Dialogue & Discourse 13(2) (2022) 49–78 doi: 10.5210/dad.2022.202

# The effect of domain knowledge on discourse relation inferences: Relation marking and interpretation strategies

**Marian Marchal** MARCHAL@COLI.UNI-SAARLAND.DE
*Language Science and Technology*
*Saarland University, Germany*

**Merel C.J. Scholman** M.C.J.SCHOLMAN@COLI.UNI-SAARLAND.DE
*Language Science and Technology*
*Saarland University, Germany*

**Vera Demberg** VERA@COLI.UNI-SAARLAND.DE
*Computer Science, Language Science and Technology*
*Saarland University, Germany*

**Editor:** Manfred Stede

Submitted 10/2021; Accepted 08/2022; Published online 09/2022

## Abstract

It is generally assumed that readers draw on their background knowledge to make inferences about information that is left implicit in the text. However, readers may differ in how much background knowledge they have, which may impact their text understanding. The present study investigates the role of domain knowledge in discourse relation interpretation, in order to examine how readers with high vs. low domain knowledge differ in their discourse relation inferences. We compare interpretations of experts from the field of economics and biomedical sciences in scientific biomedical texts as well as more easily accessible economic texts. The results show that high-knowledge readers from the biomedical domain are better at inferring the correct relation interpretation in biomedical texts compared to low-knowledge readers, but such an effect was not found for the economic domain. The results also suggest that, in the absence of domain knowledge, readers exploit linguistic signals other than connectives to infer the discourse relation, but domain knowledge is sometimes required to exploit these cues. The study provides insight into the impact of domain knowledge on discourse relation inferencing and how readers interpret discourse relations when they lack the required domain knowledge.

**Keywords:** Discourse relations; relational signals; discourse inferences; domain knowledge

## 1. Introduction

To successfully comprehend and learn from a text, readers need to construct a coherent mental representation of the information in the text. This requires readers to understand how the various concepts in a text are related and to integrate the text with background knowledge already available to the reader (see Van den Broek, 2010). Inferring discourse relations is an essential part of establishing coherence in text (Sanders et al., 1992). Prior studies have suggested that background knowledge supports the inference of discourse relations, assuming that this knowledge is activated to fill in information that is missing in the text (Noordman et al., 2015). The role of domain knowledge in interpreting discourse relations is still unclear. Earlier work has often focused on the role of background knowledge on text comprehension or recall (for an overview, see Smith et al., 2021), but how discourse relation inferences differ between high- and low-knowledge readers has not been investigated systematically.

Moreover, it is unclear what other factors guide the interpretation of discourse relations for low-knowledge compared to high-knowledge readers. Most studies in the field of discourse relations have focused on the effect of textual cues on relational inferences. Most notably, studies have shown the impact of connectives on the interpretation process. Agreement on explicit discourse relations is higher than on relations in which no connective is present (e.g., Demberg et al., 2019; Hoek et al., 2021; Kishimoto et al., 2018; Miltsakaki et al., 2004) and appropriate connectives facilitate the integration of upcoming material, whereas inappropriate connectives disrupt processing (e.g. Murray, 1997; Canestrelli et al., 2013). More recently, studies have focused on identifying other constructions and words that tend to correlate with certain discourse relations (Asr and Demberg, 2015; Webber, 2013; Grisot and Blochowiak, 2021). However, these cues are highly ambiguous and likely still need to be supplemented with non-textual information to infer the relation. How these cues interact with domain knowledge has not been taken into consideration. Furthermore, it is unclear how the type of inferences that readers make may depend on the knowledge that they have on the domain of the text.

The goal of the present study is therefore two-fold. First, the current paper aims to investigate whether domain knowledge leads to more correct interpretations of discourse relations. This will be assessed by eliciting discourse relation interpretations from high- and low-knowledge readers and comparing them to a gold label annotation. Second, this research sets out to explore how readers infer the discourse relation if they lack the necessary domain knowledge. In the next section, we will first review previous research on the role of domain knowledge in discourse inferences and discuss which factors influence discourse relation interpretation and could help low-knowledge readers to infer discourse relations for which domain knowledge is required. The hypotheses are outlined in Section 3, followed by a description of the methodology. The results are presented in Section 5. These are subsequently discussed in the final section.

## 2. Background

### 2.1 The role of domain knowledge in discourse inferences

Several models of language comprehension suggest that readers exploit their knowledge base about the concepts in the text to create a coherent representation of the text (e.g. *Construction-Integration model*, Kintsch and Van Dijk, 1978, *Landscape Model*; Van den Broek, 2010). This knowledge is activated when reading about relevant concepts in the texts, after which the information is retrieved from the long-term memory and can then be integrated with the representation that has been made of the text so far. In addition, reading about these concepts activates additional relevant information in the knowledge base, which can in turn influence predictions about subsequent text (cf. Venhuizen et al., 2019; Ferreira and Chantavarin, 2018). For example, comprehenders adopt general world knowledge in a similar way to linguistic cues to predict event structures within sentences (Milburn et al., 2016). With respect to inter-sentential discourse coherence, studies indicate that readers also create expectations based on linguistic cues (e.g. Köhne-Fuetterer et al., 2021; Scholman et al., 2017) as well as general world knowledge (Kuperberg et al., 2011). Reading comprehension is thus a dynamic process in which bottom-up and top-down processes are combined. If a discourse relation is not expressed linguistically, readers can utilize information from the knowledge base on how the events in the text are related to establish coherence.

There are different types of non-linguistic knowledge that readers can have. In the literature, a distinction is sometimes made between background knowledge (i.e. all the knowledge the reader can bring to the text), world knowledge, and domain knowledge (e.g., Smith et al., 2021). Domain knowledge is a type of background knowledge about a specific area (e.g. *apoptosis* is natural cell death). In this sense, it could be distinguished from general world knowledge (e.g. the sky is blue), which is considered to be available to almost every reader. It should be noted that we do not assume that if a reader has domain knowledge about the topic of the text, they will know how all concepts in the text are related, nor do we assume that this knowledge is required to infer every relation. Some concepts in the text might still be unknown to high-knowledge readers, and some textual relations can also be inferred without knowledge about the domain of the text. However, we do hypothesize that readers with domain knowledge might find it easier to infer the discourse relations in a text from their domain of expertise compared to readers without this specific knowledge, because they are more familiar with the concepts discussed in the text and can rely on an already existing knowledge structure.

Empirical evidence that readers benefit from domain knowledge in making discourse inferences comes from various studies on the influence of coherence marking on reading comprehension. This line of research has repeatedly shown a 'reverse cohesion' effect: in general, low-knowledge readers benefit from texts with high coherence marking, whereas high-knowledge readers show better comprehension after reading a low-cohesive text (McNamara et al., 1996; O'Reilly and McNamara, 2007; Kamalski et al., 2008; McNamara, 2001). Linguistic marking of coherence enables low-knowledge readers to understand how the concepts in a text are related. In the absence of such cues, comprehension will be impaired. For high-knowledge readers, on the other hand, a text with low cohesion induces them to employ their knowledge base to fill in the gaps in the text. Connecting the concepts from the text with those in their long-term memory then leads to deeper comprehension (McNamara et al., 1996). These studies have focused on the role of domain knowledge in text comprehension in general, but do not reveal how this influences the interpretation of discourse relations. Examining how low- and high-knowledge readers interpret discourse relations differently can provide more insights into the qualitative differences in text comprehension for these groups of readers. In addition, little is known about strategies that low-knowledge readers may have to comprehend an out-of-domain text. This will be addressed in the current study.

## 2.2 Strategies for inferring discourse relations

In addition to discourse connectives and background knowledge, several other factors have been suggested to influence discourse inferences. In cases where readers lack the domain knowledge to infer the discourse relation, and no connective is available to signal the relation, readers might resort to other strategies to establish coherence. More specifically, readers might (i) use non-connective linguistic signals for coherence relations, (ii) rely on cognitive biases for relational inferences, or (iii) process the text more shallowly. How these factors influence discourse relation inferences and how they might guide the interpretations of low-knowledge readers is outlined in more detail below.

### 2.2.1 RECOGNIZING DISCOURSE RELATIONAL CUES

Discourse connectives (e.g *because, however,* or *while*) are rather strong signals of discourse relations. Most relations, however, are not signaled explicitly by connectives (e.g., only 43% of the relations in the Penn Discourse Treebank are explicitly marked by a connective (Webber et al., 2016)). For relations that are not signaled by a connective, it is not necessarily the case that they contain no linguistic cue at all. Das and Taboada (2018) found that more than 80% of the relations in the RST Signalling Corpus were signaled by means of other cues. Non-connective linguistic cues for discourse relations (i.e. words or phrases that tend to co-occur statistically with certain discourse relations) might thus play a role in inferring discourse relations. Recent work has focused on identifying non-connective linguistic cues that co-occur frequently with certain discourse relations. For instance, a negation marker is present in more than half of the CHOSEN ALTERNATIVE relations, in which the second relational argument provides an alternative to the situation described in the first argument (see Example 1).

(1) Kevin is not going to Italy this year. Instead, he is going to visit family in New York.

Another example of a well-known linguistic cue is tense, which has been found to correlate strongly with different types of temporal relations (Grisot and Blochowiak, 2021).

Furthermore, corpus research has shown that the distribution of such cues seems to be different in explicit and implicit relations (cf. Sporleder and Lascarides, 2008). More specifically, non-connective linguistic signals for discourse relations appear to be more frequent in implicit relations than in explicit relations (Hoek et al., 2018). A case in point is the negation marker in Example 1 above: when such a cue is present, CHOSEN ALTERNATIVE relations are more likely to be implicit. In addition, Crible (2020) found that non-connective cues are more likely to occur in explicit relations if the connective is ambiguous. This is in line with Gricean's maxim of quantity to not make a message more informative than necessary (Grice, 1975) (see Asr and Demberg, 2012, 2015, for an explanation based on the Uniform Information Density Hypothesis).

