INSTANTIATION/SPECIFICATION sense of the relation between Arg1 and Arg2. However, in relations for which Arg2 can also be seen as evidence for Arg1, the following context sentence can also function to emphasize the causal sense of the relation by expanding on the argument in Arg2. Consider Example 2, taken from the class SPECIFICATION.

(2) Like Lebanon, and however unfairly, Israel is regarded by the Arab world as a colonial aberration. Its best hope of acceptance by its neighbours lies in reaching a settlement with the Palestinians.
Arg1: Like Lebanon, Israel is being remade by demography //
Arg2: in Greater Israel more than half the children under six are Muslims.
Within 25 years Jews will probably be the minority. wsj_1141

In this example, the context sentence following Arg2 expands on Arg2. Together, they convey the information that although Jews are the majority now, within 25 years Muslims will be the majority. Without the context, one could imagine that the text would go on to list more instances of how the demography is changing. Subjects in the no-context condition indeed seem to have interpreted it this way: 75% of the inserted connective phrases were *as an illustration*, and the remaining insertions were *even though* and *because*. By contrast, in the context condition subjects mainly interpreted a causal relation (64% of insertions), together with the specification sense (17%). The marker *as an illustration* only accounted for 7% of completions. Hence, with context present subjects interpreted Arg2 as providing evidence for Arg1, but without context it was interpreted as an INSTANTIATION relation.

4.3 Effect of context: Entropy per item

Another way of analyzing the influence of the presence of context on the participants' response, is to look at the entropy of the distribution of insertions. In the context of the current study, entropy is defined as a measure of the consistency of connective insertions. When the majority of insertions for a certain item are the same, the entropy will be low, but when a certain item receives many different types of insertions, the entropy will be high.

For every item, we calculated Shannon's entropy. We then compared the conditions to determine whether entropy of an item increased or decreased depending on the presence of the context. Here we discuss items that have a difference of at least 1 bit of entropy between the conditions. This set consists of 18 items. Interestingly, presence of context only leads to lower entropy (higher agreement) in 10 items. For the other 8 items, subjects showed more agreement when the context was not presented.

When context is beneficial An analysis of items for which presence of context led to higher agreement has revealed two common characteristics. First, similar to what we found in the previous section, presence of context is helpful when the context introduces important background information, or when the first argument refers to an entity or event in the context.

Second, we observed that agreement was higher in the context condition when Arg1 consists of a subordinate clause that attaches to another clause in the context. In these cases, the dependancy of Arg1 to the context possibly hinders a correct interpretation of Arg1. Consider the following SPECIFICATION relation:

(3) The spun-off concern "clearly will be one of the dominant real estate development companies with a prime portfolio," he said. For the last year, Santa Fe Pacific has redirected its real estate operations toward longer-term development of its properties,
Arg1: hurting profits that the parent had generated in the past from periodic sales from its portfolio //
Arg2: real estate operating income for the first nine months fell to $71.9 million from $143 million a year earlier.
In a statement late yesterday, Santa Fe Pacific's chairman, Robert D. Krebs, said that Santa Fe Pacific Realty would repay more than $500 million in debt owed to the parent before the planned spinoff. wsj_1330

In this example, Arg1 is a deranked subordinate clause, which cannot be used as an independent clause. All subjects in the context condition inserted a causal connective. However, in the no-context condition only 58% inserted a causal connective, and 33% of inserted connectives were *in addition*. Hence, the dominant response remained the same, but the amount of agreement decreased when the context was absent.

When context is disadvantageous Of the 8 items for which absence of context led to more agreement, 7 had a common characteristic: The relation between the context and Arg1 is not strong, for example because Arg1 is also the start of a new paragraph, or because there is a topic change. It is likely that in these cases, the presence of context took the focus away from the relation.

5 Discussion

The annotations obtained using the connective insertion task have the potential to better reflect the average readers' interpretations because the naïve annotators don't rely on implicit expert knowledge. Moreover, it is easier, more affordable and faster to obtain many annotations for the same item via crowdsourcing than via traditional annotation methods. Collecting a large number of annotations for the same item furthermore reveals a probability distribution over relation senses. This can give researchers more insight into the readings of ambiguous relations, and into how dominant each sense is for a specific relation.

The procedures of traditional annotation methods often lead to implicit annotation biases that are implemented to achieve inter-annotator agreement (see, for example, Rehbein et al., 2016). However, annotations that contain biases are less useful from a linguistic or machine learning perspective, as relevant information about a second or third interpretation is obscured. Asking a single, trained annotator to annotate several senses also does not solve this issue: The annotations would still depend on expert knowledge and the annotation process would take more time. In this paper, we have shown that crowdsourcing can be a solution.

However, it should be noted that the design of the experiment was somewhat simplified compared to traditional annotation tasks, largely due to

two factors. First, all items were known to be related to one of the six senses under investigation, that is, participants were not presented with items that did not actually contain a relation (similar to PDTB's NOREL), or that belonged to a different class from those under investigation (for example, TEMPORAL relations). A second constraint on the current study is that participants were presented with tokens that only marked the six classes under investigation. Including more classes and therefore also more connectives in an annotation study could result in lower agreement between the coders. Future research will therefore focus on whether other relations (including NOREL) can also be annotated reliably by naïve coders.

