Dialogue & Discourse 8(2) (2017) 56–83 doi: 10.5087/dad.2017.203

Examples and Specifications that Prove a Point: Identifying Elaborative and Argumentative Discourse Relations

Merel C.J. Scholman M.C.J.SCHOLMAN@COLI.UNI-SAARLAND.DE
Department of Language Science and Technology
Saarland University

Vera Demberg VERA@COLI.UNI-SAARLAND.DE
Department of Computer Science
Department of Language Science and Technology
Saarland University

Editor: David Traum

Submitted 11/2016; Accepted 03/2017; Published online 07/2017

Abstract

Examples and specifications occur frequently in text, but not much is known about how how readers interpret them. Looking at how they're annotated in existing discourse corpora, we find that annotators often disagree on these types of relations; specifically, there is disagreement about whether these relations are elaborative (additive) or argumentative (pragmatic causal). To investigate how readers interpret examples and specifications, we conducted a crowdsourced discourse annotation study. The results show that these relations can indeed have two functions: they can be used to both illustrate / specify a situation and serve as an argument for a claim. These findings suggest that examples and specifications can have multiple simultaneous readings. We discuss the implications of these results for discourse annotation.

Keywords: Coherence relations, crowdsourcing, discourse annotation, inter-annotator agreement, signalling

1. Introduction

Discourse relations (also referred to as coherence relations) are semantic links between two (or more) discourse units (cf. Hobbs, 1979; Mann & Thompson, 1988; Sanders, Spooren & Noordman, 1992). They can be explicitly signalled by connectives (also referred to as discourse markers, or DMs) such as *because* or *for example*. However, many relations are implicit, that is, they are not marked by a connective. This is illustrated in Example 1.

(1) Packaging has some drawbacks. The additional technology, personnel training and promotional effort can be expensive. wsj_0085

In order to be able to do quantitative investigations of discourse relations, text corpora with annotated discourse relations are necessary. In recent years, several discourse annotated corpora have been developed; some inspiring examples are the Penn Discourse Treebank (PDTB; Prasad et al., 2008) and the Rhetorical Structure Theory Discourse Treebank (RST-DT; Carlson, Marcu & Okurowski, 2003). Both corpora contain articles from the Wall Street Journal, and for a substantial amount of articles, annotations from both resources are available. In order to determine whether

annotations from these corpora are compatible, i.e., whether the frameworks agree on the sense of a relation, Demberg, Asr & Scholman (2017) mapped the annotations of overlapping texts of these two treebanks onto each other. We will refer to the texts with annotations from both discourse corpora as the "WSJ-Aligned corpus". The mapping allows for a comparison between the labels from PDTB and RST-DT, which reveals how much the two frameworks agree with each other as well as potentially interesting patterns of disagreement. While there is usually no one-to-one correspondence between PDTB and RST-DT relation labels; it is possible to determine the correspondences between labels based on the annotation guidelines of the two frameworks (also see Sanders et al., 2016, Submitted). Demberg et al. (2017) found that inter-framework agreement on explicit relations was reasonable (ca. 60% agreement), while agreement on implicit relations was much lower (roughly 35% agreement). The difference in the agreements in annotation between explicit and implicit relations can be explained by the presence of connectives: Achieving high inter-annotator agreement on implicit relations is harder than on explicit relations because annotators cannot rely on connectives, which represent a strong cue.

But even when taking into account the difficulty of annotating discourse relations, the level of agreement in annotation between frameworks on the same texts might seem astonishingly low. The goal of this article is to establish whether there are systematic reasons for the disagreement, whether the different annotations are justified, and what the implications for discourse annotation are. In order to answer these questions, we took a closer look at two types of relations for which inter-framework agreement is particularly poor, and investigate the factors affecting interpretation of these relations in more detail. The relations under investigation are PDTB's INSTANTIATION and SPECIFICATION relations (32% and 14% agreement, respectively). These relations do not have many prototypical connectives and are therefore hard to identify (Taboada & Das, 2013; Vergez-Couret & Adam, 2012).They are nevertheless very frequent: Together, they make up 24% of all implicit relations in the PDTB. Both relations are subtypes of the PDTB class EXPANSION. INSTANTIATION is a second-level label in the PDTB hierarchy, whereas SPECIFICATION is a third-level label in the PDTB type RESTATEMENT (see Appendix A for the PDTB and RST-DT taxonomies). For convenience, EXPANSION.RESTATEMENT.SPECIFICATION will be referred to as SPECIFICATION in this paper. In both of these relation types, one segment further specifies a set or situation described in the other segment (Halliday, 1994). INSTANTIATIONS and SPECIFICATIONS are by definition primarily considered as elaborative (also referred to as additive) kinds of relations, i.e. they are discourse relations that connect utterances describing the same situation from different angles (Jasinskaja, 2013). A large proportion of the mismatch with the corresponding RST-DT annotations stems from these same pairs of arguments being annotated as argumentative (causal) relations; namely EXPLANATION-ARGUMENTATIVE and EVIDENCE. In such argumentative relations, one segment gives support to the other segment (Jasinskaja & Karagjosova, 2011).

Blakemore (1997) argues that INSTANTIATIONS and SPECIFICATIONS can indeed have two functions: they can be used to both illustrate / specify a situation and serve as an argument to a claim. PDTB allows the annotation of more than one relation sense, but we found that the argumentative function was usually not annotated. Annotators could also assign an argumentative label such as CONTINGENCY.PRAGMATIC CAUSE or a causal label such as CONTINGENCY.CAUSE, but in practice, PDTB rather focuses on the illustrative / specification aspect of the relation. If readers do indeed infer the argumentative reading of the INSTANTIATIONS and SPECIFICATIONS, it is however important that this reading is also reflected in the label that the relation receives in a corpus. Otherwise, the annotation of those INSTANTIATIONS and SPECIFICATIONS cannot be considered as

fully descriptively adequate. Similarly, when RST-DT annotates the argumentative aspect of the relation (by labeling relations as EXPLANATION-ARGUMENTATIVE and EVIDENCE), the illustrative / specification aspect of the relation does not become visible in the annotations (which is reflected by the labels EXAMPLE and GENERAL-SPECIFIC).

Considering that the labels INSTANTIATION and SPECIFICATION occur so often, the current study sets out to investigate how readers actually interpret these relations. Do the instances of INSTANTIATION and SPECIFICATION relations from the PDTB vary in the degree to which they can be interpreted as argumentative? Can we identify specific characteristics of the relations that have additional argumentative functions? Are there also other alternative interpretations that are inferred?

We aim to answer these questions by asking participants via a crowdsourcing platform to insert connectives from a predefined list that are considered to be relatively unambiguous in between the segments of coherence relations. The insights from this investigation can be used to help to better understand the functions of INSTANTIATION and SPECIFICATION relations, and, for annotation purposes, to more reliably identify and classify INSTANTIATION and SPECIFICATION relations, which in turn can improve the agreement on these frequently-occurring classes.

In sum, the current paper deals with the following distinct issues: (i) the mismatch between annotations of INSTANTIATION and SPECIFICATION relations in the PDTB and RST-DT are analysed, (ii) the variability of presumably valid interpretations of INSTANTIATION and SPECIFICATION relations (elaborative vs. argumentative), and (iii) the usability of crowdsourced annotations for these research purposes. The lay-out of the paper is as follows: in the next section, we discuss INSTANTIATION and SPECIFICATION relations and their signalling. We then present a crowdsourcing experiment, showing that INSTANTIATION and SPECIFICATION relations are indeed often interpreted by our participants as argumentative. The paper concludes with a discussion of the results and their implications for discourse annotation.

2. Background

INSTANTIATION and SPECIFICATION both belong to the PDTB class EXPANSION, which is primarily an elaborative class. The difference between the labels is quite subtle. A relation is labeled as INSTANTIATION when "the connective indicates that Arg1 evokes a set and Arg2 describes it in further detail" (Prasad et al., 2007, p. 34).[1] The set described in Arg1 may be a set of events, reasons, behaviors, etc. INSTANTIATIONS are typically marked by *for example* and *as an illustration*. Example (2) illustrates this type of relation. In SPECIFICATION relations, Arg2 also typically describes Arg1 in more detail, but Arg2 is also a logical implication of Arg1. SPECIFICATIONS are typically marked by *specifically* and *in fact*. Example (3) presents an example of a SPECIFICATION relation. The labels are therefore similar in that Arg2 expands on Arg1, but they differ in the content of Arg1 (presence or absence of a set and logical implication).

(2) In an age of specialization, *the federal judiciary is one of the last bastions of the generalist.* **A judge must jump from murder to antitrust cases, from arson to securities fraud, without missing a beat.**[2] wsj_601

1. PDTB uses a lexical grounding approach, for which connectives are annotated for both explicit and implicit relations. In the case of implicit relations, annotators insert a connective and annotate the corresponding sense of the relation.
2. In line with the PDTB, the first segment of a relation (Arg1) will be indicated with italic font, and the second segment (Arg2) will be indicated with bold font.

(3) *In the cornucopia of go-go apples, the Fuji's track record stands out.* **During the past 15 years, it has gone from almost zilch to some 50% of Japan's market.** wsj_1128

The equivalent of these relation labels in RST are EXAMPLE for INSTANTIATION and ELABORATION-GENERAL-SPECIFIC[3] for SPECIFICATION. However, the WSJ-Aligned corpus shows that often PDTB's INSTANTIATION and SPECIFICATION relations fall into other classes in RST-DT. As illustrated in Figure 1, other common labels for INSTANTIATION relations are GENERAL-SPECIFIC (which maps onto PDTB's SPECIFICATION label), ELABORATION-ADDITIONAL (the most basic relation in RST), EXPLANATION-ARGUMENTATIVE, and EVIDENCE (both typically causal labels in RST). These same RST labels were used for relations that received the SPECIFICATION label in PDTB. Hence, RST often assigns causal labels to relations that PDTB labels as elaborative INSTANTIATION or SPECIFICATION (25% of PDTB INSTANTIATION and 21% of PDTB SPECIFICATION relations receive causal labels in RST).

