Is there less annotator agreement when the discourse relation is underspecified?

Jet Hoek
Centre for Language Studies
Radboud University Nijmegen
jet.hoek@ru.nl

Merel C.J. Scholman
Language Science and Technology
Saarland University
m.c.j.scholman@coli.uni-saarland.de

Ted J.M. Sanders
Utrecht Institute of Linguistics OTS
Utrecht University
t.j.m.sanders@uu.nl

Abstract

When annotating coherence relations, inter-annotator agreement tends to be lower on implicit relations than on relations that are explicitly marked by means of a connective or a cue phrase. This paper explores one possible explanation for this: the additional inferencing involved in interpreting implicit relations compared to explicit relations. If this is the main source of disagreements, agreement should be highly related to the specificity of the connective. Using the CCR framework, we annotated relations from TED talks that were marked by a very specific marker, marked by a highly ambiguous connective, or not marked by means of a connective at all. We indeed reached higher inter-annotator agreement on explicit than on implicit relations. However, agreement on underspecified relations was not necessarily in between, which is what would be expected if agreement on implicit relations mainly suffers because annotators have less specific instructions for inferring the relation.

1 Introduction

Discourse-annotated corpora allow coherence researchers to study the distribution and linguistic realization of coherence relations. Such sources of information enable us to take the study of coherence relations an important step forward. However, discourse annotation has proven to be a difficult task, which is reflected in low inter-annotator agreement (IAA) scores (Artstein and Poesio, 2008; Spooren and Degand, 2010). One explanation for this observation is that coherence is a feature of the mental representation that readers form of a text, rather than of the linguistic material itself (e.g., Sanders et al., 1992). Discourse annotation thus relies on annotators' interpretation of a text, which makes it a particularly difficult task.

In order to gain a deeper understanding of the difficulties associated with reaching sufficient inter-annotator agreement on coherence relation annotations, we need more data on the agreement on different types of relations. Unfortunately, many annotation studies report only overall agreement scores (not distinguishing between different connectives or relation types), or only report agreement scores after the annotators have reconciled disagreements.

The few studies that did report separate agreement statistics have shown that annotators tend to agree more when annotating explicit coherence relations, which are signalled by a connective or cue phrase (e.g. *because*, *as a result*; we will use 'connectives' as a shorthand for the combined category), than when annotating implicit coherence relations, which contain no or less explicit linguistic markers on which annotators can base their decision (e.g., Miltsakaki et al., 2004; Prasad et al., 2008).

This can be considered an expected finding, given that connectives provide comprehenders with "processing instructions" on how to connect incoming text inputs to previously read segments (Britton, 1994; Canestrelli et al., 2013; Gernsbacher, 1997; Sanders and Noordman, 2000). However, it does raise concerns about the validity and added value of coherence annotation efforts: if annotators need connectives in order to reach sufficient agreement on the sense of the relation at hand, is the annotation focused on the coherence relation or rather on the connective? If discourse annotation is mainly focused on connectives, one can wonder how valuable the annotated label is? After all, annotations for explicit connectives such as *if* can, depending on the discourse annotation framework, be done largely or completely automatically. The value of manual annotation comes from disambiguating between relational senses when more than one reading could be inferred. This can occur when the connective is ambiguous (underspecified relative to the inferred relation, Spooren, 1997) or when a

relation is not explicitly marked with a connective at all.

The current study functions as an initial investigation of agreement on relations with various markers. By annotating relations marked by specific connectives (*because*, *in addition* and *even though*), highly ambiguous connectives (*and* and *but*), or no connective, we aim to investigate to what extent agreement between annotators is dependent on the specificity of the connective that marks the coherence relation. If the amount of inferencing involved in interpreting a coherence relation is the main source of differences in IAA scores between implicit and explicit relations, we expect IAA to decrease as a function of connective specificity: lowest IAA on implicit relations, intermediate IAA on underspecified relations, and highest IAA on relations marked by a specific connective.

2 Method

2.1 Materials

The data set contained 350 relations taken from transcribed English TED talks: 100 implicit relations, 100 relations marked by underspecified connectives (*and/but*), and 150 relations marked by more specific connectives (*because/in addition/even though*). TED talks are highly structured speeches that are minutely prepared and are meant to provide targeted information on various topics.

The 100 implicit coherence relations were randomly selected from the English part of the TED-MDB corpus (Zeyrek et al., 2019), as well as 50 relations marked by *and* and all relations marked by *but* (n=47). We used the Ted Corpus Search Engine (Hasebe, 2015) to randomly select 50 coherence relations each marked by *because*, *in addition*, and *even though*, plus 3 additional *but*-relations. [1] The selected relations were displayed in their original context during annotation.