Even though studies are increasingly showing the prevalence of non-connective linguistic cues, it remains unclear whether readers are sensitive to such cues. Given that the primary function of non-connective cues is to convey propositional content (rather than signaling a discourse relation, as connectives do), they could be considered to be much subtler cues. Furthermore, these elements are more ambiguous than discourse connectives in that they can correlate with a large variety of discourse relations. Thus, even though such patterns provide cues about the discourse relation, readers might not pick up on them. Still, there is some prior literature showing that readers are sensitive to cues other than connectives in processing discourse relations. For example, implicit causality verbs, like *admire* in (2) below, have been shown to elicit expectations for causal discourse relations (Kehler et al., 2008; Koornneef and Sanders, 2013).

(2) Laura admires Mo, because he won the competition.

Similarly, Crible (2021) found that the processing of CONCESSION relations is facilitated by the presence of overt negation in the first segment, whereas RESULT relations are read slower when the first segment contains negation. These findings suggest that comprehenders do draw on non-connective cues to infer discourse relations. In a study by Scholman et al. (2020), quantity expressions (e.g. *a few*, *several*) in the preceding context yielded more LIST continuations from participants in a study completion paradigm than when these signals were not present. Their research also showed that participants with much reading experience were more sensitive to these cues. This suggests that not all readers pick up on discourse relational cues equally well.

The present study aims to investigate how readers' sensitivity to discourse relational cues interacts with their domain knowledge. To manipulate the degree to which linguistic signals might be present in the text, both originally implicit and originally explicit relations will be used in the current study.[1] For the originally explicit relations, we remove the connective to create implicit versions. We call these instances in which the original connective has been removed *implicitated* relations (following e.g. Hoek et al., 2017, on implicitation in translation). Previous studies on discourse relation inferences have used either implicit (Scholman and Demberg, 2017a; Yung et al., 2019) or implicitated relations (Sanders et al., 1992), but have not compared readers' accuracy on the two types of relations. Since relations that are implicit likely contain more discourse relational cues than relations that have been implicitated, we expect readers to be better at inferring implicit relations than implicitated relations. We thus predict that agreement on implicitated relations will be lower than on originally implicit relations. Note that implicit relations might also be left unmarked because they are easy to infer based on general world knowledge and will therefore yield higher accuracy. Likewise, implicitated relations might be more difficult because the meaning changes when the connective is removed. We will return to this issue in the discussion.

Crucially, we also predict a possible interaction with domain knowledge here. If the text contains relational cues, both high- and low-knowledge readers might be able to employ these to infer the discourse relation. The effect of domain knowledge will then be moderated. However, if the amount of linguistic information for the discourse relation is limited, as in the case of implicitated relations, high-knowledge readers can still rely on their domain knowledge to infer the relation. Low-knowledge readers, on the other hand, do not have this information at their disposal and will then struggle with inferring the intended relation, leading to lower agreement. The effect of domain knowledge is therefore hypothesized to be larger for implicitated than for implicit relations.

### 2.2.2 COGNITIVE BIASES IN RELATION INFERENCES

Another way in which readers might infer discourse relations in the absence of other information is by relying on cognitive biases towards certain discourse relation inferences. According to the *continuity hypothesis* (Segal et al., 1991; Murray, 1997), readers prefer to interpret information in a text as being temporally and causally continuous. More specifically, it suggests that readers tend to relate sentences in an additive, temporal or causal way. Similarly, the *causality-by-default hypothesis* (Sanders, 2005) states that readers have a bias to infer causal relations between the segments in a text. Several corpus-based and experimental studies have provided evidence for these hypotheses. For example, continuous relations have been shown to be marked less frequently by a connective (Asr and Demberg, 2012), which is argued to be caused by the fact that these relations will be inferred regardless of coherence marking. In addition, causal relations are processed faster than additive relations (Sanders and Noordman, 2000), which in turn are processed faster than discontinuous discourse relations, such as contrastive relations (Murray, 1997).

1. Manipulating the materials by adding or removing these cues was not deemed suitable for the present study, given the relatively limited insights that are currently available regarding the variety of non-connective relational cues and their effects. Furthermore, readers with different levels of domain knowledge might make use of different types of signals, but we do not know beforehand what these signals might be. We therefore use natural text to be able to explore what such cues might be in a qualitative analysis of the results.

With respect to discourse inference strategies in cases where domain knowledge is required to interpret the relation correctly, it can then be hypothesized that readers will default to inferring a causal or another type of continuous discourse relation if no other relation can be inferred. Only in cases where readers have the necessary background knowledge to infer the relation, they might not rely on these cognitive biases and infer a less-expected discourse relation.

The hypothesis that readers infer expected relations is also supported by a shallow processing account for discourse relations in the absence of domain knowledge. According to Graesser et al. (1994), readers with less background knowledge process text less deeply and might even abandon their search for coherence. Several experimental studies have shown that readers are indeed less likely to make inferences during reading when they lack background knowledge (Noordman et al., 2015, 1992). This suggests that low-knowledge readers process discourse relations more shallowly. Scholman (2019) shows that shallow processing might lead to a higher susceptibility for cognitive biases in relation interpretation. In her study, readers interpret INSTANTIATION and SPECIFICATION relations less often as being argumentative when being forced to process the relation more deeply (i.e. by first summarizing the text). Thus, according to a shallow-processing account, low-knowledge readers who process the text more shallowly might therefore have a stronger preference for continuous and causal discourse relations.

### 2.2.3 UNDERSPECIFIED INTERPRETATIONS

Finally, low-knowledge readers could abstain from committing to a specific discourse relation, but rather make an approximate assumption about the meaning. For example, readers might infer that there is a negative, or adversative, relation between the segments in (3), but not determine whether it is a CONCESSION (i.e., one of the segments raises an expectation that is denied in the other segment) or a CONTRAST (i.e., the two segments present two different concepts). In the CONCESSION reading, Juan knew that his girlfriend would be satisfied with just a drink, but ordered much more despite that. In the CONTRAST interpretation, Juan's extensive order is compared to the small order of his girlfriend.

(3) Juan ordered everything on the menu. His girlfriend only wanted something to drink.

Such underspecified interpretations can arise from two causes. On the one hand, it might be a result of shallow processing, similar to a preference for cognitively expected relations. If readers process a text shallowly, they might be satisfied with only inferring that the relation is negative and not wish to specify it further, as this would require more effort. On the other hand, such underspecified interpretations might also arise from uncertainty about the discourse relation. Even if a low-knowledge reader processes the text deeply, they might still remain uncertain about the specific relation sense when they lack the required domain knowledge. For example, readers might not be able to determine whether the relation in (3) above, is a CONCESSION or a CONTRAST relation, despite wishing to do so. If they are nevertheless able to infer some features of the discourse relation (e.g. that it is a negative relation), they might still infer such an underspecified relation, rather than committing to a specific relation that could be wrong.

Participants can express uncertainty about the relation sense through their connectives. For example, the connective 'but' is underspecified regarding its relational sense; it can be used to express both CONTRAST and CONCESSION. A connective like *by contrast*, on the other hand, is more specific, as it can only be used in CONTRAST relations. Similarly, *nevertheless* specifies the relation for CONCESSION. Readers who retain underspecified interpretations might therefore prefer to provide ambiguous connectives, such as *but*. If readers make specific relation interpretations, they will insert more specific connectives, like *nevertheless*.

## 3. Present study

Background knowledge has been assumed to help readers to infer discourse relations, but it is still unclear how discourse relation inferences differ between high- and low-knowledge readers, both with respect to the quality of the inference as well as the cues that these different types of readers use. In this study, we manipulate domain knowledge by presenting experts from economics and biomedical sciences with texts that either stem from their domain of expertise (e.g. biomedical research papers in the case of biomedical experts) or from the other domain (e.g. biomedical research papers for economists). The biomedical texts included in this study stem from research papers, which were written for experts in the field. These texts are likely difficult to understand for readers without domain knowledge. The economic texts stem from newspaper articles, which were written for a broader audience. The effect of domain knowledge may therefore be less strong in this genre. Nevertheless, since the topic of the newspaper texts focuses on a specific domain, these texts may still be easier to understand for readers who are familiar with that domain than those who are not.

Discourse interpretations were elicited using a connective insertion task (Yung et al., 2019) and compared to gold label annotations. To examine the use of non-connective linguistic signals for discourse relations (see Section 2.2.1), the relations were either originally implicit or implicitated for the purposes of the current study.

The first research question that this study will address is:

- Do high-knowledge readers make more accurate discourse inferences than low-knowledge readers?

If high-knowledge readers employ their knowledge base when inferring how segments in a text are related (cf. Noordman et al., 2015), high-knowledge readers are expected to infer the relation correctly more often than low-knowledge readers.

Secondly, when required to make an inference about a discourse relation, low-knowledge readers might take several approaches to establish coherence in the text. The second aim of the study is therefore to investigate:

- What inferences do readers make if their domain knowledge is insufficient to infer the discourse relation?

Based on the discussion above, we can formulate three hypotheses about what readers will do in the absence of domain knowledge:

(a) Readers use non-connective linguistic signals to infer the discourse relation.

(b) Readers resort to default interpretation strategies based on cognitive biases for continuity and causality.

(c) Readers make less precise interpretations about the discourse relation.

Since implicit relations have been suggested to contain more non-connective linguistic signals than implicitated relations, we hypothesize that these relations will be easier to infer. Moreover, we predict that this effect is stronger for low-knowledge readers, as they are hypothesized to be unable to compensate for their lack of domain knowledge in implicitated relations. Note that it might also be the case that relations are left implicit for other reasons, for example because they are easier to infer on the basis of general world knowledge. We will therefore also examine qualitative differences in the presence of linguistic signals in items on which high- and low-knowledge readers differ.