Crowdsourcing the data also presents possible confounding factors for the design of an annotation study. More specifically, one has to be aware of the effect of motivation on the results. For example, we found that the participants rarely inserted multiple connectives for the same relation. It is possible that motivation played a role in this. Participants were only required to insert one connecting phrase; the second one was optional. Since inserting a second phrase takes more time, participants might have neglected to do so, even if they interpreted multiple readings for some relations. For future experiments, this effect can be avoided by asking subjects to explicitly indicate that they don't see a second reading.

Regarding the influence of context, the findings from our experiment do not support the general consensus that presence of context is a necessary requirement for discourse annotation. The lack of a clear positive effect of context on agreement could be due to general ambiguity of language. As Spooren and Degand (2010) note, "establishing a coherence relation in a particular instance requires the use of contextual information, which in itself can be interpreted in multiple ways and hence is a source of disagreement." Nevertheless, we do suggest to include context in discourse annotation tasks if time and resources permit it. Generally context does not lead to worse annotations when the fragments are presented in their original formatting, and the presence of context might facilitate the inference of the intended relation.

6 Conclusion

The current paper addresses the question of whether a crowdsourcing connective insertion task can be used to obtain reliable discourse annotations, and whether the presence of context influence the reliability of the data.

Regarding the influence of context, the results showed that the presence of context influenced the annotations when the fragments contained at least one of the following characteristics: (i) the context introduced the topic, (ii) the context sentence following the relation expands on the second argument of the relation; or (iii) the first argument of the relation is a subordinate clause that attaches to the context. The presence of context led to less agreement when the connection between the context and the first argument was not strong due to a paragraph break or a topic change.

Regarding the reliability of the task, we found that the method is reliable for acquiring discourse annotations: The majority of inserted connectives converged with the original label, and this convergence was almost perfectly replicable, in the sense that a similar pattern was found in both conditions. The results also showed that subjects often converged on two types of insertions. This indicates that multiple interpretations are possible for a single relation. Based on these results, we argue that annotation by many (more than 2) annotators is necessary, because it provides researchers with a probability distribution of all the senses of a relation. This probability distribution reflects the true meaning of the relation better than a single label assigned by an annotator according to a specific framework.

Acknowledgements

This research was funded by the German Research Foundation (DFG) as part of SFB 1102 "Information Density and Linguistic Encoding" and the Cluster of Excellence "Multimodal Computing and Interaction" (EXC 284).

References

Ron Artstein and Massimo Poesio. 2005. Bias decreases in proportion to the number of annotators. *Proceedings of the Conference on Formal Grammar and Mathematics of Language (FG-MoL)*, pages 141–150.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Fatemeh Torabi Asr and Vera Demberg. 2013. On the information conveyed by discourse markers. In *Pro-*

ceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics, pages 84–93.

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.

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

Francis Cornish. 2009. "Text'' and "discourse" as context. *Working Papers in Functional Discourse Grammar (WP-FDG-82): The London Papers I, 2009*, pages 97–115.

Kordula De Kuthy, Ramon Ziai, and Detmar Meurers. 2016. Focus annotation of task-based data: Establishing the quality of crowd annotation. In *Proceedings of the Linguistic Annotation Workshop (LAW X)*, pages 110–119.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.

Jet Hoek and Sandrine Zufferey. 2015. Factors influencing the implicitation of discourse relations across languages. In *Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-11)*, pages 39–45. TiCC, Tilburg center for Cognition and Communication.

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 International Conference on Computational Linguistics (COLING)*, pages 269–278.

Alistair Knott and Robert Dale. 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62.

Klaus Krippendorff. 2004. Reliability in content analysis. *Human communication research*, 30(3):411–433.

Alex Lascarides, Nicholas Asher, and Jon Oberlander. 1992. Inferring discourse relations in context. In *Proceedings of the 30th annual meeting on Association for Computational Linguistics*, pages 1–8. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. Citeseer.

Florian Pusse, Asad Sayeed, and Vera Demberg. 2016. LingoTurk: Managing crowdsourced tasks for psycholinguistics. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Ines Rehbein, Merel C. J. Scholman, and Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Portoroz, Slovenia. European Language Resources Association (ELRA).

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

Livio Robaldo and Eleni Miltsakaki. 2014. Corpus-driven semantics of concession: Where do expectations come from? *Dialogue & Discourse*, 5(1):1–36.

Hannah Rohde, Anna Dickinson, Nathan Schneider, Christopher 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 Linguistic Annotation Workshop (LAW X)*, pages 49–58.

Merel C. J. Scholman, Jacqueline Evers-Vermeul, and Ted J. M. Sanders. 2016. Categories of coherence relations in discourse annotation: Towards a reliable categorization of coherence relations. *Dialogue & Discourse*, 7(2):1–28.

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 (EMNLP)*, pages 254–263. Association for Computational Linguistics.

Lichao Song. 2010. The role of context in discourse analysis. *Journal of Language Teaching and Research*, 1(6):876–879.

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

Sandrine Zufferey and Liesbeth Degand. 2013. Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory*, 1:1–24.

Sandrine Zufferey, Liesbeth Degand, Andrei Popescu-Belis, and Ted J. M. Sanders. 2012. Empirical validations of multilingual annotation schemes for discourse relations. In *Proceedings of the 8th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation*, pages 77–84.