2.2 Annotation framework

The Cognitive approach to Coherence Relations (CCR) was used to annotate all relations (Sanders et al., 1992, see Hoek et al., 2019 for an up-to-date version). CCR depicts coherence relations in terms of cognitive primitives. Crucial primitives are POLARITY, BASIC OPERATION, SOURCE OF COHERENCE, and ORDER OF THE SEGMENTS.

[1]The full annotated data set can be accessed at https://tinyurl.com/rgdjear.

POLARITY distinguishes between positive and negative relations. A relation is positive if the propositions P and Q (expressed in the discourse segments S_1 and S_2) are linked without a negation of one of these propositions. A relation is negative if the negative counterpart of either P or Q functions in the relation.

BASIC OPERATION distinguishes between causal and additive relations. In causal relations, an implication relation (P $\rightarrow$ Q) can be deduced between the two segments. In additive relations, the segments are connected as a conjunction (P & Q). Temporal relations, in which the segments are ordered in time, are considered a subclass of additive relations. Conditional relations are considered a subclass of causal relations.

SOURCE OF COHERENCE distinguishes between objective and subjective relations. Subjective relations express the speaker's opinion, argument, claim, or conclusion. Objective relations, on the other hand, describe situations that occur in the real world. Temporal relations are assumed to always be objective.

ORDER OF THE SEGMENTS applies to causal and conditional relations. In a basic order relation, the antecedent (P) is S_2, followed by the consequent (Q) as S_1. In a non-basic order relation, P maps onto S_2 and Q onto S_1. The ordering of events in temporal relations (chronological, anti-chronological, synchronous) is captured by TEMPORALITY (see Evers-Vermeul et al., 2017).

2.3 Connective choice

Because is a typical, specific marker of causal coherence relations. *In addition* is a typical, specific marker of additive coherence relations. *Even though* is considered a prototypical connective for negative causal relations.

And is considered an underspecified connective: it can mark a variety of relations, including positive additive relations, as in Example (1), and positive causal relations, as in Example (2), but it can also mark negative additive and causal relations (see Crible et al., 2019). It tends to be used most frequently in positive additive relations, however.

(1) I am terrible at playing darts and I don't know how to play pool.

(2) I missed the dart board and someone lost an eye.

But is also considered an underspecified connective: it can mark negative additive relations, as in Example (3), as well as negative causal relations, as in Example (4). Its distribution is different to *and*, in that it has a less strongly associated default interpretation.

(3) I am terrible at playing darts, but I am a champion in pool.

(4) I missed the dart board, but everybody is safe.

2.4 Annotation procedure

The first two authors, both expert coders, annotated discourse relations according to the CCR framework, without specific within-genre training or intermediate discussion. They assigned single values for every primitive. In cases where the two annotators disagreed, the third author provided an additional annotation. The majority vote was then chosen as the true value. This was used to establish ambiguity of connective usage.

2.5 Inter-annotator agreement metrics

In order to evaluate inter-annotator agreement and gain a comprehensive overview of the agreement, we use different metrics and methods.

Regarding the metrics, we report on three different measures: percentage agreement (also known as observed agreement), Cohen's Kappa κ (Cohen, 1960) and AC_1 (Gwet, 2001). Kappa is the most commonly used agreement measure, but it can behave erratically in certain situations; a problem known as Kappa's Paradox (Feinstein and Cicchetti, 1990). Specifically, when data sets are characterized by an uneven distribution of categories, Kappa's values can be relatively low, despite a higher percentage of observed agreement (see also Hoek and Scholman, 2017). AC_1 was introduced to address this issue. Since some types of relations will likely occur more frequently than others in our data set per connective, we consider both Kappa and AC_1 in order to get a full overview of the agreement.

Regarding the method of the inter-annotator agreement, we consider the agreement on the full "label" of the relation (the combination of all values on the dimensions). Full labels give a straightforward impression of a connective's specificity (i.e., the more types of labels, the more ambiguity) and make for better comparison to annotation efforts in other frameworks, which only use end labels (although there is not necessarily a 1:1 correspondence between the full CCR relation labels and relation labels from other approaches, see Sanders et al., 2018).

Connective		%	κ	AC_1
explicit	*because*	84	.68	.68
	in addition	82	.57	.69
	even though	78	.58	.74
underspecified	*and*	74	.58	.71
	but	58	.39	.46
implicit	Ø	66	.58	.64

Table 1: IAA per connective type and connective

3 Results

We exclude the ORDER OF THE SEGMENTS from our analyses. Determining ORDER is largely trivial for specific connectives (indeed, we reached 100% agreement) and the only source of disagreement for the underspecified and implicit relations was the direct result of a disagreement on BASIC OPERATION (i.e., NA order for additive relations versus basic/non-basic order for causal relations).