In addition, if low-knowledge readers' interpretations are guided by their cognitive biases, it is predicted that these readers will infer more continuous and causal discourse relations than high-knowledge readers. If the third hypothesis is true, low-knowledge participants are predicted to insert more ambiguous connectives than high-knowledge readers, as they reflect their underspecified interpretation better than specific connectives.

These hypotheses are not mutually exclusive. Readers might attempt to use linguistic correlates for discourse relations to make the inference, but leave the relation underspecified if these cues are not sufficient to make a precise inference. Similarly, if these discourse relational signals are absent, readers may interpret the relation as being continuous, but not further commit their interpretation to a specific type of continuous relation.

## 4. Method

### 4.1 Participants

We recruited students and graduates in the field of economics and biomedical sciences on Prolific for a prescreening study. More specifically, five hundred workers participated that had registered on Prolific that their subject of study was in the field of economics or biomedical sciences[2]. The prescreening study served to further ensure expertise in one of the two domains and assess familiarity with each domain. Participants were presented with a short questionnaire assessing their demographic background and familiarity with both fields. The latter involved questions about their study and work experience in the field. In addition, we assessed participants' knowledge by asking them to indicate their familiarity with 10 concepts that are specific to the two domains extracted from the two corpora (e.g. *volatility*, *phosphorylation*). For each of these concepts, participants could indicate that they either did not know the term (which was coded as 1), had heard of the term while not being able to describe it (coded as 2), or would also be able to provide a description of the term (coded as 3). Five filler concepts for each domain consisting of terms that were deemed familiar for non-experts as well (e.g. *interest, DNA*) were included as an attention check.

In order to ensure that our study participants were knowledgeable in their own field of expertise, but not in the other field, we selected only those participants that met the following criteria for the final experiment: (a) they were working or studying in one (and only one) of the two fields, (b) they had high familiarity with the terms in their own field of expertise (top 40% compared to all participants), (c) they had low familiarity in the other field (bottom 50%), and (d) they did not consider themselves novices in the field (i.e. they did not rate their own familiarity with the field compared to other people working or studying in the field lower than 3 on a 7 point Likert scale).

2. In other words, when registering on Prolific, participants had indicated that their field of study was either in economics, accounting and/or finance or in biomedical sciences, genetics, biology, biological sciences and/or biochemistry.

Participants' average familiarity with the concepts in the domain they were an expert in was higher than that of the other domain (see Table 1). In addition, each individual participant had higher familiarity with the terms from their own domain than with the terms from the other domain. Note that the biomedical experts have higher familiarity with the terms from the economics domain than vice versa. We will return to this issue in the discussion. In short, the experts in our study had academic experience in the relevant subject (as shown by their registration on Prolific and their responses to our pretest) and considered themselves knowledgeable in the field (as indicated in our pretest) and their expertise was also reflected in their familiarity with the specialized language used in the texts.

Table 1: Mean scores on the concepts by domain and expertise. Scores on a scale from 1 (*I have not heard of the term*) to 3 (*I would be able to describe the term*).

| | biomedical terms | economic terms |
|---|---|---|
| biomedical experts | 2.86 | 2.09 |
| economics experts | 1.62 | 2.94 |

In the final experiment, 106 participants, all native speakers of English, took part (age range, 19-47 years; mean age, 24.7 years; 60 female). Of these, 89 participants were students; 56 had completed an undergraduate degree or had obtained a higher education level.

### 4.2 Materials

Ninety-six relations were sourced from the Penn Discourse Treebank 2.0 (PDTB; Prasad et al., 2008), a discourse-annotated corpus containing Wall Street Journal texts. Only those sections from the PDTB that were classified as news articles were selected. In addition, we only included texts that covered economic or financial topics. An additional set of 96 relations was extracted from the Biomedical Discourse Relation Bank (BioDRB; Prasad et al., 2011). This corpus contains discourse annotations of 24 biomedical research articles from the GENIA corpus, using an adapted version of the PDTB annotation framework. The latter texts are likely more specialized than the newspaper texts, which are written to be accessible to a broader audience. Still, financial newspapers, like biomedical research papers, target a specific group of readers (i.e. people working in the field of economics) and some degree of domain-specific knowledge is presumed by the writers of these texts as well. We will elaborate on this issue in the Discussion. Different items could come from (different parts of) the same texts, but the items in each corpus came from at least twenty different texts so that writing styles were varied.

The set of experimental items contained an equal amount of implicitated (i.e. originally explicit discourse relations from which the connective has been removed) and implicit relations. To balance the items with respect to the cognitive complexity and expectedness of the relation sense, four different relation senses were selected for the purposes of the present study: RESULT, CONTRAST, CONCESSION and INSTANTIATION. More specifically, we selected CONTRA-EXPECTATION as the subcategory of CONCESSION relations.[3] Each relation sense occurred equally often in the experiment.

Only items for which both arguments were single full sentences were included. The context, consisting of one or two full sentences before and after the arguments, was also presented. An example of an implicitated CAUSE item can be found in Passage 4. The relational arguments, separated by \\ are presented in boldface here. To the participants, the context was presented in grey and the target sentences in black.

(4) Convertible debentures – bonds that can later be converted into equity shares – are the most popular instrument this year, though many companies are also selling non-convertible bonds or equity shares. These mega-issues are being propelled by two factors, economic and political. **In the past, the socialist policies of the government strictly limited the size of new steel mills, petrochemical plants, car factories and other industrial concerns to conserve resources and restrict the profits businessmen could make \\ industry operated out of small, expensive, highly inefficient industrial units.** When Mr. Gandhi came to power, he ushered in new rules for business. He said industry should build plants on the same scale as those outside India and benefit from economies of scale.

For each of the four relation senses and each relation marking, 12 items were extracted from the PDTB and 12 items from the BioDRB, resulting in 192 items in total.

### 4.3 Procedure

The task was an updated version of the two-step connective insertion task developed by (Yung et al., 2019).[4] Participants were presented with each item one by one and were asked to complete two steps. In the first step, participants were asked to freely insert a connective in the blank that reflects the relation between the arguments best. They could only continue to the next step if they had typed something in the blank and were instructed to type the word *nothing* if they could not think of a linking phrase connecting the sentences.[5] They were then provided with a list of connectives in the second step and asked to select the connective that fits the relation best. The options presented in the second step were based on the insertion in the first step, and were unambiguous alternatives for the relations that can be signaled by the connective in the first step. For example, if *but* was inserted in the first step, the options in the second step consisted of (among other options) *despite this* and *on the contrary* to disambiguate between the CONCESSION and CONTRAST relation sense that can be marked by *but*. If the option inserted in the first step was not present in the connective bank, a default list was presented: *therefore, in addition, despite this, in more detail, even though, for example, by contrast, due to, this example illustrates that, in other words*. This default list thus contained a target connective for each of the target relations included in the item.

3. Within the class of contrast relations, the PDTB2 distinguishes between JUXTAPOSITION and OPPOSITION. Since no such distinction was made in the BioDRB, this distinction was disregarded when selecting materials for the present experiment.
4. Three adaptations were made to Yung et al. (2019)'s task. Firstly, participants were always presented with the second step in this experiment, regardless whether the connective they inserted in the first step was unambiguous or not, to discourage the use of only very specific connectives in the first step. Secondly, the connective bank and the mapping of ambiguous connectives to the options in the second step was updated based on follow-up experiments. Thirdly, the default list was adapted for the purposes of the current study.
5. Participants avoided this restriction in 1.2% of all data points by inserting punctuation or a whitespace.

The experiment was hosted on Lingoturk (Pusse et al., 2016) and distributed via Prolific. Participants first received instructions. They then saw two practice items, after which they received feedback on possible answers for these items. The items were divided across three batches per relation marking, with four items per relation sense per domain. Thus, each participant saw 32 experimental items. Four additional filler items were included as attention checks. These items were taken from the PDTB and did not require economic domain knowledge. Performance for these items was at ceiling in previous experiments. After completing the study, participants were asked to rate the difficulty of the texts on economic and biomedical topics. The study took around 30 minutes to complete and participants were given £3.50 as compensation for their participation.

### 4.4 Data analysis

Data from participants who provided less than five different types of connectives in the first step (n = 2) or selected that they wanted to insert a different connective in the second step in more than half of the cases (n = 4) were excluded from further analysis. In addition, participants who failed to select a connective that belonged to the same relational class as the gold label (see below) for more than half of the filler items were also removed (n = 6). The final dataset (n = 2,976) contained observations of 48 experts in the domain of biomedical sciences (implicit: 23, implicitated: 25) and 44 experts in the domain of economics (implicit: 24, implicitated: 22). Trials for which participants answered that none of the options provided in the second step were suitable, were coded as missing data (n = 179 observations, 6.0%).

To determine whether participants had inferred the relation correctly, the connectives in the second step were categorized as signaling eight different relational classes: (1) cause, (2) temporal, (3) contrast, (4) concession, (5) positive expansion (e.g. INSTANTIATION), (6) negative expansion (e.g. DISJUNCTION), (7) condition, (8) no relation. An overview of which PDTB3 relation senses are included in each relational class can be found in the Appendix.

We recoded a new variable, correctness, which was 1 when the inserted connective in the second step matched the relational class of the target relation sense (i.e. agreed with the gold standard), and 0 when it signaled a different relational sense. The correctness variable is used as the dependent variable in subsequent analyses, unless stated otherwise.

During data exploration, we discovered that performance on the CONTRAST relations was much lower in the PDTB than in the BioDRB (11.2% vs. 31.9%) as well as compared to other relation senses in the PDTB (60.0%). For these contrastive PDTB items, participants frequently provided a concessive connective. Note that the distinction between CONTRAST and CONCESSION is notoriously difficult (see e.g. Robaldo and Miltsakaki, 2014; Zufferey and Degand, 2017). In fact, the manual of the updated version of the PDTB2 (PDTB3, Webber et al., 2019) states that they addressed this issue in PDTB3 by reclassifying many CONTRAST relations as CONCESSION. We compared the labels of our CONTRAST items between PDTB2 and PDTB3 and found that 21 out of 24 CONTRAST relations were relabelled as CONCESSION. We therefore decided to use the updated PDTB 3.0 labels as the gold label.[6] We will come back to this issue in the discussion.