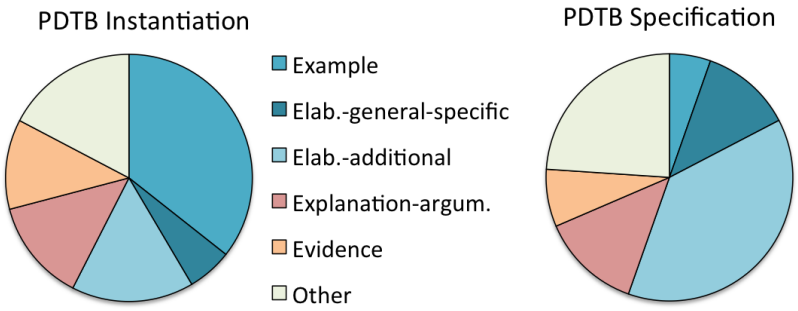

Figure 1: 306 PDTB INSTANTIATION and 426 PDTB SPECIFICATION relations (explicit and implicit) annotated according to RST labels. Blue RST labels are primarily elaborative, orange RST labels are primarily causal.

A similar pattern does not occur for the RST classes EXAMPLE and GENERAL-SPECIFIC: Figure 2 shows that PDTB annotators did not assign a causal label as often to RST's EXAMPLE and GENERAL-SPECIFIC relations (7% and 12%, respectively). Hence, for relations that RST annotators consider to be EXAMPLES and GENERAL-SPECIFICATIONS, PDTB annotators tend annotate the same reading.

A possible explanation for this difference between frameworks reagrding the annotation of INSTANTIATIONS and SPECIFICATIONS is that the frameworks' procedures influence the resulting annotations. PDTB has a connective-based approach: Annotators are instructed to annotate the connective if one is present, or insert one and then annotate the relation, when the relation is implicit. Annotators are not instructed to systematically try to insert certain connectives before trying others. They can choose from a list containing many connectives, and often multiple connectives can be used for the same relation. For example, PDTB annotators inserted the discourse marker *specifically* for Example 3 above, but *because* could have also been inserted in this relation. The choice of inserted connective therefore depends entirely on the annotator, and this can vary from annotator to annotator. It is plausible that different annotators have different biases for inferring one

3. ELABORATION-GENERAL-SPECIFIC will be abbreviated to GENERAL-SPECIFIC in the remainder of this paper.

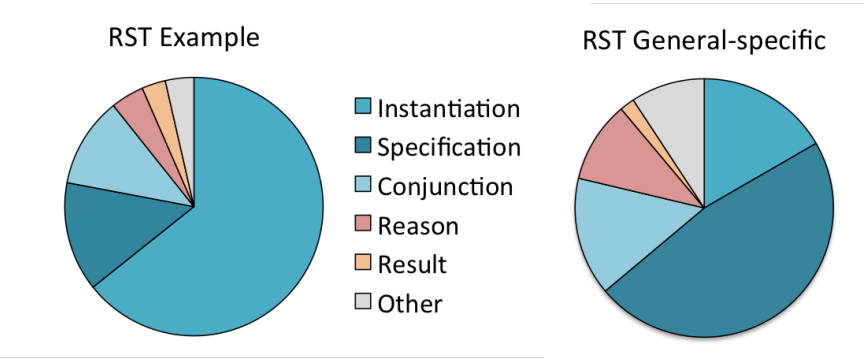

Figure 2: 168 RST EXAMPLE and 108 RST GENERAL-SPECIFIC relations (explicit and implicit) annotated according to PDTB labels.

sense over another, but this is not retraceable in the corpus itself. It is therefore difficult to determine what the framework's bias is.

RST, by contrast, doesn't explicitly make use of connectives at all; rather, the framework instructs annotators to focus on the writer's intentions. From this viewpoint, it is likely that even though *specifically* or *for example* can be used to express the relation, the intention of the writer in fact has an argumentative nature: to convince the reader of a point by providing evidence for a claim. Indeed, PDTB's INSTANTIATION and SPECIFICATION classes contain instances that can be analysed as consisting of a claim and an argument. In the case of INSTANTIATION relations, the segment containing the instantiation can be seen as supporting a claim in the other argument. Consider the following example, annotated as PDTB INSTANTIATION – RST EVIDENCE:

(4) *"If you are born to give parties, you give parties.* **Even in Russia we managed to give parties."** wsj_1367

The first segment of this example is a claim, and the second segment gives an illustration of the claim. This reading is referred to as the ideational (or semantic) reading, which involves the relation between the information conveyed in consecutive elements of a coherent discourse (cf. Moore & Pollack, 1992). Besides this ideational reading of Example 4, the second segment can also be interpreted as a premise that underpins the validity of the claim. With this reading, the relation is in fact pragmatic causal, or argumentative.[4] Argumentative relations are relations in which the writer attempts to affect the addressee's beliefs, attitudes, desires etc. by means of language (cf. Hovy & Maier, 1995).

SPECIFICATION relations have a similar ambiguity between an ideational and an argumentative reading: The second segment can serve as merely providing more information about a concept or situation in the first segment, or it can provide support for a claim in the first segment. This double function of INSTANTIATIONS and SPECIFICATIONS was brought up by Carston (1993, p. 164), who noted that "exemplification is a common way of providing evidence to support a claim, or, equivalently, of giving a reason for believing something." Building on this, Blakemore (1997) ar-

4. The terms ideational and argumentative are also known as informational and intentional (Moore & Pollack, 1992), subject matter and presentational (Mann & Thompson, 1988), and semantic and interpersonal (Hovy & Maier, 1995).

gues that INSTANTIATION and SPECIFICATION relations can have different functions in a text, and that classifying them as only ideational or only argumentative does not do justice to the way that these relations are interpreted. The double function of INSTANTIATION and SPECIFICATION relations has also been noted in other descriptive work on ELABORATION relations (see, for example, Cuenca, 2003; Hyland, 2007; Jasinskaja & Karagjosova, 2011), but it is not reflected in discourse annotation frameworks. Blakemore (1997) however argues that the classification of these relations in frameworks is irrelevant; rather, one should focus on how the utterance achieves relevance with the reader. The current study argues against this claim: The classification of INSTANTIATIONS and SPECIFICATIONS is important considering that corpora should be descriptively adequate.

Returning to the classification of these relations, we can conclude (based on Figures 1 and 2) that PDTB's INSTANTIATION and SPECIFICATION classes and RST's GENERAL-SPECIFIC class contain a mixture of both argumentative and non-argumentative relations. This finding is actually not surprising: It has been noted that these frameworks do not have distinct labels for specific argumentative relations since they are focused on identifying general discourse structures (Peldszus & Stede, 2013; Stab & Gurevych, 2014). Moreover, as Biran & Rambow (2011) note, argumentation is not characterised by a single discourse relation; instead, it can be realised by a large number of discourse relation types. This does not imply that all relation types necessarily have an argumentative and a non-argumentative reading, but it does seem to apply to the INSTANTIATION and SPECIFICATION type of relations. If readers can actually infer an argumentative reading for such non-causal relation types, one could argue that discourse annotation frameworks would need to distinguish an argumentative counterpart of these labels in order to be able to represent the the argumentative reading of those non-causal relations.

Another way of incorporating the argumentative reading of a non-causal relation in its annotation, is to annotate multiple senses for the same relation. Consider Example 4: Relations like this example actually have multiple readings (both elaborative and argumentative), and multiple labels would therefore adequately express the readings of this relation. Unfortunately, the annotated relations that are currently available do not carry double annotations.[5] This is because these corpora are built on the assumption that at most a *single* discourse relation holds between two segments. However, properties of the discourse adverbial *instead* (Webber, 2013) have shown that some relations can actually express multiple meanings. *Instead* is a strong marker for EXCEPTION relations, but when *instead* occurs at the beginning of a sentence, another relation can often be inferred as well, as illustrated in 5-7, taken from Rohde et al. (2016).

(5) *I planned to make lasagna.* Instead **I made hamburgers.**
→ But instead **I made hamburgers.**

(6) *I don't know how to make lasagna.* Instead **I made hamburgers.**
→ So instead **I made hamburgers.**

(7) *Surprisingly, they ignored the lasagna.* Instead **they just ate the salad.**
→ And instead **they just ate the salad.**

As Examples 5-7 show, two segments marked by *instead* can express different relations, including the EXCEPTION relation that the adverbial *instead* marks. Building on these observations, Rohde

5. PDTB does allow for double annotations, but this is rarely applied in practice: less than 5% of instances in the PDTB carry two labels. The PDTB group plans to release a new version, PDTB 3.0, in which double labels occur more frequently (Webber et al., 2016).

et al. (2016) showed that relations marked by discourse adverbials other than *instead* can also express multiple meanings. These results indicate that multiple discourse relations can hold between two discourse segments marked by an adverbial. Based on these findings, we expect INSTANTIATION and SPECIFICATION relations to also have multiple readings.

In sum, the fact that human annotators often disagree on the elaborative or argumentative nature of PDTB's INSTANTIATION and SPECIFICATION relations can be due to the operationalization of the frameworks, or to an actual ambiguity or double-function of these relations that is not captured in PDTB's framework. The current study therefore investigates how comprehenders interpret these relations. The obtained annotations will also be used to identify certain cues that might have influenced the readers' interpretations (in other words, cues that signal the elaborative or argumentative type of INSTANTIATION and SPECIFICATION relations). We will now turn to a more detailed explanation of the methodology.