Connectives and their assigned senses First, we focus on the annotated labels per connective to answer the question of whether underspecified connectives are truly underspecified, when compared to the specific connectives. Moreover, we examine the different senses assigned to implicit relations to determine how "underspecified" such relations are.

Figure 1 shows the distribution of relations per connective. As assumed, *and* and *but* were more ambiguous than *because*, *in addition*, and *even though*. The largest variety of relation labels was used for the implicit relations.

Agreement per connective Next, we compare the inter-annotator agreement of the two coders, to determine whether agreement on underspecified connectives differs from agreement on specific connectives and from agreement on implicit relations.

Table 1 shows the inter-annotator agreement for each connective. In line with IAA statistics from other annotation efforts (e.g., Miltsakaki et al., 2004; Prasad et al., 2008), agreement was lower on the implicit relations than on the explicit relations. Note that this difference is smaller according to

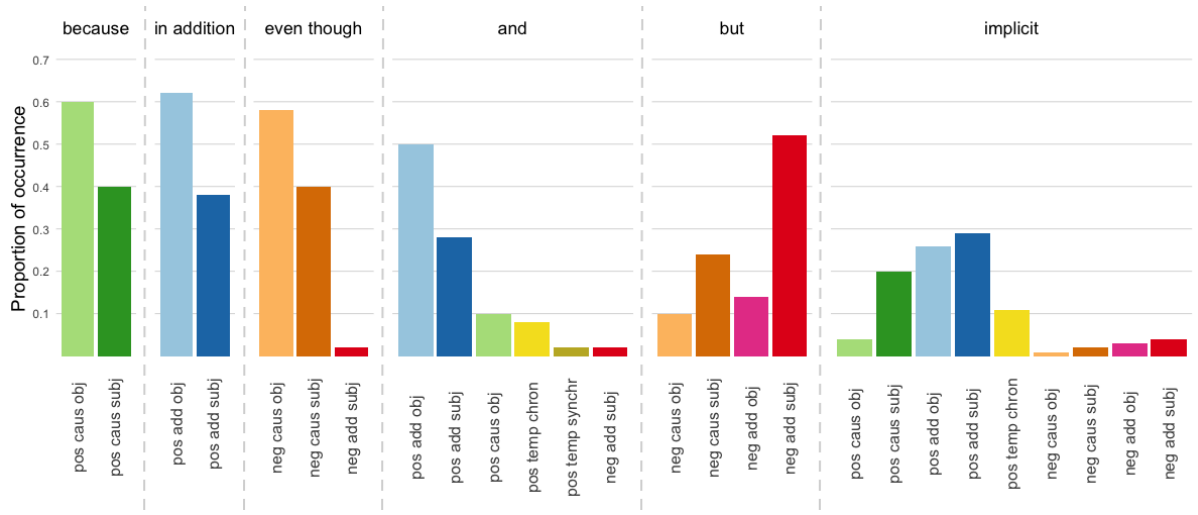

Figure 1: Distribution of relations, per connective.

the two agreement statistics that correct for chance agreement (Kappa and AC_1), but that the difference is still considerable in AC_1, which better takes into account the prevalence of the various categories. However, agreement on the underspecified relations was not necessarily in between. While IAA on relations marked by *and* was comparable with IAA on the explicit relations, agreement was much lower on relations marked by *but*.

4 Discussion

Much like the IAA statistics reported for other annotation efforts, we reached less agreement on implicit relations than on relations that were explicitly marked. However, the level of agreement reached for relations marked by ambiguous connectives *and* and *but* suggests that lower IAA on implicit relations cannot be straightforwardly explained by the specificity of the marker: We did not find an intermediate IAA score on the underspecified relations, and both in the absence of a connective and in the presence of *and*, more types of relations are available than in the presence of *but*, the connective for which we reached the lowest IAA.

Regarding the higher agreement on *and* compared to *but*, we could in part attribute this to the difference in default interpretations. Even though *and* can mark a larger variety of relations than *but*, it is associated more strongly with a default interpretation (78% of *and*-occurrences were positive additive, compared to 66% of *but*-occurrences being negative additive). This stronger default interpretation of *and* likely resulted in more agreement between the annotators. It emphasizes the need for further studies investigating a larger variety of underspecified connectives.

We can further interpret the low agreement on *but* using the primitive-specific annotations: the majority of disagreements on *but*-relations was on BASIC OPERATION, as was for instance the case for example (5).

(5) The US government says it doesn't use torture, and we condemn other countries, like Iran and North Korea, for their use of torture. **But** some people think the so-called worst of the worst deserve it: terrorists, mass murderers, the really "bad" people.