Binomial mixed-effects analyses were used to examine the data. Corpus, expertise and relation marking, but not relation sense, were deviation coded for ease of interpretation of the model with the PDTB corpus, economic experts and implicit relations at -1 and their counterparts at 1. For relation

6. The label for CONTRA-EXPECTATION and INSTANTIATION relations also differ between these two versions of the PDTB. These items were all labeled as ARG2-AS-DENIER and ARG2-AS-INSTANCE respectively in the PDTB 3.0.

Table 2: Confusion matrix of the gold relation senses and the inserted categories (% per relation sense).

| | cause | positive expansion | concession | contrast | other |
|---|---|---|---|---|---|
| Result | 64.7 | 17.5 | 8.7 | 2.7 | 6.3 |
| Instantiation | 21.9 | 56.2 | 9.9 | 5.7 | 6.4 |
| Concession | 19.1 | 15.0 | 48.8 | 8.1 | 9.0 |
| Contrast | 15.0 | 25.7 | 21.2 | 31.4 | 6.7 |

sense, treatment coding was used with CONCESSION as the intercept, as this was hypothesized to be one of the most difficult relations to infer. In addition, we were interested in its comparison with CONTRAST relations, due to these relations often being confused. Because of convergence issues, the BOBYQA optimizer was used with 10,000 iterations. The models were always first constructed with maximal random effect structure. In case of non-convergence, the model was reduced (Barr et al., 2013). The random slope for relation sense never converged. Unless specified otherwise, the models therefore contained random intercepts for participants and items and random slopes for corpus and expertise. All materials and data can be found online [7].

## 5. Results

### 5.1 Convergence with the gold label

On average, the correct relation sense was inferred in 52.1% of the insertions, as shown by convergence of the insertions in the second step with the gold label. Although this performance is relatively low, discourse relation classification is a notoriously difficult task and these numbers are comparable with similar studies using crowd-sourcing for discourse relation annotation (cf. Rohde et al., 2016; Kishimoto et al., 2018; Scholman et al., 2022). When the majority label per item is taken (i.e. aggregating responses of all participants per item to obtain a single annotated label), performance is much higher (74.2%).

As can be seen in Table 2, performance is much higher on some relation senses than on others (see Yung et al., 2019; Scholman and Demberg, 2017a, for similar results). Connective insertions for RESULT items were correct in 64.7% of cases, followed by the INSTANTIATION relations (56.2%). These two relational classes were often confused, suggesting that participants did not always know whether the relation was causal or not. Another possibility is that these relations were ambiguous for these two relation senses, since INSTANTIATION relations can often also be causal (Scholman and Demberg, 2017b). CONCESSION (48.8% correct) showed significantly lower accuracy than performance on RESULT relations as shown in a binomial mixed-effects analysis (see Table 3 below). The difference with INSTANTIATION relations was not significant. CONCESSION relations were sometimes confused with RESULT, but also with positive expansion relations. The latter is surprising, since that means that participants neither infer the causal nor the negative relation between the arguments. Finally, performance on CONTRAST relations was even lower than on CONCESSION relations, with only 31.4% of insertions falling in the same category of the gold label. In many of

7. https://osf.io/gq59w/?view_only=04d85b1d499b4379b3015693d71c6fcc

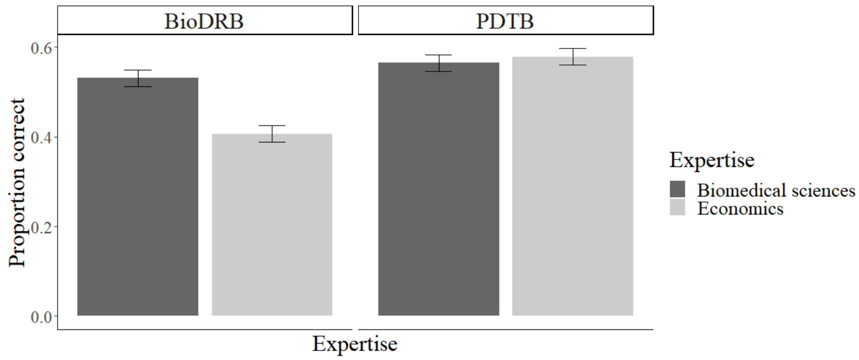


Figure 1: Convergence with gold label per domain and expertise with error bars showing the standard error.

these items, a connective that signals positive expansion or CONCESSION was inserted. CONTRAST relations have been shown to be difficult to annotate in other studies as well (cf. Kishimoto et al., 2018). In addition, CONTRAST and CONCESSION relations are known to often be confused with each other (e.g. Robaldo and Miltsakaki, 2014). Given the effect of relation sense on accuracy, relation sense was included as a covariate in all models presented in this paper.

### 5.2 The effect of domain knowledge on discourse relation inferences

As can be seen in Figure 1, performance was higher in the PDTB (57.1%) than in the BioDRB (47.1%). This effect was confirmed in the model, as shown by a significant main effect of corpus (see Table 3). In addition, overall, biologists converged with the gold label significantly more often than economists (54.8% vs. 49.4%). Indeed, expertise was also a significant predictor in the regression analysis.

The main question that this study aims to answer, however, is whether high-knowledge readers infer the correct relation sense more often than low-knowledge readers and how readers interpret discourse relations in the absence of domain knowledge. As can be seen in Figure 1, for each corpus, highest performance was obtained by the experts from that domain. The binomial mixed-effects analysis shows that the interaction between corpus and expertise is significant, suggesting that domain knowledge leads to a higher accuracy on relational inference. To examine this interaction more closely, we performed a subset analysis on the two corpora. Experts from the biomedical sciences converged with the gold label on items from the BioDRB (53.1%) more often than the economic experts (40.7%). In the BioDRB, expertise was indeed a significant predictor of correctness ($\beta$ = 0.30, SE= 0.09, $z$ = 3.47, $p < .001$). For the economic texts, however, this difference was minimal: biologists converged with the gold label in 56.4% and economists in 57.8% of cases. Expertise was not found to significantly predict accuracy in the PDTB texts. The full output of these models, as well as of all models below can be found in the Appendix.

Table 3: Output of the full model. Model specification: `correctness ~ relationsense + corpus*expertise*relationmarking + (1 + corpus | workerid) + (1 + expertise | questionid)`

| | **Estimate** | **Std. Error** | **Z value** | **P value** |
|---|---|---|---|---|
| Intercept | -0.11 | 0.14 | -0.81 | .42 |
| Relationsense RESULT | 0.84 | 0.20 | 4.20 | **<.001** |
| Relationsense CONTRAST | -0.75 | 0.26 | -2.92 | **<.01** |
| Relationsense INSTANTIATION | 0.39 | 0.20 | 1.94 | .05 |
| Corpus | -0.17 | 0.08 | -2.07 | **.04** |
| Expertise | 0.14 | 0.07 | 1.97 | **.05** |
| Relationmarking | 0.09 | 0.09 | 0.98 | .33 |
| Corpus:expertise | 0.17 | 0.05 | 3.55 | **<.001** |
| Corpus:relationmarking | 0.04 | 0.08 | 0.54 | .59 |
| Expertise:relationmarking | -0.01 | 0.07 | -0.18 | .86 |
| Corpus:expertise:rel...marking | -0.09 | 0.05 | -1.92 | .06 |

## 5.3 Interpretation strategies in the absence of domain knowledge

### 5.3.1 EXPLOITING RELATION MARKING

We hypothesized that readers use discourse relational cues in the text to infer the relation. More specifically, we assumed that implicit relations contain more of these cues than implicitated relations and are therefore easier to infer. In addition, low-knowledge readers were hypothesized to rely on these cues more than high-knowledge readers and therefore perform better on implicit than on implicitated relations. Table 4 shows the mean accuracy per corpus and relation marking by expertise. Overall, performance on implicit relations was slightly higher (54.0%) than on implicitated relations (50.5%). However, relation marking was not a significant predictor for convergence with the gold label (see Table 3). There was also no three-way interaction between corpus, expertise and relation marking. We thus find no evidence that the implicit relations are easier to infer than implicitated relations, nor that the effect of domain knowledge is different in implicit and implicitated relations.

To examine the role of discourse relational cues more closely, we performed a qualitative analysis on the 30 items for which the difference in accuracy between high-knowledge readers and low-knowledge readers was largest and examined the insertions by both groups. This allowed us to distinguish three different types of relations, which are presented below.

Table 4: Mean percentage of correct answers per corpus and relation marking by expertise.

| | **BioDRB** | | **PDTB** | |
|---|---|---|---|---|
| | implicitated | implicit | implicitated | implicit |
| biomedical sciences | 52.2 | 54.1 | 54.3 | 58.7 |
| economics | 36.2 | 45.4 | 58.7 | 57.0 |
| *mean* | *44.4* | *49.9* | *56.4* | *57.9* |

**Relations without linguistic cues require domain knowledge** For some items, the relation could only be inferred using domain knowledge. For instance, in Passage (5), a reader needs to know what 'Treg activities' are like in murine systems in order to know whether a reduction in human systems is similar or not. However, no linguistic cues are present to signal this relation. As a result, low-knowledge readers often interpreted this item as a CAUSE relation, instead of CONCESSION.

(5) In human infectious, neoplastic, and autoimmune diseases, Treg activities often mirror those in murine systems ___ numbers of Treg are reportedly reduced in human autoimmune diseases, (...)