3. Method

We conducted a crowdsourcing experiment for which naïve (non-trained, non-expert) annotators were asked to insert connectives from a predefined list into coherence relations taken from WSJ-Aligned. Naïve subjects were chosen over expert annotators for several reasons. First, naïve annotators are not influenced by their prior experience with a specific framework, thereby eliminating the possibility of a framework bias or instruction bias occurring in the annotations. Moreover, they are not experts in discourse annotation, and their annotations therefore do not rely on implicit, expert knowledge (Riezler, 2014). Third, working with naïve annotators has the practical advantage that they are easier to come by. This makes it easier to employ a larger number of annotators, which decreases the effect of annotator bias (Artstein & Poesio, 2005). Studies employing naïve annotators have found high agreement between these annotators and expert annotators for Natural Language tasks (for example, Snow, O'Connor, Jurafsky & Ng, 2008). More recently, they have been employed successfully in coherence relation classification tasks (Kawahara, Machida, Shibata, Kurohashi, Kobayashi & Sassano, 2014; Rohde, Dickinson, Clark, Louis & Webber, 2015; Rohde, Dickinson, Schneider, Clark, Louis & Webber, 2016; Scholman, Evers-Vermeul & Sanders, 2016).

However, 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 specific relation labels, we ask them to insert a connective from a predefined list of connective phrases. This is not to say that a single connective cannot mark multiple types of relations. In fact, connectives are well known to be ambiguous and multifunctional (see Asr & Demberg, 2013; Degand, 1998; Hovy, 1995; Maschler & Schiffrin, 2015; Versley, 2011, among many others). In order to ensure that the connectives included in this experiment are as unambiguous as possible, we chose connectives based on a classification of connective substitutability by Knott & Dale (1994). It is assumed that these connectives are typical markers of the relational classes included in this experiment. However, we do not assume that, for example, an instantiation marker cannot also imply an argumentative reading. Rather, the method is based on the assumption that participants choose the connective that matches the strongest reading that they infer. Hence, if they infer an argumentative reading for a specific relation annotated as INSTANTIATION by PDTB annotators, they will choose a causal connective. If they believe both readings hold, they can choose multiple connectives. Similar methodologies have been used to obtain in-

sights into readers' interpretations of relations by Rohde et al. (2015, 2016); Sanders et al. (1992) and Scholman et al. (2016).

Inserting a connecting phrase is of course not the same as assigning a relation label to two arguments. Using connecting phrases to obtain annotations leads to more coarse-grained observations. However, this method can be used to tap into the interpretation of a relation by a large group of annotators, thereby reducing the effect of individual biases. Moreover, the current method reveals a distribution of the senses of the discourse relation at hand, making this method more suitable for reflecting a situation in which a relation can have multiple interpretations (cf. Cuenca & Marín, 2009; Rohde et al., 2015, 2016; Webber et al., 2001).

3.1 Participants

111 native English speakers (47 female) completed one or more batches of this experiment. They were recruited via Prolific Academic and reimbursed for their participation (2 GBP per batch). Participants came from the United States, United Kingdom, Ireland, and Australia. Their education level ranged from an undergraduate degree to a doctorate degree.

3.2 Materials

The experimental passages were implicit INSTANTIATION and SPECIFICATION relations taken from the WSJ-Aligned corpus, which contains fragments from the Wall Street Journal newspaper. These relations were chosen to enable a comparison between the PDTB label and the RST label. The INSTANTIATION and SPECIFICATION relations that were included in this experiment fell into one of four RST classes: EXAMPLE, GENERAL-SPECIFIC, EVIDENCE, and EXPLANATION-ARGUMENTATIVE. These categories were chosen because relations that received PDTB's INSTANTIATION or SPECIFICATION label fall in these four RST classes most often. In total, there were 54 INSTANTIATION and 60 SPECIFICATION items included in this experiment. For each of the four relevant RST classes, 15 INSTANTIATION and 15 SPECIFICATION items were taken, with the exception of the INSTANTIATION – GENERAL-SPECIFIC combination: there were only 9 items the PDTB label INSTANTIATION and the RST label GENERAL-SPECIFIC.

When the WSJ-Aligned corpus contained more than 15 items annotated with the target PDTB and RST label, preference was given to items that (i) differed less than 60 characters between the RST and PDTB segmentation, (ii) did not contain attribution in one of the arguments, and (iii) dealt with a non-economic topic. The first criterion relates to the size of the segments: PDTB and RST have different segmentation rules, resulting in different segment sizes. We chose relations that differed as little as possible in segmentation. In the experiment, we adhered to the PDTB segmentation because the PDTB annotates the minimal amount of information necessary to infer the intended relation. The second criterion relates to attribution in the segments, that is, the explicit reference to the source (e.g., *John said that*). PDTB does not annotate attribution as part of an argument, but rather as a feature of the relation. Therefore, if the source of the attribution was reported in between the two arguments, it was moved to the context sentence following the second argument, to ensure that participants did not treat it as part of the argument. The third criterion was composed for the motivation of participants: Non-economical topics were considered more interesting than economical topics.

5. Crowdsourcing platform, `www.prolific.ac`

The fillers in this experiment consisted of 24 causal, 24 conjunction (additive), 36 concessive and 36 contrastive relations. Causal relations are characterized by the presence of an implication between the two discourse segments; in other words, one segment contains a logical cause for a situation or event in the other segment. Typical connectives are *because* and *so*. Conjunction relations do not have an implication relation and can be expressed by *and* or *also*. Concessive relations are considered as "negative causals" (Konig & Siemund, 2000), since they establish a similar causal relation but differ in their polarity (Sanders et al., 1992). Concessive relations are characterized by the presence of a denial of expectation: one segment contains a consequence despite a situation or event in the other segment. Typical connectives for concessive relations are *even though* and *nevertheless*. Contrastive relations are considered to be negative counterparts of additives (cf. Sanders et al., 1992), since the segments are also connected in a conjunction, but a difference is highlighted between the segments. Prototypical lexical markers for these relations are *but* and *by contrast*.

To ensure that the fillers in this experiment were clear cases of a specific type of relation, they were selected based on the criterion that the PDTB and RST label were in agreement (for example, when PDTB gave an item a causal label but RST gave it an additive label, it was not chosen).[6] All relations included in the study were originally implicit, except for concessive and some contrastive relations, since there were no implicit CONCESSION relations and too few implicit CONTRAST relations in the WSJ-Aligned corpus. Table 1 shows all PDTB and RST labels that these fillers carried per type of relation.

Relation type	**PDTB label**	**RST label**	**Freq.**
Causal	REASON, RESULT	CAUSE-RESULT, CONSEQUENCE, EXPLANATION-ARGUMENTATIVE, REASON, RESULT, EVIDENCE	24
Conjunction	CONJUNCTION	ELABORATION-ADDITIONAL	24
Concessive	CONTRA-EXPECTATION, EXPECTATION	CONCESSION, ANTITHESIS	36
Contrastive	CONTRAST, JUXTAPOSITION, OPPOSITION	CONTRAST, ANTITHESIS	36

Table 1: Type of filler and the PDTB and RST labels that it can carry.

The total of 234 items were divided into 12 batches, with 4 or 5 INSTANTIATION, 5 SPECIFICATION, 2 causal, 2 conjunction, 3 concessive, and 3 contrastive 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. Due to presentation errors in one conjunction, two causal, and two concessive items, the final dataset for analysis including fillers consists of 229 items.

Connecting phrases – The list of connecting phrases consisted of: *as an illustration*, *more specifically*, *in addition*, *because*, *as a result*, *even though*, *nevertheless*, and *by contrast*. With the exception of the first two phrases, these connectives were chosen based on a classification of connectives in Knott and Dale (1994).

6. The RST label ANTITHESIS is ambiguous between a causal and an additive reading. It was nevertheless used as a criterion for both concessive and contrastive items because the subset was too small without the inclusion of RST ANTITHESIS.

Knott & Dale (1994) list *for example* as a typical marker of INSTANTIATION relations. However, we decided to choose *as an illustration* instead. The two connecting phrases are interchangeable, but *for example* is used more commonly than *as an illustration*. We hypothesized that inserting *as an illustration* would therefore require more active reasoning than *for example*.

Knott & Dale (1994) did not list markers for SPECIFICATION relations. In the PDTB, SPECIFICATION relations are most often marked by *in fact* and *indeed*. These markers can also be used to mark causal relations. We decided to choose *more specifically* as a typical marker because it is less ambiguous. *In addition* was chosen over *and* for the same reason: *and* is an underspecified connective and can be used to mark several different types of relations, whereas *in addition* only marks additive relations.

Because and *as a result* were chosen to represent causal relations. Both connectives can be used to mark argumentative relations (Cohen, 1987). *Even though* and *nevertheless* were included as markers of concession relations. *By contrast* was chosen as a typical marker of contrastive relations. Other typical markers, such as *but*, *although* and *however* were not included because they can be used to mark both concessive and contrastive relations. Similar to the order of the items, the order in which the connecting phrases were presented was also randomized for every item.

3.3 Procedure

The experiment was distributed via Prolific Academic and hosted on LingoTurk (Pusse, Sayeed & Demberg, 2016). First, participants were presented with instructions for the study. Next, they were presented with the experiment interface, which consisted of three parts: a short summary of the instructions, a box with predefined connectives, and the text passage (see Figure 3 for an example of the interface). The text passage contained two context sentences preceding the first segment and one context sentence following the second segment. These context sentences were taken from the original text and were not altered in any way. The two segments of the target relation were shown in black text while the context sentences were displayed in grey text. 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 between the two segments of the coherence relation was a green box (see Figure 3). Participants were instructed to "drag and drop" the connecting phrase that "best reflected the meaning of the connection between the sentences" (cf. Rohde et al., 2015) into this green box. Participants could also choose two connecting phrases if both phrases reflected the meaning of the relation, using the option "add another connective". Moreover, they could manually insert a connecting phrase by clicking "none of these", if they felt that none of the predefined options suited the relation.