(BioDRB:CONCESSION:Implicit)

**Relational cues allow relational inferences in the absence of domain knowledge** A number of items on which experts and non-experts diverged contained non-connective cues that could help readers to infer the correct relation, such as hyponyms for INSTANTIATION relations and antonyms for CONTRAST relations. More specifically, the majority of the fourteen INSTANTIATION and CONTRAST items that yielded a large difference between experts and non-experts contained such a cue. To illustrate these cues, consider (6) and (7), which yielded high accuracy from both high- and low-knowledge readers. The relational cues in these items are signalled linguistically by repeating words (e.g. *magazine*) or based on general world knowledge (left vs. right). This allows readers to infer relations even in the absence of domain knowledge.

(6) Other **magazine publishing companies** have been moving in the same direction ___ **the New York Times Co.'s Magazine Group** earlier this year began offering advertisers extensive merchandising services built around buying ad pages in its Golf Digest magazine.

(PDTB:INSTANTIATION:Implicit)

(7) The core biopsy of the **left breast** revealed infiltrating ductal carcinoma in 2 of 5 core fragments; high nuclear grade, with no lymphatic invasion seen ___ the core biopsy of the **right breast** demonstrated benign pathology, specifically, fibrosis with focal ductal epithelial hyperplasia.

(BioDRB:CONTRAST:Implicit)

**Relational cues sometimes require domain knowledge** In the items where there was a large difference between experts and non-experts, low-knowledge readers did not always pick up on these cues. The reason for this is that domain knowledge was often required to exploit the cue. This was especially the case for the INSTANTIATION relations. In about half of the cases in which a hyponym was present, this cue could only be exploited with domain knowledge. For example, in (8) below, the reader needs to know that *orthologous genes* are genes in different species that have a similar descent. The second argument provides a specific example of this, but if a reader does not have the required domain knowledge, they will likely also not understand that these genes are instances of orthologous genes.

(8) In particular, we assumed that the transcriptional regulation is conserved for **orthologous genes ___ the mouse gene Myh1 and the human gene MYH1** are assumed to share expression patterns and to share important cis-regulatory sequences.

(BioDRB:INSTANTIATION:Implicit)

Interestingly, the largest difference between experts and non-experts in convergence with the gold label in the full dataset can be found in implicitated INSTANTIATION relations in the BioDRB. Experts performed 30 percentage points higher than non-experts in this condition (see Table 6 in the Appendix). The implicit INSTANTIATION items in the BioDRB and implicitated INSTANTIATION items in the PDTB also yielded higher accuracy for experts than for novices. This suggests that cues for INSTANTIATION relations are more easily exploited by experts. In a post-hoc analysis, we therefore examined whether the effect of domain knowledge was different per relation sense. Adding the three-way interaction between relation sense, corpus and expertise did not significantly improve model fit when compared to the same model without this interaction. Since examining differences between the relation senses was not the purpose of the present study and power for finding such a three-way interaction effect with the current study design is likely to be low, further quantitative research is necessary to examine the effect of domain knowledge on different relation senses and different relational cues. The present qualitative analysis provides directions for future research.

Furthermore, it is interesting to point out that low-knowledge readers do not always exploit relational cues that do not require domain knowledge. More specifically, the three antonyms in the CONTRAST relations that were more challenging for low-knowledge readers could also be detected with general world knowledge, contrasting concepts that are accessible for low-knowledge readers as well (see (10) for an example). In addition, we found instances of hyponyms in our qualitative analysis that do not require specific domain knowledge to infer the INSTANTIATION relations, but were nevertheless not detected by low-knowledge readers, as in (9).

(9) More recently, **several groups** have demonstrated the feasibility of hybridizing metabolically labeled mRNAs directly from nuclear run-on (NRO) reactions to nylon filter microarrays in order to investigate nascent transcripts ___ **Schuhmacher et al.** used a B cell line carrying a conditional, tetracycline-regulated myc gene, and found that myc induction resulted in only a small overlap in regulated mRNAs at 4 hours post-induction when comparing polyA mRNA and NRO RNA on microarrays.

(BioDRB:INSTANTIATION:Implicitated)

(10) E2 **inhibits** apoptosis in different cell types (cardiac myocytes and others) ___ androgens have been found to **induce** apoptosis.

(BioDRB:CONTRAST:Implicitated)

To sum up, non-connective cues seem to play a role in discourse relation inferences, although we do not find evidence that the presence of these cues (or the extent to which they are used to infer the discourse relation) depends on whether or not the relation is marked. In addition, the qualitative analysis shows that adopting these cues sometimes requires domain knowledge. However, even if domain knowledge is not required, low-knowledge readers do not always adopt these cues.

#### 5.3.2 COGNITIVE BIAS FOR CONTINUITY AND CAUSALITY

A second hypothesis was that readers might be guided by cognitive biases for causality and continuity in case their background knowledge was insufficient to determine the relation sense. To examine whether low-knowledge readers resorted to default interpretation strategies, we coded the connective insertions for whether they were signals of continuous relations (*cause, positive expansion,*

Table 5: Mean effect (sd) of corpus and expertise on the insertion of a causal or continuous connective.

| | **Estimate** | **Std. Error** | **Z value** | ***P value*** |
|---|---|---|---|---|
| corpus | 0.28 (0.04) | 0.14 (0.01) | 2.02 (0.32) | .06 (.06) |
| expertise | 0.13 (0.04) | 0.10 (0.01) | 1.27 (0.41) | .24 (.16) |
| corpus:expertise | 0.01 (0.04) | 0.09 (0.00) | 0.06 (0.44) | .72 (.18) |

*temporal, condition*) or not (*contrast, concession, negative expansion)*. We only included incorrect insertions in this analysis, because correct continuous or discontinuous interpretations cannot be attributed with high certainty to a cognitive bias for continuity and causality, as they are likely guided by the true sense of the relation. In addition, we sampled an equal amount of insertions per relation sense from each corpus and domain of expertise. This was done to ensure an equal balance between relation senses, since certain relation senses might yield more default interpretations than other relation senses (e.g. incorrect interpretations of CONTRAST relations are more likely to be CONCESSION than RESULT). The sampling was repeated 100 times to examine whether the effects in the binomial mixed-effects analysis were stable. The results are presented in Table 5. No main effect of corpus, domain of expertise nor of the interaction between corpus and expertise was found. There is thus no evidence that readers have a bias to infer a causal or continuous relation if they lack the domain knowledge to interpret the relation correctly.

#### 5.3.3 UNDERSPECIFIED INTERPRETATIONS

If readers are not certain about the discourse relation between two arguments, they might resort to making an underspecified inference, rather than committing to a specific interpretation. In the paradigm used in this experiment, this would mean that participants insert more ambiguous connectives in the first step when they have little knowledge about the domain of the text, compared to when they are experts in that domain. The connective insertions in the first step were therefore annotated as indicating relations from one vs. multiple relational classes. The most frequent ambiguous first step insertions were *however* (11.8%), *and* (6.5%) and *but* (5.9%). In addition, participants typed *nothing* in 3.4% of cases, which indicated that they could not come up with a linking phrase connecting the sentences. *For example* (6.4%), *therefore* (5.1%) and *because* (3.3%) were the most frequent specific connectives.

A binomial mixed-effects logistic regression analysis showed an interaction between corpus and expertise ($\beta$ = -0.14, SE = 0.05, $z$ = -3.08, $p < .01$). This effect of domain knowledge on the insertion of ambiguous connectives is visualized in Figure 2. Splitting up the data by corpus revealed that, compared to experts from the domain of biomedical sciences, economic experts inserted significantly fewer specific connectives in the BioDRB ($\beta$ = -0.29, SE = 0.09, $z$ = -3.03, $p < .01$), but not in the PDTB. Besides inferring more incorrect relation types, low-knowledge readers thus also leave the relation underspecified by inserting ambiguous connectives in the first step, when reading the BioDRB.

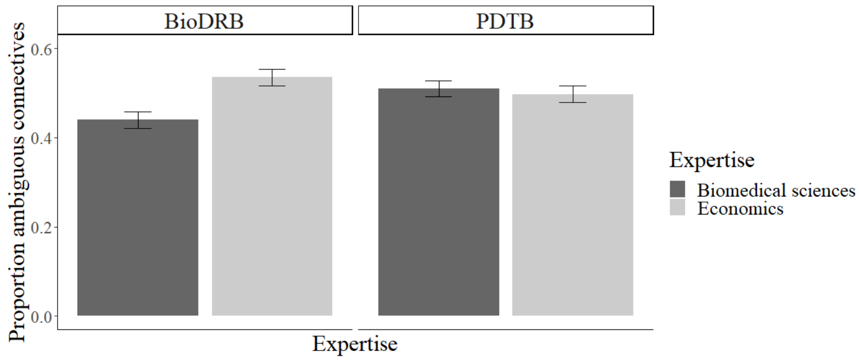


Figure 2: Proportion of ambiguous insertions in the first step with error bars showing standard error.

## 6. Discussion and conclusion

Background knowledge has often been assumed to play a role in correctly interpreting discourse relations, but this has never been investigated experimentally. The current study filled this gap by assessing discourse relation interpretations of high- and low-knowledge readers. We aimed to examine whether domain knowledge contributes to inferring the correct discourse relation, as well as which factors guide discourse relation interpretation in the absence of connectives and domain knowledge. The first main finding of this research is that high-knowledge readers were better at inferring the discourse relation, as measured by convergence with the gold label, than low-knowledge readers. Thus, domain knowledge can, in some instances, facilitate establishing coherence and readers are able to employ their knowledge base to interpret the relation correctly. However, this effect was modulated by the corpus from which the text was taken: The effect of expertise was significant for the items from the BioDRB, but not for the PDTB (see Section 6.1). In addition, we identified non-connective linguistic signals for discourse relations, showing that domain knowledge influences how readers adopt these cues (see Section 6.2).

### 6.1 Text genre and the influence of domain knowledge

One possible reason for why there was only an effect of domain knowledge for texts from the BioDRB and not the PDTB is the difference between these specific genres. Even though economics newspaper texts are targeted at readers with a specific interest in economics, they are intended for a broader audience with various levels of expertise. Research papers, on the other hand, are often not accessible to a general audience. Instead, they specifically target experts in that domain. They contain more specialized vocabulary and focus on topics that only a limited amount of people are familiar with.[8] Also note that the two texts differ in that they are written by journalists, who are not experts themselves, versus researchers. Discourse relations in the biomedical texts therefore likely required more domain knowledge than those in the economics newspaper texts.

8. This was also confirmed in a post-hoc analysis of the perplexity of the items using a generic language model (GPT (Radford et al., 2018), a transformer model that is trained on a 1B word book corpus). The perplexity of the PDTB items (89.0) is lower than of the BioDRB items (101.4).

Another explanation for this pattern could be related to the level of expertise of our participants. We recruited the participants via a crowd-sourcing platform, but their expertise was assessed in various ways (among others their subject of study as indicated on Prolific and their familiarity with specialized terms as determined during prescreening). This ensured that they were indeed high vs. low-knowledge readers with respect to the texts presented in this study. We note here again that experts were not expected to be familiar with all the information in the text. Domain knowledge was hypothesized to help in interpreting discourse relations correctly, because text processing is facilitated by an existing knowledge structure. This knowledge base does not need to be exhaustive, as the information in the text fills gaps in existing knowledge. Still, the experts from the domain of biomedical sciences seemed to know more about economics than vice versa, as measured by their self-rated familiarity with specialized terms from texts from that domain. They therefore might have also been able to rely on their background knowledge of some economic topics, when interpreting the relations. The finding that the effect of domain knowledge is smaller for the items from the PDTB can therefore not be considered surprising.

This interaction also raises the question of what constitutes expertise. Experts were assumed to be more knowledgeable with respect to the topic of the text. This knowledge would have been gained through reading texts typical to the domain. In the case of the biomedical experts, this would more likely be research papers than newspapers; in the case of economics experts, this would more likely be economic newspapers than research papers. In our post-test questionnaire, biomedical experts indeed indicated that they read research papers more often than economists (mean 3.57 vs. 2.59 on a 1-5 Likert scale ranging from 'never' to 'daily'). This difference was even more distinct for biomedical research papers (3.41 vs. 1.30). Economics experts, on the other hand, read newspapers (and specifically business newspapers) more often than biomedic experts (3.5 vs. 2.35 for newspapers in general; 3.09 vs. 1.33 for business newspapers). Domain knowledge might thus consist not only of topic knowledge, but also of text genre familiarity. Such familiarity might help readers to infer the discourse relations in that genre. For example, newspaper texts are characterized by a so-called *inverted pyramid scheme*, where the first paragraph is followed by an ELABORATION in the subsequent paragraphs (Das et al., 2018). Methodology sections of research papers often contain many TEMPORAL relations (Bachand et al., 2014). Even readers who are not familiar with the domain of the text (e.g. psycholinguistic researchers when reading biomedical research papers) might use genre familiarity with the text structure to infer the discourse relation. A future line of research could attempt to further tease apart the influence of topic knowledge and genre familiarity on the impact of domain knowledge, and how these two factors separately contribute to inferring discourse relations.

### 6.2 Discourse relational cues

Besides examining the role of domain knowledge in inferring the correct discourse relation, we also set out to explore how readers infer discourse relations in the absence of domain knowledge. The first prediction was that non-connective linguistic cues for discourse relations would be used. We therefore varied the presence of a connective in the original text. Discourse relational cues were hypothesized to be more frequent in implicit than in (originally explicit) implicitated relations, which is why we expected implicit relations to be easier to infer than implicitated relations. However, no effect of the presence of a connective in the original text was found. In addition, high- and low-knowledge readers were not affected differently by whether the relation was originally marked. We can therefore not confirm the hypothesis that discourse relational cues in implicit relations facilitate discourse relational inferences.

The by-item analysis revealed that some discourse relations contained cues for the relation, even in the absence of a connective. For example, antonyms were present in CONTRAST relations and hypernyms in INSTANTIATION relations. Low-knowledge readers sometimes successfully retrieved the relation when such cues were present. However, they were not always sensitive to these cues. For example, a cue such as hypernyms did not always help low-knowledge readers to infer an INSTANTIATION relation, and nor did antonyms in CONTRAST relations. Instead, readers strongly diverged in the interpretations of these items. There are several possible explanations for these findings. First of all, non-connective linguistic signals for discourse relations are highly ambiguous, with many signaling a large variety of discourse relations. The cue might then exclude some possible relation interpretations, but not provide only one single likely interpretation. As a result, different readers might interpret the cue differently, diminishing the facilitative effect of additional relational cues in implicit relations. Secondly, signals for discourse relations (other than connectives) may require domain knowledge to interpret them. To illustrate, antonyms often occur in CONTRAST relations, but in order to know that two concepts are opposite, the reader should know what the words mean. This could explain why low-knowledge readers do not always pick up on these cues.

Another explanation for the diverging interpretations of items containing non-connective linguistic signals might lie outside the scope of the text itself and be influenced by characteristics of the reader. Scholman et al. (2020) show that some readers are more sensitive to contextual list signals than others. More specifically, participants in their study who had more reading experience (as measured by an Author Recognition Test), picked up on these cues more than participants who were less experienced readers. Only some readers might therefore have been able to employ these signals in inferring the relation, leading to differences in how relations containing such cues are interpreted. With respect to domain knowledge, high-knowledge readers had access to two strategies in interpreting the relation: non-connective linguistic signals and their knowledge base. The high-knowledge readers who were not sensitive to these non-connective discourse relational cues could then use their domain knowledge to infer the relation, whereas low-knowledge readers would not be able to interpret the relation if they did not detect these signals.

More research is needed on what these non-connective linguistic cues consist of as well as the extent to which readers are sensitive to them. For example, is a hypernym a reliable cue for INSTANTIATION relations? In addition, the examples provided in this study focused on the two relational arguments, but discourse structure might also guide relational inferences, for example when the arguments are part of a longer LIST structure. Furthermore, even though this study was conducted in English, we expect this effect to replicate in other languages. Discourse relational inferences are part of high-level discourse processes and the facilitative effect of connectives has also been replicated in various languages (Kamalski et al., 2008; Blumenthal-Dramé, 2021; Lyu et al., 2020). However, which discourse relational cues are present and the extent to which readers rely on this information may differ between languages (cf. Blumenthal-Dramé, 2021; Schwab and Liu, 2020). Moreover, further research should examine readers' sensitivity to these cues. The present study did not find a difference between implicit and implicitated relations, even though the latter has been argued to contain more discourse relational cues. Do readers notice non-connective signals for discourse relations? Or do they only adopt this information if they are experts on the domain of the text (as suggested in the present study) or have much reading experience (Scholman et al., 2020)? Such research would provide more insight in how readers establish coherence in a text.

### 6.3 Inferences in the absence of domain knowledge

The present study also set out to investigate what low-knowledge readers do when they lack the domain knowledge that is required to infer the correct discourse relation. Apart from using non-connective linguistic signals, we hypothesized that participants might resort to a default interpretation strategy and have a preference for causal and continuous relation interpretations in cases in which the relation was not inferred correctly. However, low-knowledge readers did not insert causal and continuous connectives in incorrect items to a greater extent than high-knowledge readers. We thus did not find evidence that domain knowledge influences readers' cognitive biases for causality and continuity.

In addition, we predicted that low-knowledge readers would prefer to leave the discourse relation underspecified. Rather than committing to a certain interpretations that might be incorrect, readers were hypothesized to make underspecified discourse interpretations and therefore provide connectives reflecting this underspecification. We found some evidence for this hypothesis, since low-knowledge readers inserted more ambiguous connectives in the first step than high-knowledge readers. Low-knowledge readers thus seem to avoid making a specific relation interpretation. However, it remains unclear what the reason for these underspecified interpretations is. On the one hand, it is possible that low-knowledge readers were unable to specify the relation further. On the other hand, low-knowledge readers might have processed the text less deeply and therefore not committed to a specific relation because they did not wish to do so. Future research could examine whether low-knowledge readers perform better when they are forced to process the text more deeply (cf. Scholman, 2019) to disentangle these two factors.

### 6.4 Limitations

Finally, we note some limitations of the present research. Firstly, the study aimed to balance the items among the different discourse relation senses, since different relation senses were hypothesized to yield differences in accuracy and interpretation biases. We did indeed find that CAUSE and INSTANTIATION relations were easier to infer for participants than CONCESSION relations and that INSTANTIATION and CONCESSION relations were often interpreted as being causal. Many of the initially selected CONTRAST relations had been annotated as CONCESSION in the PDTB 3.0. The lower performance on this relation sense could therefore partly be attributed to the disagreement about the gold label, since a CONCESSION interpretation might also have been possible. However, since relation sense was included as a covariate in the analysis, this does not affect the conclusions about the role of domain knowledge.

Another limitation is our manipulation of relation marking. It is possible that the relation might have become impossible to identify or has changed by removing the connective. In the first case, we would find floor effects on the implicitated relations, even for the high-knowledge readers. Overall, there were ten (out of 190) items for which none of the high-knowledge readers converged with the gold label. However, these were equally distributed over the implicit and implicitated condition. This suggests that the original relation could still be retrieved, even when the connective had been removed, also in the implicitated condition. In the case of multiple interpretations, convergence to the gold label is not reliable anymore. To account for the problem of multiple interpretations, we examined those implicitated relation items where several participants agreed on the same non-gold relation interpretation and assessed whether this interpretation was also possible. Including these alternative answers as correct still revealed the same pattern as above: high-knowledge readers interpreted the relation correctly more often than low-knowledge readers. Nevertheless, a study manipulating discourse relational cues specifically would provide further insight on this matter.

Furthermore, despite carefully selecting our participants, we cannot be sure that they were indeed as knowledgeable as they said they were. Nevertheless, we found a clear effect of domain knowledge in the BioDRB, suggesting that the biomedical experts were indeed more familiar in this domain than the economic experts. There is no reason to believe that the experts from the field of economics would be less knowledgeable than the participants from the biomedical domain.