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 (for example, not insert the connective *because* because of a full stop after the first segment). The second segment always started with a lowercase letter.

4. Results

Prior to analysis, the data of 4 participants were removed because these participants had very short completion times ($<$10 minutes for 20 passages of 5 sentences each) and showed high disagreement on causal and concessive items with other participants. The following analyses do not take the responses of these participants into consideration, leaving us with a total of 2962 observations. In

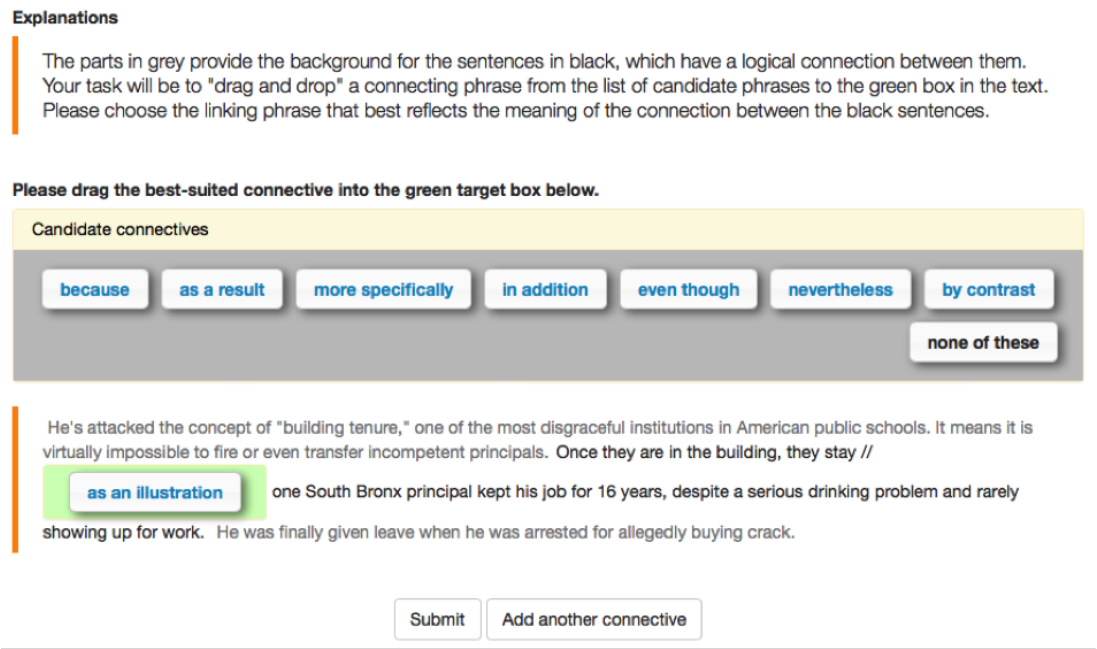

Figure 3: Example of the interface of the experiment – in this case, the participant chose *as an illustration* to indicate the meaning between the segments.

total, each list was completed by 12 to 14 participants. 25 participants completed between two and six batches; the others (86 participants) completed only one batch.

For the following analyses, we aggregated frequencies of the connectives that fell into the same class. In other words, *because* and *as a result* were aggregated as causal connectives, and *even though* and *nevertheless* were aggregated as concessive connectives. In 3.7% of the instances, participants made use of the 'manual answer' option, to insert connectives that were not on the provided list or to avoid inserting a connective (by leaving a blank or inserting punctuation). We discuss these manual insertions separately in Section 4.4, and aggregate the class 'manual answer' and 'no answer' for our analyses in Sections 4.1 to 4.3.2.

As with any discourse annotation task, some variation in the distribution of insertions can be expected. We are therefore interested in larger patterns in the distribution of inserted connectives. Crucially, we do not assume that there is one single correct label for each of our experimental items (cf. Rohde et al., 2016). In cases where we for example observe a distribution across two connectives in the insertions, we take this to indicate that this particular class or relation might have multiple senses. In order to determine the reliability of our method, we calculate agreement between connective insertions from our experiment and a replication experiment, and also evaluate the agreement with the originally annotated discourse treebank labels for these items. We report percentages of agreement and Krippendorf's Alpha[7] (α, Krippendorff, 1980). For the filler items, alpha was calculated by comparing the PDTB label to the dominant response of the crowdsourced participants.

7. Alpha was calculated using the R package agree.coeff2.r.

First, we discuss the results of the experiment overall, to determine whether the method was effective (Section 4.1). Next, we turn to a more comprehensive analysis of PDTB INSTANTIATIONS (Section 4.2). We look at the insertions per RST class (Section 4.2.1) to investigate whether the subjects interpreted the items in line with PDTB's or RST's classification. Then we look at a few examples of clear elaborative and clear argumentative relations in more detail to be able to identify cues for these relation types (Section 4.2.2). The same is discussed for SPECIFICATION items (Sections 4.3.1 and 4.3.2). Finally, we discuss observations regarding manual answers and double insertions (Section 4.4).

4.1 Reliability of the crowd-sourced annotation method

The aim of our first analysis is to check whether the method of crowdsourcing the discourse relation labels is a valid method. To this end, our dataset includes not only SPECIFICATION and INSTANTIATION relations which are the focus of the study, but also conjunction, causal, contrastive and concessive relations. We found that our method is successful overall: The connectives inserted by the participants are consistent with the original PDTB annotation (α = .57). This is shown in Figure 4, with the bars reflecting the inserted connective per PDTB class.

78% of the inserted connectives in items with a causal PDTB label were causal connectives (*because* and *as a result*), and 67% of the inserted connectives in concessive items were the concessive connectives *even though* and *nevertheless*. For both classes, the second most frequent category of inserted connectives was their negative/positive counterpart: for CAUSE, the second most frequent category was CONCESSION (10%), and for CONCESSION, the second most frequent counterpart was CAUSE (15%). On closer inspection of the items, we find that the disagreement between crowdsourced annotations and original PDTB 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 agreement on relations with conjunction and contrastive PDTB labels is lower than agreement on causal and concessive relations, but the distribution of inserted connectives for CONJUNCTION and CONTRAST looks similar: the expected marker is used most often (40% and 44%, respectively), with the corresponding causal counterpart as the second most frequent inserted connective type (27% causal insertions and 32% concessive insertions, respectively). A closer look at the crowdsourced annotations for items in these classes reveals that this is due to genuine ambiguity

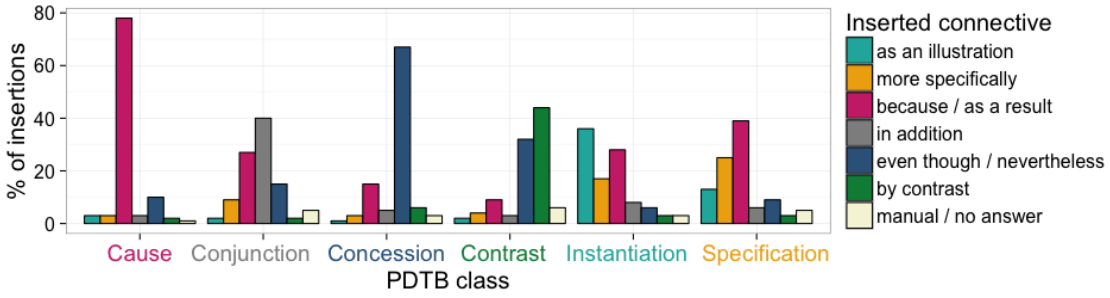

Figure 4: Distribution (%) of inserted connectives per PDTB class.

of the relation. For relations originally annotated as conjunction, we find that oftentimes a causal relation can also be inferred. The same explanation holds for contrastive 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 concessive and contrastive relations are relatively difficult to distinguish even for trained annotators (see, for example, Robaldo & Miltsakaki, 2014; Zufferey & Degand, 2013).

Finally, looking at the INSTANTIATION and SPECIFICATION relations that we set out to investigate in more detail in our study, we can see that there is more variety in terms of which connective participants inserted. This was expected, and the connectives that were most often inserted (*as an illustration, more specifically* and *because / as a result*) are consistent with our hypotheses. In order to get a clearer picture of the elaborative and argumentative types of these relations, we will turn to a more fine-grained analysis of these relations per RST class in the following sections.

The reliability of the current method was furthermore confirmed by a follow-up replication study: We repeated the experiment by presenting a new group of participants the same items without the context sentences to investigate the influence of context on interpretations (Scholman & Demberg, 2017). The connective insertions were almost a perfect replication of the results reported in this study, with the distribution being stable on an item-by-item basis. The agreement between the participants in the current study and the participants in the replication study is high: $\alpha = .71$. On average, the difference between the experiments on agreement with the PDTB label differed only by 3.7%. We ran Fisher exact tests on the insertions for each of the PDTB classes for the experiment reported here vs. the replication without context, and found no significant difference in the distribution of responses between studies for any of the PDTB classes (CAUSE: p=0.61; CONJUNCTION: p=0.62; CONCESSION: p=0.98; CONTRAST: p=0.88; INSTANTIATION: p=0.93; SPECIFICATION: p=0.85). The presence of context therefore did not significantly influence the results, see Scholman & Demberg (2017) for a more detailed discussion on this. Importantly, the results show that the distribution of insertions for every item is stable when two different crowdsourced groups take part in the experiment. Twelve insertions per item therefore seems to be an adequate amount to get a representative distribution of senses.

4.2 Analysis of INSTANTIATION relations

In this section, we look at the insertions into INSTANTIATION relations, first by RST label (Section 4.2.1) and then by item (Section 4.2.2).