### 6.5 Conclusion

To conclude, the current research provides more insight into the role of domain knowledge in discourse processing by examining discourse relation interpretations. Previous work has mainly focused on the influence of domain knowledge on text comprehension and recall (e.g. McNamara et al., 1996; Smith et al., 2021) or on whether or not discourse inferences are made in the absence of domain knowledge (e.g. Noordman and Vonk, 1998), showing that low-knowledge readers benefit more from coherence marking than high-knowledge readers and are less likely to make relational inferences during reading. However, these studies did not address how the discourse is interpreted differently by high- and low-knowledge readers. The present study shows that readers are able to interpret discourse relations correctly, even if they have little knowledge about the domain of the text. Still, high-knowledge readers make more correct (and more specific) discourse relation interpretations. This effect was established in biomedical research papers, a text type that targets a specialist audience, but not in economic newspapers, possibly because the genre is aimed to be accessible for both experts and novices in the field. Moreover, we found that readers adopt linguistic cues for inferring discourse relations, although this did not interact with the presence of a connective in the original text. A text without discourse connectives is therefore not necessarily detrimental for low-knowledge readers (cf. McNamara et al., 1996), as they can also establish coherence with other discourse cues. Still, these cues might be more challenging to low-knowledge readers as in some cases domain knowledge is required to detect them.

## Funding

This work was supported by the German Research Foundation (DFG) under Grant SFB 1102 ("Information Density and Linguistic Encoding"). The second and third author were supported by the European Research Council under Grant 948878 ("Individualized Interaction in Discourse").

## References

Fatemeh Torabi Asr and Vera Demberg. Implicitness of discourse relations. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 2669–2684, Mumbai, India, 2012. `https://aclanthology.org/C12-1163.pdf`.

Fatemeh Torabi Asr and Vera Demberg. Uniform Information Density at the level of discourse relations: Negation markers and discourse connective omission. In *Proceedings of the Inter-*

*national Conference on Computational Semantics (IWCS)*, pages 118–128, London, UK, 2015. https://aclanthology.org/W15-0117.pdf.

Félix-Hervé Bachand, Elnaz Davoodi, and Leila Kosseim. An investigation on the influence of genres and textual organisation on the use of discourse relations. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 454–468. Springer, 2014.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3): 255–278, 2013. doi:10.1016/j.jml.2012.11.001.

Alice Blumenthal-Dramé. The online processing of causal and concessive relations: Comparing native speakers of English and German. *Discourse Processes*, 58(7):642–661, 2021. doi:10.1080/0163853X.2020.1855693.

Anneloes R Canestrelli, Willem M Mak, and Ted J. M. Sanders. Causal connectives in discourse processing: How differences in subjectivity are reflected in eye movements. *Language and Cognitive Processes*, 28(9):1394–1413, 2013. doi:10.1080/01690965.2012.685885.

Ludivine Crible. Weak and strong discourse markers in speech, chat, and writing: Do signals compensate for ambiguity in explicit relations? *Discourse Processes*, 57(9):793–807, 2020. doi:10.1080/0163853X.2020.1786778.

Ludivine Crible. Negation cancels discourse-level processing differences: Evidence from reading times in concession and result relations. *Journal of Psycholinguistic Research*, 50(6):1283–1308, 2021. doi:10.1007/s10936-021-09802-2.

Debopam Das and Maite Taboada. RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52(1):149–184, 2018. doi:10.1007/s10579-017-9383-x.

Debopam Das, Tatjana Scheffler, Peter Bourgonje, and Manfred Stede. Constructing a lexicon of English discourse connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–365, 2018.

Vera Demberg, Fatemeh Torabi Asr, and Merel C. J. Scholman. How compatible are our discourse annotations? Insights from mapping RST-DT and PDTB annotations. *Dialogue & Discourse*, 10 (1):877–135, 2019. doi:10.5087/dad.2019.104.

Fernanda Ferreira and Suphasiree Chantavarin. Integration and prediction in language processing: A synthesis of old and new. *Current Directions in Psychological Science*, 27(6):443–448, 2018. doi:10.1177/0963721418794491.

Arthur C. Graesser, Murray Singer, and Tom Trabasso. Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3):371–395, 1994. doi:10.1037/0033-295X.101.3.371.

Herbert P. Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Speech acts*, pages 41–58. Brill, Leiden, 1975. doi:10.1163/9789004368811_003.

Cristina Grisot and Joanna Blochowiak. Temporal relations at the sentence and text genre level: The role of linguistic cueing and non-linguistic biases—An annotation study of a bilingual corpus. *Corpus Pragmatics*, 5:379–419, 2021. doi:10.1007/s41701-021-00104-5.

Jet Hoek, Sandrine Zufferey, Jacqueline Evers-Vermeul, and Ted J. M. Sanders. Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. *Journal of Pragmatics*, 121:113–131, 2017. doi:10.1016/j.pragma.2017.10.010.

Jet Hoek, Sandrine Zufferey, Jacqueline Evers-Vermeul, and Ted J. M. Sanders. The linguistic marking of coherence relations: Interactions between connectives and segment-internal elements. *Pragmatics & Cognition*, 25(2):276–309, 2018. doi:10.1075/pc.18016.hoe.

Jet Hoek, Merel C. J. Scholman, and Ted J. M. Sanders. Is there less annotator agreement when the discourse relation is underspecified? In *Proceedings of the Workshop in Integrating Perspectives in Discourse Annotation (DiscAnn)*, pages 1–5, Tübingen, Germany, 2021.

Judith Kamalski, Ted J. M. Sanders, and Leo Lentz. Coherence marking, prior knowledge, and comprehension of informative and persuasive texts: Sorting things out. *Discourse Processes*, 45 (4-5):323–345, 2008. doi:10.1080/01638530802145486.

Andrew Kehler, Laura Kertz, Hannah Rohde, and Jeffrey L. Elman. Coherence and coreference revisited. *Journal of Semantics*, 25(1):1–44, 2008. doi:10.1093/jos/ffm018.

Walter Kintsch and Teun A. Van Dijk. Toward a model of text comprehension and production. *Psychological Review*, 85(5):363–394, 1978. doi:10.1037/0033-295X.85.5.363.

Yudai Kishimoto, Shinnosuke Sawada, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. Improving crowdsourcing-based annotation of japanese discourse relations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Judith Köhne-Fuetterer, Heiner Drenhaus, Francesca Delogu, and Vera Demberg. The online processing of causal and concessive discourse connectives. *Linguistics*, 59(2):417–448, 2021. doi:10.1515/ling-2021-0011.

Arnout W. Koornneef and T. J. M. Sanders. Establishing coherence relations in discourse: the influence of implicit causality and connectives on pronoun resolution. *Language and Cognitive Processes*, 28:1169–1206, 2013. `https://www.tandfonline.com/doi/full/10.1080/01690965.2012.699076`.

Gina R. Kuperberg, Martin Paczynski, and Tali Ditman. Establishing causal coherence across sentences: An ERP study. *Journal of Cognitive Neuroscience*, 23(5):1230–1246, 2011. doi:10.1162/jocn.2010.21452.

Siqi Lyu, Jung-Yueh Tu, and Chien-Jer Charles Lin. Processing plausibility in concessive and causal relations: Evidence from self-paced reading and eye-tracking. *Discourse Processes*, 57 (4):320–342, 2020.

Danielle S. McNamara. Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 55(1):51–62, 2001. doi:10.1037/h0087352.

Danielle S. McNamara, Eileen Kintsch, Nancy Butler Songer, and Walter Kintsch. Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14(1):1–43, 1996. doi:10.1207/s1532690xci1401_1.

Evelyn Milburn, Tessa Warren, and Michael Walsh Dickey. World knowledge affects prediction as quickly as selectional restrictions: Evidence from the visual world paradigm. *Language, cognition and neuroscience*, 31(4):536–548, 2016. doi:10.1080/23273798.2015.1117117.

Eleni Miltsakaki, Aravind Joshi, Rashmi Prasad, and Bonnie Webber. Annotating discourse connectives and their arguments. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 9–16, Boston, MA, USA, 2004.

John D. Murray. Connectives and narrative text: The role of continuity. *Memory & Cognition*, 25 (2):227–236, 1997. doi:10.3758/BF03201114.

Leo G. M. Noordman and Wietske Vonk. Memory-based processing in understanding causal information. *Discourse Processes*, 26(2-3):191–212, 1998. doi:10.1080/01638539809545044.

Leo G M Noordman, Wietske Vonk, and Henk J Kempff. Causal inferences during the reading of expository texts. *Journal of Memory and Language*, 31(5):573–590, 1992. `https://www.sciencedirect.com/science/article/pii/0749596X9290029W`.

Leo G M Noordman, Wietske Vonk, Reinier Cozijn, and Stefan Frank. Causal inferences and world knowledge. In *Inferences during reading*, pages 260–289. Cambridge University Press, 2015.

Tenaha O'Reilly and Danielle S. McNamara. Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes*, 43(2):121–152, 2007. doi:10.1080/01638530709336895.

Florian Pusse, Asad Sayeed, and Vera Demberg. LingoTurk: Managing crowdsourced tasks for psycholinguistics. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 57–61, San Diego, CA, 2016.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018. URL `https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf`.

Livio Robaldo and Eleni Miltsakaki. Corpus-driven semantics of concession: Where do expectations come from? *Dialogue & Discourse*, 5(1):1–36, 2014. doi:10.5087/dad.2014.101.

Hannah Rohde, Anna Dickinson, Nathan Schneider, Christopher Clark, Annie Louis, and Bonnie Webber. 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)*, pages 49–58, Berlin, Germany, 2016.

Ted J. M. Sanders. Coherence, causality and cognitive complexity in discourse. In *Proceedings/Actes SEM-05, First International Symposium on the Exploration and Modelling of Meaning*, pages 105–114, Toulouse, France, 2005.