4.2.1 ANALYSIS OF INSTANTIATION RELATIONS BY RST LABEL

We now turn to an analysis of only those relations that belong to the INSTANTIATION class. As shown in Figure 4, items in this category often received causal, instantiation and specification markers in our study. To investigate whether this variance of interpretation within the PDTB INSTANTIATION class is reflected in the RST annotation, we analyse these INSTANTIATION relations separately by their RST label. Hence, we repeat the analysis that is shown in Figure 4, but we include only INSTANTIATION relations and separate them by the four RST classes. This allows us to see whether the insertions that we obtained in our experiment agrees with the reading that is annotated by RST annotators (elaborative or argumentative). Figure 5 shows the distribution of inserted connectives in INSTANTIATION relations per RST class.

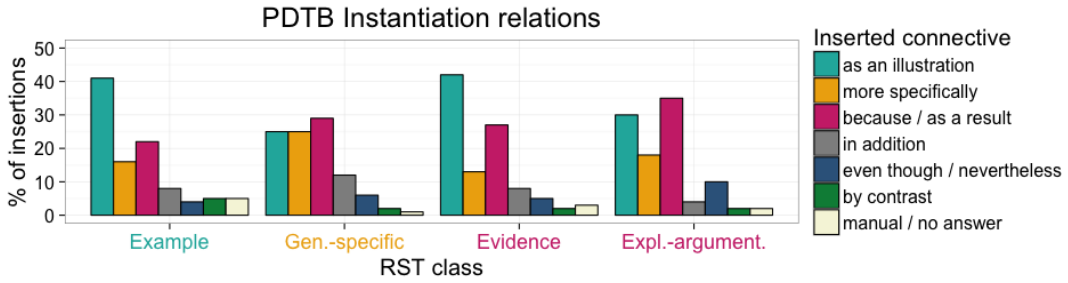

Figure 5: Distribution (%) of inserted connectives in INSTANTIATION relations per RST class.

The first RST class under investigation, EXAMPLE, is considered to be similar to PDTB's INSTANTIATION. As these two labels are direct correspondences to one another, we predicted that the participants in our study would also agree with this label and hence most frequently choose the connective *as an illustration*. As Figure 5 shows, the instantiation marker *as an illustration* was indeed the most frequent connective chosen (41% of all insertions), but the distribution is quite broad: there were also many causal connectives (22% of inserted connectives) and many SPECIFICATION insertions (16% of insertions).

The second class, GENERAL-SPECIFIC, is considered to be inconsistent with PDTB's class INSTANTIATION, since PDTB's INSTANTIATION corresponds to RST's EXAMPLE in RST, and PDTB's SPECIFICATION corresponds to GENERAL-SPECIFIC (also see Section 2). We predicted that these relations might be genuinely ambiguous between an INSTANTIATION or a SPECIFICATION reading, and that we would see both INSTANTIATION and SPECIFICATION markers inserted. The results confirm this prediction: items in the GENERAL-SPECIFIC class received an equal amount of INSTANTIATION and SPECIFICATION insertions (both 25%). However, the most frequently inserted type of connectives is causal, taking up 29% of all inserted connectives in this class, and we also see more conjunction relations than for other subgroups of INSTANTIATION relations. This group of relations therefore seems to be quite ambiguous.

The third RST class under investigation, EVIDENCE, is generally considered to be a causal class. We therefore predicted that these relations may have two functions: being an example that also serves as evidence for a claim. We expected that both INSTANTIATION and causal markers would both be inserted often. Figure 5 confirms this prediction: the INSTANTIATION marker is inserted most often (42% of all insertions), with causal connectives as the second most frequently inserted type (27%).

Finally, for the items annotated as PDTB INSTANTIATION and RST EXPLANATION-ARGUMENTATIVE, it was also expected that both INSTANTIATION and CAUSE markers would be inserted. This was indeed the case: 30% of inserted markers were *as an illustration*, and 35% were *because / as a result*. Furthermore, we see a higher number of connectives expressing concession relations for this subgroup, which may reflect the causal aspect of these relations. Again, the SPECIFICATION marker was also inserted relatively often, accounting for 18% of all insertions.

In sum, we find that all of the subclasses had a substantial amount of INSTANTIATION, SPECIFICATION and CAUSE interpretations. The results from our study on PDTB INSTANTIATION relations could be interpreted as evidence that INSTANTIATION items have both an elaborative *and* an argumentative function. However, it is also possible that rather than items being complex or ambiguous, subjects interpret a proportion of the INSTANTIATION items as expressing an elaborative relation,

and another proportion presenting an argumentative relation. The next section provides more insight into this issue.

4.2.2 BY-ITEM INSERTIONS IN INSTANTIATION ITEMS

For the analyses in the previous section, all INSTANTIATION items were grouped together and the percentages represented an average amount of insertions (over all items) per RST class. This grouping obscures any possible differences between items of the same RST class. In this section, we look at the distribution of insertions per item. The distribution per item can be used to observe genuine ambiguity in the interpretation of some items, but also to derive the dominant interpretation for each item.

Figure 6 provides a detailed picture by showing the percentage of inserted connectives per RST class and item. Every stacked bar on the x-axis represents an item, and the colours on the bars represent the inserted connectives. One way to analyse this data is to assign to each relation the label corresponding to the connective that was inserted most frequently by our participants, referred to as the *dominant response* (in Figure 6, this corresponds to the largest bar per item). After assigning relations the label corresponding to the dominant response, we can calculate how many items received a dominant response that is the same as the PDTB or RST label. In other words, we can calculate agreement between the dominant response per item and the PDTB label and RST label. This result is reported in Table 2.

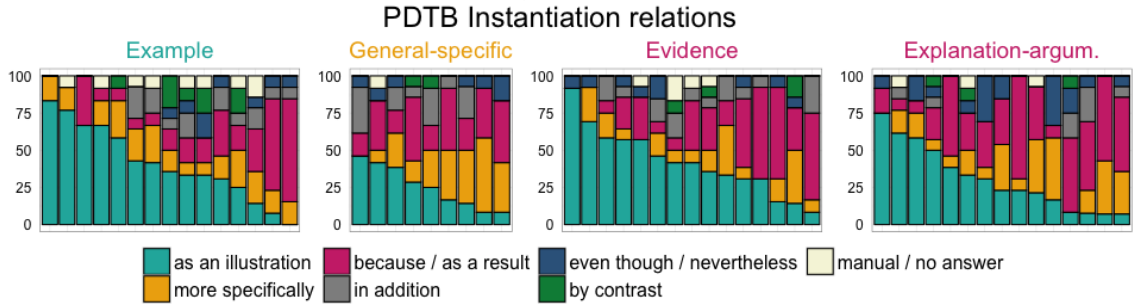

Figure 6: Distribution (%) of inserted connectives in INSTANTIATION relations per RST class and item. Plots are arranged according to the amount of INSTANTIATION insertions.

Relation type	**Agr. with PDTB**	**Agr. with RST**
INSTANTIATION – EXAMPLE	73	73
INSTANTIATION – GENERAL-SPECIFIC	33	22
INSTANTIATION – EVIDENCE	67	27
INSTANTIATION – EXPLANATION-ARGUM.	33	60

Table 2: Percentage agreement between the dominant response and the PDTB label INSTANTIATION, per RST class.

Table 2 shows that the dominant response converges with the PDTB label relatively often for items in the RST classes EXAMPLE and EVIDENCE. For items in the class INSTANTIATION – GENERAL-SPECIFIC, the dominant response does not converge with the PDTB label more, or less, often than with the RST label. This supports the hypothesis that these relations are ambiguous. The other common dominant response for these INSTANTIATION – GENERAL-SPECIFIC items is causal. For items in the class INSTANTIATION – EXPLANATION-ARGUMENTATIVE, the dominant response is most often causal, thereby converging with the RST label.

The visualization of the by-item analysis in Figure 6 reveals some interesting trends that we will discuss in more detail. Items that elicited many INSTANTIATION insertions mainly belong to the RST classes EXAMPLE and EVIDENCE, with a few cases also occurring in the class EXPLANATION-ARGUMENTATIVE. These items are considered clear examples of the class INSTANTIATION, and we expect there to be a cue present that indicates to readers that the item is an INSTANTIATION relation. A closer look at these items revealed a common characteristic: Often, a larger set is mentioned in the first argument, and one member of the set is explicitly referred to in the second argument. This larger set is referred to by a quantifier such as 'many', by plural noun phrases such as 'glossy brochures' and 'larger department stores', or by a combination of a quantifier and a plural noun phrase. This is illustrated in example (8), which is taken from the INSTANTIATION – EXAMPLE class and is presented in the same way as it was presented to participants. In Arg1 of Example (8), the set 'glossy brochures' is mentioned. Arg2 then refers to one member of the set ('one handout') and gives a more specific example of the phenomenon described in Arg1.

(8) But that's for the best horses, with most selling for much less – as little as $100 for some pedestrian thoroughbreds. Even while they move outside their traditional tony circle, race-horse owners still try to capitalize on the elan of the sport.
Glossy brochures circulated at racetracks gush about the limelight of the winner's circle and high-society schmoozing // **one handout promises: Pedigrees, parties, post times, parimutuels and pageantry.**
"It's just a matter of marketing and promoting ourselves," says Headley Bell, a fifth-generation horse breeder from Lexington. wsj_1174

The items that elicited mainly causal insertions occurred predominantly in the RST classes EVIDENCE and EXPLANATION-ARGUMENTATIVE (with two items in the class EXAMPLE). A common trait of these causal INSTANTIATIONS is that the first segment consists of a subjective utterance that can be interpreted as a claim and the second segment contains an argument for this claim, as in Example (9), taken from the group INSTANTIATION – EVIDENCE: The speaker makes a claim in the first segment, and provides evidence for this claim in the second segment, as well as in the context following the second segment. The majority of the subjects interpreted this relation as causal (62%).

(9) That done, Ms. Volokh spoke with rampant eloquence about the many attributes she feels she was born with: an understanding of food, business, Russian culture, human nature, and parties. "Parties are rather a state of mind," she said, pausing only to taste and pass judgment on the Georgian shashlik ("a little well done, but very good").
"If you are born to give parties, you give parties // **even in Russia we managed to give parties.**
In Los Angeles, in our lean years, we gave parties." wsj_1367

Another characteristic of relations that elicited causal insertions is that the second segment can be interpreted as a result of the situation described in Arg1, as in Example (10), taken from the group INSTANTIATION – EXAMPLE. In this example, the INSTANTIATION reading can be inferred when the reader interprets the second segment as an example of how international competition is heating up. However, when the reader interprets these segments as occurring chronologically, he will get the reading that the situation in the second segment happens as a result of the situation in the first segment. Indeed, 54% of insertions in this item were *as a result* (and 15% of insertions were because).

(10) The goal of most U.S. firms – joint ventures – remains elusive. Because the Soviet ruble isn't convertible into dollars, marks and other Western currencies, companies that hope to set up production facilities here must either export some of the goods to earn hard currency or find Soviet goods they can take in a counter-trade transaction.
International competition for the few Soviet goods that can be sold on world markets is heating up, however // **West German companies already have snapped up much of the production of these items.**
Seeking to overcome the currency problems, Mr. Giffen's American Trade Consortium, which comprises Chevron Corp., RJR, Johnson & Johnson, Eastman Kodak Co., and Archer-Daniels-Midland Co., has concocted an elaborate scheme to share out dollar earnings, largely from the revenues of a planned Chevron oil project. wsj_1368

Finally, certain items received many different types of insertions without showing a clear dominant response. Manual inspection revealed that these items often revolve around topics of economics that typically require background knowledge about the stock markets. The lack of agreement in annotation of these relations may hence be due to participants not having enough background information to judge the relations in the text. Given that even professionally trained discourse relation annotators are often not experts on the topic of the text that is being annotated, it is possible that this domain problem also affects the original PDTB and RST-DT annotations (also see Martins, Kigiel & Jhean-Larose, 2006; McNamara, Kintsch, Songer & Kintsch, 1996).

4.3 Analysis of SPECIFICATION relations

In this section, we look at the insertions into SPECIFICATION relations, first by RST label (Section 4.3.1) and then by item (Section 4.3.2).

4.3.1 ANALYSIS OF SPECIFICATION RELATIONS BY RST LABEL

For items from the PDTB class SPECIFICATION, the most frequently inserted connective type was not the marker that would be expected based on the PDTB label, but a causal marker (39%). The connective *more specifically* was the second most frequent type (25%) and *as an illustration* was the third most frequent type (13%). We again split up the dataset by RST labels for a more detailed analysis, see Figure 7.

The same predictions that held for INSTANTIATION items per RST class also hold for SPECIFICATION items: For items annotated as PDTB SPECIFICATION and RST EXAMPLE, we predicted that this disagreement between PDTB and RST annotations would be reflected in a similar split of inserted connectives by our participants. As Figure 7 shows, this is indeed the case: Subjects inserted the INSTANTIATION marker in 24% of the cases, and the SPECIFICATION marker in 22%

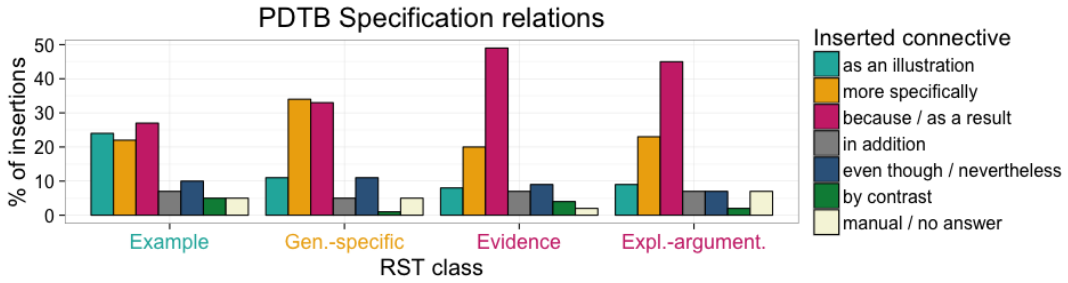

Figure 7: Distribution (%) of inserted connectives in SPECIFICATION relations per RST class.

of the cases. Somewhat surprisingly, we find a large proportion of causal insertions (27%) for these instances. Similar to the findings in Section 4.2.1, we find that relations for which PDTB and RST annotators did not agree on the INSTANTIATION or SPECIFICATION sense are ambiguous.

Items annotated as PDTB SPECIFICATION and RST GENERAL-SPECIFIC received a nearly equal amount of SPECIFICATION and causal insertions (34% and 33%, respectively). This brings up the question whether these instances are in fact both elaborative and argumentative. This will be discussed in the by-item analysis in Section 4.3.2 below.

For items that received the PDTB SPECIFICATION and RST EVIDENCE label, we predicted that participants would insert a high amount of SPECIFICATION and CAUSE markers. As Figure 7 shows, nearly half of all insertions were causal (49%), while only 20% of insertions were *more specifically*. Hence, participants seem to pick up on the same reading as RST annotators for these items. A similar pattern occurs for SPECIFICATION items that received the RST label EXPLANATION-ARGUMENTATIVE : 45% of the insertions were causal, and 23% were *more specifically*.

These results indicate that naïve subjects tend to interpret SPECIFICATION items as expressing a causal relation. The by-item analysis in the next section will show that items often received both types of insertions, and not only one or the other type.

4.3.2 BY-ITEM INSERTIONS IN SPECIFICATION ITEMS

Figure 8 displays the distribution of inserted connectives in PDTB SPECIFICATION relations per RST class. From these distributions of answers per items, we again calculate dominant responses

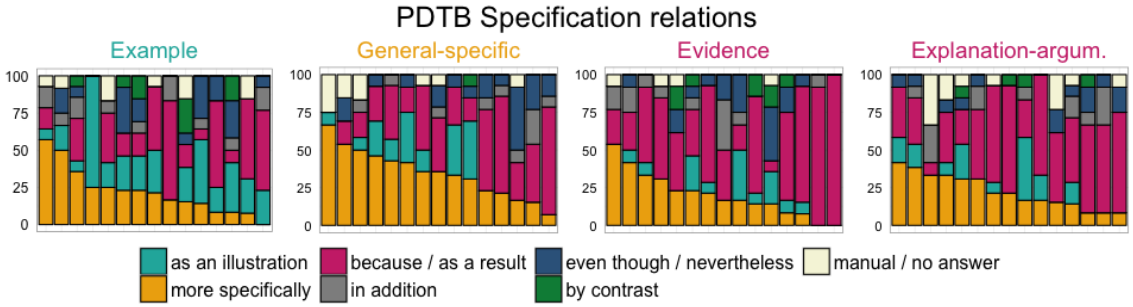

Figure 8: Distribution (%) of inserted connectives in SPECIFICATION items per RST class and item. Plots are arranged according to the amount of SPECIFICATION insertions.

Relation type	Agr. with PDTB	Agr. with RST
SPECIFICATION – EXAMPLE	20	27
SPECIFICATION – GENERAL-SPECIFIC	40	40
SPECIFICATION – EVIDENCE	13	73
SPECIFICATION – EXPLANATION-ARGUM.	27	67

Table 3: Percentage agreement between the dominant response and the PDTB label SPECIFICATION, per RST class.

and their agreement with the original PDTB and RST labels. The results of this analysis are shown in Table 3. The dominant response only rarely converges with the PDTB label. Most agreement is observed for items on which the PDTB and RST annotations agree (RST GENERAL-SPECIFIC).

A common characteristic of items receiving a high number of SPECIFICATION insertions is that the first segment contains a reference to a general or vague concept, such as 'one thing' in Example (11).

(11) The LDP won by a landslide in the last election, in July 1986. But less than two years later, the LDP started to crumble, and dissent rose to unprecedented heights.
The symptoms all point to one thing // **Japan does not have a modern government.**
Its government still wants to sit in the driver's seat, set the speed, step on the gas, apply the brakes and steer, with 120 million people in the back seat.

For items in the class SPECIFICATION – EXAMPLE, we find that the dominant response does not converge with either of the labels very often. For items in the class SPECIFICATION – EVIDENCE and SPECIFICATION – EXPLANATION-ARGUMENTATIVE, the dominant response is often causal. It thereby converges with the RST label (also shown in Table 3). A manual analysis of these items showed a similar finding to that discussed in Section 4.2.2: Often, the first segment of the relation contains a subjective claim. It is likely that readers interpreted the more specific information in the second segment as evidence for the claim.

4.4 Manual answers and double insertions

When participants did not think any of the provided connecting phrases suited the relation, they were allowed to provide a manual answer. 2.5% of all insertions in INSTANTIATION and SPECIFICATION items were manual answers (a raw count of 37 insertions). There was no clear pattern in these manual answers: only a few items received manual answers, and these items received at most two manual answers. The type of manual answer was also variable: A few of them were connectives (*although* and *however* were inserted once, *while* was inserted three times), some seemed to be related to the syntax of the items (for example, *as of*, *in which*, and *with*), while others aimed at attributing information between the two arguments to a speaker (for example, *saying*, *stating*, *adding*). No clear conclusions can be drawn from these insertions.

An additional 1.2% 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: Only a few items received a blank insertion, and there were no more than two blank insertions per item.