Ted J. M. Sanders and Leo G. M. Noordman. The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, 29(1):37–60, 2000. doi:10.1207/S15326950dp2901_3.

Ted J. M. Sanders, Wilbert P. M. S. Spooren, and Leo G. M. Noordman. Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35, 1992. https://www.tandfonline.com/doi/abs/10.1080/01638539209544800.

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

Merel C. J. Scholman and Vera Demberg. 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)*, pages 24–33, Valencia, Spain, 2017a.

Merel C. J. Scholman and Vera Demberg. Examples and specifications that prove a point: Identifying elaborative and argumentative discourse relations. *Dialogue & Discourse*, 8(2):56–83, 2017b. doi:10.5087/dad.2017.203.

Merel C. J. Scholman, Hannah Rohde, and Vera Demberg. "On the one hand" as a cue to anticipate upcoming discourse structure. *Journal of Memory and Language*, 97:47–60, 2017. doi:10.1016/j.jml.2017.07.010.

Merel C. J. Scholman, Vera Demberg, and Ted J. M. Sanders. Individual differences in expecting coherence relations: Exploring the variability in sensitivity to contextual signals in discourse. *Discourse Processes*, 57(10):844–861, 2020. doi:10.1080/0163853X.2020.1813492.

Merel C. J. Scholman, Tianai Dong, Frances Yung, and Vera Demberg. Discogem: A crowdsourced corpus of genre-mixed implicit discourse relations. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC'22)*, Marseille, France, 2022. European Language Resources Association (ELRA).

Juliane Schwab and Mingya Liu. Lexical and contextual cue effects in discourse expectations: Experimenting with German 'zwar... aber'and English 'true/sure... but'. *Dialogue & Discourse*, 11(2), 2020. doi:10.5087/dad.2020.203.

Erwin M. Segal, Judith F. Duchan, and Paula J. Scott. The role of interclausal connectives in narrative structuring: Evidence from adults' interpretations of simple stories. *Discourse Processes*, 14 (1):27–54, 1991. doi:10.1080/01638539109544773.

Reid Smith, Pamela Snow, Tanya Serry, and Lorraine Hammond. The role of background knowledge in reading comprehension: A critical review. *Reading Psychology*, 42(3):214–240, 2021. doi:10.1080/02702711.2021.1888348.

Caroline Sporleder and Alex Lascarides. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(3):369–416, 2008. doi:10.1017/S1351324906004451.

Paul Van den Broek. Using texts in science education: Cognitive processes and knowledge representation. *Science*, 328:453–456, 2010. doi:10.1126/science.1182594.

Noortje J. Venhuizen, Matthew W Crocker, and Harm Brouwer. Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. *Discourse Processes*, 56 (3):229–255, 2019. doi:10.1080/0163853X.2018.1448677.

Bonnie Webber. What excludes an alternative in coherence relations? In *Proceedings of the International Conference on Computational Semantics (IWCS)*, pages 921–950, Potsdam, Germany, 2013. `https://aclanthology.org/W13-0124.pdf`.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. A discourse-annotated corpus of conjoined VPs. In *Proceedings of the 10th Linguistic Annotation Workshop (LAW X)*, pages 22–31, Berlin, Germany, 2016.

Frances Yung, Vera Demberg, and Merel C. J. Scholman. Crowdsourcing discourse relation annotations by a two-step connective insertion task. In *Proceedings of the 13th Linguistic Annotation Workshop (LAW)*, pages 16–25, 2019. `https://aclanthology.org/W19-4003.pdf`.

Sandrine Zufferey and Liesbeth Degand. Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory*, 13(2):399–422, 2017. doi:10.1515/cllt-2013-0022.

## Appendix A. Relation sense classification

(i) Cause

- CAUSE
- CAUSE+BELIEF
- CAUSE+SPEECHACT
- PURPOSE

(ii) Temporal

- SYNCHRONOUS
- ASYNCHRONOUS

(iii) Contrast

(iv) Concession

- CONCESSION
- CONCESSION+SPEECHACT

(v) Positive expansion

- SIMILARITY
- CONJUNCTION
- EQUIVALENCE
- INSTANTIATION
- LEVEL-OF-DETAIL
- MANNER

(vi) Negative expansion

- DISJUNCTION
- EXCEPTION
- SUBSTITUTION

(vii) Condition

- CONDITION
- CONDITION+SPEECHACT
- NEGATIVE-CONDITION
- NEGATIVE-CONDITION+SPEECHACT

## Appendix B. Means across conditions

Table 6: Mean percentage of correct answers per condition.

| | | **BioDRB** | | **PDTB** | | |
|---|---|---|---|---|---|---|
| | | bio | eco | bio | eco | *mean* |
| **Implicitated** | Result | 57.3 | 47.2 | 72.6 | 72.1 | *62.4* |
| | Instantiation | 71.4 | 41.0 | 46.8 | 49.5 | *52.0* |
| | Concession | 48.4 | 28.6 | 48.9 | 58.6 | *48.6* |
| | Contrast | 34.7 | 27.6 | 42.9 | 14.3 | *31.2* |
| | *mean* | *52.2* | *36.2* | *54.3* | *58.7* | |
| **Implicit** | Result | 70.3 | 69.1 | 64.1 | 65.1 | *67.1* |
| | Instantiation | 65.1 | 51.9 | 64.1 | 60.5 | *60.7* |
| | Concession | 41.0 | 36.1 | 54.9 | 54.2 | *49.1* |
| | Contrast | 39.1 | 25.9 | 33.3 | 20.0 | *31.7* |
| | *mean* | *54.1* | *45.4* | *58.7* | *57.0* | |

## Appendix C. Model output summaries

### Models described in Section 5.2

Table 7 shows the model output of the subset analysis on the PDTB items, in which the effect of corpus was not significant. The effect was significant in the items in the BioDRB, which is displayed in Table 8.

Table 7: Subset analysis on PDTB items. Model specification: `correctness ~relationsense + expertise + relationmarking + (1 | workerid) + (1 + expertise | questionid)`

| | **Estimate** | **Std. Error** | **Z value** | **P value** |
|---|---|---|---|---|
| Intercept | 0.19 | 0.17 | 1.11 | .27 |
| Relationsense:CAUSE | 0.73 | 0.27 | 2.70 | **<.001** |
| Relationsense:CONTRAST | -1.28 | 0.67 | -1.92 | .06 |
| Relationsense:INSTANTIATION | 0.07 | 0.27 | 0.25 | .80 |
| Expertise | -0.04 | 0.08 | -0.52 | .60 |
| Relationmarking | 0.06 | 0.12 | 0.52 | .60 |

Table 8: Subset analysis on BioDRB items. Model specification: `correctness ~ relationsense + expertise + relationmarking + (1 | workerid) + (1 + expertise | questionid)`

| | **Estimate** | **Std. Error** | **Z value** | **P value** |
|---|---|---|---|---|
| Intercept | -0.54 | 0.21 | -2.55 | **.01** |
| Relationsense:CAUSE | 1.11 | 0.28 | 3.92 | **<.001** |
| Relationsense:CONTRAST | -0.40 | 0.29 | -1.40 | .16 |
| Relationsense:INSTANTIATION | 0.81 | 0.29 | 2.80 | **<.01** |
| Expertise | 0.30 | 0.09 | 3.47 | **<.001** |
| Relationmarking | 0.15 | 0.12 | 1.24 | .21 |

### Models described in Section 5.3.3

Table 9 shows the output of the model in which the ambiguity of the first step insertions was predicted. To examine the interaction between corpus and expertise, separate analyses were performed for the items from the PDTB (Table 10) and the BioDRB (Table 11).

Table 9: Output of the model predicting ambiguity of the insertions in the first step. Model specification: `ambiguity ~ relationsense + corpus*expertise + (1 + corpus | workerid) + (1 + expertise | questionid)`

| | Estimate | Std. Error | Z value | P value |
|---|---|---|---|---|
| Intercept | 0.33 | 0.12 | 2.86 | **<.01** |
| Relationsense:CAUSE | -0.97 | 0.15 | -6.36 | **<.001** |
| Relationsense:CONTRAST | 0.05 | 0.19 | 0.28 | .78 |
| Relationsense:INSTANTIATION | -1.47101 | 0.16 | -9.24 | **<.001** |
| Corpus | -0.08 | 0.06 | -1.30 | .20 |
| Expertise | -0.15 | 0.08 | -1.95 | .05 |
| Corpus:expertise | -0.14 | 0.05 | -3.08 | **<.01** |

Table 10: Subset analysis on items from the PDTB. Model specification: `ambiguity ~ relationsense + corpus*expertise + (1 | workerid) + (1 | questionid)`

| | Estimate | Std. Error | Z value | P value |
|---|---|---|---|---|
| Intercept | 0.40 | 0.14 | 2.88 | **<.01** |
| Relationsense:CAUSE | -0.98 | 0.21 | -4.62 | **<.001** |
| Relationsense:CONTRAST | -0.09 | 0.50 | -0.19 | .85 |
| Relationsense:INSTANTIATION | -1.39 | 0.22 | -6.38 | **<.001** |
| Expertise | -0.01 | 0.09 | -0.09 | .93 |

Table 11: Subset analysis on items from the BioDRB. Model specification: `ambiguity ~ relationsense + corpus*expertise + (1 | workerid) + (1 + expertise | questionid)`

| | Estimate | Std. Error | Z value | P value |
|---|---|---|---|---|
| Intercept | 0.29 | 0.17 | 1.76 | .08 |
| Relationsense:CAUSE | -1.02 | 0.22 | -4.71 | **<.001** |
| Relationsense:CONTRAST | 0.04 | 0.21 | 0.19 | .85 |
| Relationsense:INSTANTIATION | -1.59 | 0.24 | -6.72 | **<.001** |
| Expertise | -0.29 | 0.09 | -3.03 | **<.01** |