Participants were given the option of inserting two connecting phrases if they thought that both phrases reflected the meaning of the relation. A double insertion can give us more insight into whether subjects thought that two senses held for a relation. For INSTANTIATION and SPECIFICATION items, 4.1% of all consisted of two connecting phrases. For most items that received a double insertion, only one answer consisted of a double insertion. For a few items, two or at most three participants provided a double insertion. Looking at the amount of double insertions per participant, we find that only a few participants inserted multiple connectives (18 of 112 participants). The data on multiple insertions therefore does not allow us to draw any strong conclusions. This will be discussed further in the next section.

5. Discussion and conclusion

INSTANTIATION and SPECIFICATION are two of the most frequent implicit types of relations in the PDTB, making up 24% of all implicit relations in the PDTB. Considering that these labels occur so frequently, the current study was designed to investigate how readers interpret these relations. More specifically, we examined whether readers interpret them as elaborative, argumentative, or complex, and searched for characteristics that are shared by relations which are interpreted to be argumentative.

The results showed that both INSTANTIATION and SPECIFICATION items received many causal insertions: 28% of insertions in INSTANTIATION items and 39% of insertions in SPECIFICATION items were causal. We found that causal connectives were particularly prevalent in PDTB SPECIFICATION relations with RST EVIDENCE and EXPLANATION-ARGUMENTATIVE annotations. Importantly, a by-item analysis revealed that items rarely received only one type of insertion; rather, there were often two or more main types of insertions. These findings are consistent with a recent line of research that has focused on multiple readings of coherence relations: Rohde et al. (2015, 2016) have shown that certain relations can have more than one single reading. The current study has provided more evidence for this hypothesis, showing systematic ways in which different types of discourse relations can occur simultaneously. In future work, we plan to investigate whether originally explicit INSTANTIATION and SPECIFICATION relations (that is, those marked explicitly with connectives such as *for example*, *for instance* or *more specifically*) also have an additional argumentative reading that is not annotated (and is possibly even harder to detect for annotators, due to the presence of the explicit connective).

A manual analysis of items that elicited an INSTANTIATION or SPECIFICATION connective as the dominant response revealed that these items often contained one of the following characteristics: (i) a larger set is mentioned in the first segment, and one member of the set is explicitly referred to in the second argument, or (ii) the first segment contained a reference to a general or vague concept. By contrast, items that were often assigned a causal label shared one of the following characteristics: (i) the first segment contains a claim, and the second segment contains evidence or an argument for this claim, or (ii) the second segment can be interpreted as a result of the situation described in the first segment. These results go beyond previous work that has identified signals of INSTANTIATIONS and SPECIFICATIONS and other ELABORATION relations (e.g., Li & Nenkova, 2016; Taboada & Das, 2013; Vergez-Couret & Adam, 2012). For example, Taboada & Das (2013) have shown through manual annotation that relations from the RST class EXAMPLE can be signalled by individual words that indicate a relation without linking the two arguments (for example, the word *explaining*). GENERAL-SPECIFIC relations are most often marked by entity features and lexical

chains or overlap markers. Taboada & Das (2013) also find that EXPLANATION-ARGUMENTATIVE and EVIDENCE relations often remain unsignalled. Through computational corpus analysis, Li & Nenkova (2016) showed that first segments in INSTANTIATION relations are often shorter than other sentences, and the second segments are often longer. Moreover, first segments contain fewer out-of-vocabulary words than other sentences, and they contain more gradable adjectives (such as *high, likely*). The current study differs from these efforts in that the results are based on naïve readers' interpretations of relations, rather than expert judgments.

5.1 Implications for the annotation of INSTANTIATION and SPECIFICATION items

The results of the current study show that many discourse relations annotated in the PDTB as INSTANTIATION and SPECIFICATION also have an argumentative reading. This finding supports the hypothesis that INSTANTIATIONS and SPECIFICATIONS are sometimes used to illustrate / specify a situation and to serve as an argument to a claim. PDTB does not annotate this argumentative function, but rather focuses only on the ideational relation between the arguments; that is, on the elaborational reading of the relation. By contrast, RST does classify these relations into separate elaborative and argumentative classes, which better matches the dominant responses of participants in our study for some relations. However, neither framework fully captures the double reading of these items that was reflected in the results. In particular, RST-DT does not make provisions for annotating more than one reading of a discourse relation. PDTB annotators, while allowed to annotate more than one label per relation, hardly ever make use of this option for the INSTANTIATION and SPECIFICATION relations.

Classifying INSTANTIATIONS and SPECIFICATIONS as elaborative relations disregards the finding that many of these relations have an argumentative function as well. But classifying them as argumentative relations results in a disregard of their elaborative function (whether it's instantiating or specifying). In order to make the annotation of these relations more reliable and to ensure that annotations reflect actual interpretations by readers, we recommend that both the ideational function of a relation (for example, that one segment provides an example of what is said in the other segment) and its rhetorical function (for example, that the example is used to justify a claim) be annotated (also see Crible & Degand, in press; González, 2005; Redeker, 1990).

The way that the discourse-annotated corpora are currently structured, they do not contain multiple relation labels. But given that readers can obtain two different readings for the same relation, corpora would be descriptively more adequate if relations with multiple readings would receive double annotations. This could improve inter-annotator agreement as well: When using data annotated by only two coders, differences in the annotations might be interpreted as annotator error or disagreements. However, if at least one coder would annotate both senses, agreement would improve and the resulting annotations would better reflect the full meaning of the relation. Of course, a double annotation process raises issues as well: It would need to be clear whether a single coder sees both senses, or whether different annotators have different interpretations that are alternatives to one another but can't hold at the same time.

An alternative to annotating two separate labels for reflecting the ideational and the rhetorical function would be to create separate sub-classes for causal INSTANTIATIONS and causal SPECIFICATIONS. This solution would set these relations apart from purely ideational relations, and would as a result increase the label set. Adding more subtypes to the set of discourse relations instead of adding double annotations would be in line with the traditional assumptions that only one relation

holds between two discourse segments. Each of these solutions is likely to improve the descriptive adequacy of the labels for these relations, and thereby also the validity of the frameworks.

5.2 Methodological remarks on crowdsourcing discourse relation annotations

The crowdsourcing method used in the current study was shown to be relatively reliable for acquiring discourse annotations: The participants were able to insert the predicted connectives in filler items with high accuracy. Furthermore, we showed that replicating the study with a new set of participants lead to the same results, providing evidence that our type of crowdsourced annotations are reliable and reproducible.

Existing corpora have mostly been annotated by a small set of trained, expert annotators. Even after receiving a lot of training, agreement on the resulting annotations can be low (within and between frameworks). In Sections 1 and 2, we have shown that agreement on implicit relations in particular is very low between frameworks (roughly 35% agreement). This disagreement can partially be attributed to a difference in operationalization: The way that a discourse annotation task is designed and formalised naturally influences the resulting data. At a general level, the method we propose here is similar to PDTB's approach to discourse annotation: Participants are asked to insert a connective that signals the relation between two segments. Nevertheless, there are crucial differences between the two approaches, the most important difference being the use of naïve, untrained individuals in our study, the lack of an annotation stage that labels the relation, and the much larger number of judgments in our study. Additionally, our participants had the choice between only a small set of connectives, which are less ambiguous than many of the connectives that PDTB annotators could choose from. The participants also only had a few context sentences available to them (in the replication experiment they even had no context available), in contrast to PDTB annotators who can choose to read the entire text.

The most crucial difference between traditional annotation tasks and the task described in the current study is the resulting data. The method of inserting connectives instead of assigning discourse relation labels does lead to more coarse-grained annotations compared to annotations of trained, expert annotators. However, our annotations have the potential of better reflecting the average readers' interpretations, because they don't rely on rules and biases introduced by the annotation frameworks that are supposed to increase inter-annotator agreement. 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 allows researchers to obtain a distribution of relation senses. This distribution can give researchers more insight into the readings of ambiguous relations, and into how dominant each sense is for a specific relation. The method can therefore be used to investigate comparable issues with other relational classes as well.

We however also encountered limits in interpretability that are due to the experimental design. For instance, we can't decide based on our results whether relations that received different insertions were genuinely ambiguous to a single participant (i.e. both readings were possible and the participant decided for expressing only one of them with a connective) or whether different participants had different interpretations of the same relation (but did not think that the connective that another participant inserted was suitable). Even though our participants were provided with the option of inserting two connectives into a relation, they hardly ever made use of this option. It is possible that they avoided inserting a second connective because they only had one reading of the item; for ex-

ample, they either interpreted the relation as elaborative or argumentative, but not both at the same time. However, it is also possible that motivation played a role. 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. If the double sense of relations is the focus in a future experiment, this can be solved by asking subjects to explicitly indicate that they don't see a second reading. Hence, it is possible to make it obligatory to choose two connectives, with the option of "No other connective fits" as a second connective (a similar approach is taken in the construction of a new version of the PDTB (Webber et al., 2016)). Another option woud be to present the items with one of the connectives, and ask participants to indicate whether the connective accurately expresses the relation.

Finally, we also found that some instances received a lot of very different annotations, and that these instances could likely be due to participants' lack of domain knowledge in economics. We would like to draw attention to the fact that such a lack of domain knowledge might not only affect participants recruited via crowdsourcing platforms, but may also affect annotations of linguistically trained annotators who may not be very familiar with the textual domain. It is possible that, as a community, we are underestimating the effect of familiarity with a domain on discourse relation annotation quality and reliability.

Based on the high level of replicability between our original study and the replication study, we conclude that there is merit in the crowd-sourcing method, and believe it can potentially be used to create a corpus. Related approaches using crowdsourcing for discourse relation annotation have been put forward by Kawahara et al. (2014) and Rohde et al. (2016). These studies also advocate crowdsourcing as it is fast and cheap, the resulting data are reliable, and the method can provide valuable insights that traditional annotation tasks do not. What separates the current method from previous work is that it is designed to be comparable to PDTB's annotations, and the connectives were chosen to match PDTB's classes. Even though these connectives lead to more coarse-grained annotations, it is conceivable that the method can be extended to lead to more fine-grained distinctions in interpretations. In order to be able to apply this method to more general discourse annotation tasks, we recommend more research into which connectives can be added to represent more distinctions (e.g., TEMPORAL relations), as well as more research into the lower agreement for CONJUNCTION and CONTRAST relations. If these issues are dealt with, we believe that the task presented in this paper has the potential to function as a method to create a discourse annotated corpus that embraces multiple interpretations.

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). We are grateful to Jacqueline Evers-Vermeul and Jet Hoek for fruitful discussions.

Appendix A.

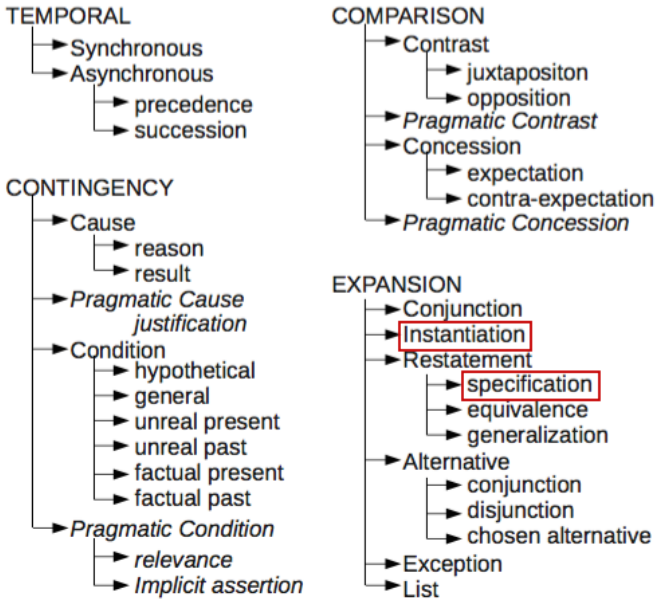

Figure 9: PDTB hierarchy (Prasad et al., 2008)

- ***Attribution:*** attribution, attribution-negative
- ***Background:*** background, circumstance
- ***Cause:*** cause, result, consequence
- ***Comparison:*** comparison, preference, analogy, proportion
- ***Condition:*** condition, hypothetical, contingency, otherwise
- ***Contrast:*** contrast, concession, antithesis
- ***Elaboration:*** elaboration-additional, elaboration-general-specific, elaboration-part-whole, elaboration-process-step, elaboration-object-attribute, elaboration-set-member, example, definition
- ***Enablement:*** purpose, enablement
- ***Evaluation:*** evaluation, interpretation, conclusion, comment
- ***Explanation:*** evidence, explanation-argumentative, reason
- ***Joint:*** list, disjunction
- ***Manner-Means:*** manner, means
- ***Topic-Comment:*** problem-solution, question-answer, statement-response, topic-comment, comment-topic, rhetorical-question
- ***Summary:*** summary, restatement
- ***Temporal:*** temporal-before, temporal-after, temporal-same-time, sequence, inverted-sequence
- ***Topic Change:*** topic-shift, topic-drift

Figure 10: RST-DT tagset (Carlson & Marcu, 2001)

References

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

Asr, F. T., & Demberg, V. (2013). On the information conveyed by discourse markers. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics* (pp. 84–93).

Biran, O., & Rambow, O. (2011). Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, *5*, 363–381.

Blakemore, D. (1997). Restatement and exemplification: A relevance theoretic reassessment of elaboration. *Pragmatics & Cognition*, *5*, 1–19.

Carlson, L., & Marcu, D. (2001). Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, *54*, 1–56.

Carlson, L., Marcu, D., & Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Current and new directions in discourse and dialogue* (pp. 85–112). Springer.

Carston, R. (1993). Conjunction, explanation and relevance. *Lingua*, *90*, 27–48.

Cohen, R. (1987). Analyzing the structure of argumentative discourse. *Computational Linguistics*, *13*, 11–24.

Crible, L., & Degand, L. (in press). Reliability vs. granularity in discourse annotation: What is the trade-off? *Corpus Linguistics and Linguistic Theory*, .

Cuenca, M.-J. (2003). Two ways to reformulate: a contrastive analysis of reformulation markers. *Journal of Pragmatics*, *35*, 1069–1093.

Cuenca, M.-J., & Marín, M.-J. (2009). Co-occurrence of discourse markers in Catalan and Spanish oral narrative. *Journal of Pragmatics*, *41*, 899–914.

Degand, L. (1998). On classifying connectives and coherence relations. In *Proceedings of the 1998 ACL Workshop on Discourse Relations and Discourse Markers* (pp. 29–35).

Demberg, V., Asr, F., & Scholman, M. (2017). How consistent are our discourse annotations? Insights from mapping RST-DT and PDTB annotations. *ArXiv e-prints*, *April*. arXiv:1704.08893.

González, M. (2005). Pragmatic markers and discourse coherence relations in English and Catalan oral narrative. *Discourse Studies*, *7*, 53–86.

Halliday, M. A. (1994). Functional grammar. *London: Edward Arnold*, .

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science*, *3*, 67–90.

Hovy, E. H. (1995). The multifunctionality of discourse markers. In *Proceedings of the Workshop on Discourse Markers* (pp. 1–11). Citeseer.

Hovy, E. H., & Maier, E. (1995). Parsimonious or profligate: How many and which discourse structure relations. *Unpublished manuscript*, .

Hyland, K. (2007). Applying a gloss: Exemplifying and reformulating in academic discourse. *Applied Linguistics*, *28*, 266–285.

Jasinskaja, K. (2013). Corrective elaboration. *Lingua*, *132*, 51–66.

Jasinskaja, K., & Karagjosova, E. (2011). Elaboration and explanation. *Constraints in Discourse*, *4*.

Kawahara, D., Machida, Y., Shibata, T., Kurohashi, S., Kobayashi, H., & Sassano, M. (2014). Rapid development of a corpus with discourse annotations using two-stage crowdsourcing. In *Proceedings of the International Conference on Computational Linguistics* (pp. 269–278).

Knott, A., & Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, *18*, 35–62.

Konig, E., & Siemund, P. (2000). Causal and concessive clauses: Formal and semantic relations. *Topics in English Linguistics*, *33*, 341–360.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Sage.

Li, J. J., & Nenkova, A. (2016). The Instantiation discourse relation: A corpus analysis of its properties and improved detection. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 1181–1186).

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, *8*, 243–281.

Martins, D., Kigiel, D., & Jhean-Larose, S. (2006). Influence of expertise, coherence, and causal connectives on comprehension and recall of an expository text. *Current Psychology Letters. Behaviour, Brain & Cognition*, *3*.

Maschler, Y., & Schiffrin, D. (2015). Discourse markers: Language, meaning, and context. In D. S. D. Tannen, H.E. Hamilton (Ed.), *The Handbook of Discourse Analysis, Second edition, Chichester, UK: John Wiley & Sons* (pp. 189–221).

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*, 1–43.

Moore, J. D., & Pollack, M. E. (1992). A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, *18*, 537–544.

Peldszus, A., & Stede, M. (2013). From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, *7*, 1–31.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. Citeseer.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., & Webber, B. (2007). *The Penn Discourse Treebank 2.0 annotation manual*.

Pusse, F., Sayeed, A., & Demberg, V. (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)*.

Redeker, G. (1990). Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics*, *14*, 367–381.

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

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

Rohde, H., Dickinson, A., Clark, C. N., Louis, A., & Webber, B. (2015). Recovering discourse relations: Varying influence of discourse adverbials. In *Proceedings of the Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)* (pp. 1–22).

Rohde, H., Dickinson, A., Schneider, N., Clark, C. N., Louis, A., & Webber, B. (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 (LAW X)* (pp. 49–58).

Sanders, T. J., Demberg, V., Hoek, J., Scholman, M. C., Torabi Asr, F., Zufferey, S., & Evers-Vermuel, J. (Submitted). Unifying dimensions in discourse relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, .

Sanders, T. J., Demberg, V., Hoek, J., Scholman, M. C., Zufferey, S., & Evers-Vermuel, J. (2016). How can we relate various annotation schemes? unifying dimensions in discourse relations. In *TextLink Second Action Conference* (pp. 110–112).

Sanders, T. J., Spooren, W. P., & Noordman, L. G. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, *15*, 1–35.

Scholman, M. C., & Demberg, V. (2017). Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task. In *Proceedings of the 11th Linguistic Annotation Workshop (LAW)* (pp. 24–33).

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

Snow, R., O'Connor, B., Jurafsky, D., & 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)* (pp. 254–263). Association for Computational Linguistics.

Stab, C., & Gurevych, I. (2014). Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 46–56).

Taboada, M., & Das, D. (2013). Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue & Discourse*, *4*, 249–281.

Vergez-Couret, M., & Adam, C. (2012). Signaling Elaboration: Combining french gerund clauses with lexical cohesion cues. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, .

Versley, Y. (2011). Multilabel tagging of discourse relations in ambiguous temporal connectives. In *International Conference on Recent Advances in Natural Language Processing (RANLP)* (pp. 154–161).

Webber, B. (2013). What excludes an alternative in coherence relations. In *Proceedings of the International Conference on Computational Semantics (IWCS)*.

Webber, B., Knott, A., & Joshi, A. (2001). Multiple discourse connectives in a lexicalized grammar for discourse. In *Computing Meaning* (pp. 229–245). Springer.

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2016). A discourse-annotated corpus of conjoined VPs. In *Proceedings of the 10th Linguistic Annotation Workshop (LAW)* (pp. 22–31).

Zufferey, S., & Degand, L. (2013). Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory*, *1*, 1–24.