Under the negative causal reading, some people think really bad people deserve to be tortured, *even though* the US government does not support the practice; the fact that your government condemns something might plausibly lead you to condemn it too. Under the negative additive reading, this fragment presents merely two opposite viewpoints: some people condemn torture, *while* others support it (at least in some cases). The distinction between negative additive and negative causal relations corresponds to the distinction between contrast and concessive/denial-of-expectation relations in many other frameworks. Agreement statistics from other annotation efforts indicate that this distinction is a notoriously difficult one to make when coding corpus data (e.g., Robaldo and Miltsakaki, 2014; Degand and Zufferey, 2013).

While implicit relations can also express relations with negative polarity (see e.g., Figure 1), the specific interpretation problems with contrastive

relations do not seem to have a big effect on the agreement on implicit relations. Negative relations tend to be explicitly marked much more often than positive relations (e.g., Asr and Demberg, 2012; Hoek et al., 2017) and thus only make up a modest percentage of implicit relations. And while negative relations tend to be implicit more often in spoken than in written language, spoken language offers alternative ways to express contrast, such as topicalization and sentence stress (Rehbein et al., 2016).

Although the results suggest that increased inferencing does not necessarily lead to more disagreements, it is likely that the ambiguity of implicit relations does negatively impact the IAA scores. Implicit relation annotation is characterized by the added complexity of it not being clear *which* relation should be annotated, since more than one relation can hold between two segments (e.g., Rohde et al., 2018; Scholman and Demberg, 2017). For example, the originally implicit relation in (6), taken from our data set, can be interpreted in (at least) two ways: the second segment presents a reason for the first segment ('because'), or it supplies an alternative ('instead'). Note that these relations can hold at the same time.

(6) Prudent investing and finance theory aren't subordinate to sustainability. [BECAUSE INSTEAD] They're compatible.

Multiple relations can also hold between segments that are connected by an explicit connective, but in those cases, the connective supplies a clear cue as to which relation should be annotated.

In sum, the current study showed that IAA scores on underspecified relations do not necessarily fall in between the scores of explicit and implicit relations, which is what would be expected if IAA on implicit relations mainly suffers because annotators have less specific instructions for inferring the relation. Hence, our results indicate how implicit and underspecified coherence relations remain a major challenge for the field, both in terms of annotation practice and in terms of theoretical implications: how do humans deal with so many ambiguous relations in everyday communication?

Acknowledgments

The second author was supported by the German Research Foundation (DFG) as part of SFB 1102 "Information Density and Linguistic Encoding".

References

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

Fatemeh Torabi Asr and Vera Demberg. 2012. Implicitness of discourse relations. In *Proceedings of COLING 2012*, pages 2669–2684.

Bruce K. Britton. 1994. *Understanding expository text: Building mental structures to induce insights*. Academic Press.

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

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

Ludivine Crible, Ágnes Abuczki, Nijolė Burkšaitienė, Péter Furkó, Anna Nedoluzhko, Sigita Rackevičienė, Giedrė Valūnaitė Oleškevičienė, and Šárka Zikánová. 2019. Functions and translations of discourse markers in ted talks: A parallel corpus study of underspecification in five languages. *Journal of Pragmatics*, 142:139–155.

Liesbeth Degand and Sandrine Zufferey. 2013. Annotating the meaning of discourse connectives in multilingual corpora. *Corpus linguistics and linguistic theory*, pages 1–24.

Jacqueline Evers-Vermeul, Jet Hoek, and Merel C.J. Scholman. 2017. On temporality in discourse annotation: Theoretical and practical considerations. *Dialogue & Discourse*, 8(2):1–20.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. The problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549.

Morton Ann Gernsbacher. 1997. Coherence cues mapping during comprehension. *Processing interclausal relationships. Studies in the production and comprehension of text*, pages 3–22.

Kilem Gwet. 2001. *Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters*. Gaithersburg, MD: STATAXIS Publishing Company.

Yoichiro Hasebe. 2015. Design and implementation of an online corpus of presentation transcripts of ted talks. *Procedia: Social and Behavioral Sciences*, 24:174–182.

Jet Hoek, Jacqueline Evers-Vermeul, and Ted J.M. Sanders. 2019. Using the cognitive approach to coherence relations for discourse annotation. *Dialogue & Discourse*, 10(2):1–33.

Jet Hoek and Merel C.J. Scholman. 2017. Evaluating discourse annotation: Some recent insights and new approaches. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*, pages 1–13, Toulouse, France.

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

Eleni Miltsakaki, Aravind Joshi, Rashmi Prasad, and Bonnie Webber. 2004. 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.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 2961–2968, Marrakech, Marocco.

Ines Rehbein, Merel C. J. Scholman, and Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 23–28, Portoroz, Slovenia.

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

Hannah Rohde, Alexander Johnson, Nathan Schneider, and Bonnie Webber. 2018. Discourse coherence: Concurrent explicit and implicit relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2267.

Ted J. M. Sanders, Vera Demberg, Jet Hoek, Merel C. J. Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2018. Unifying dimensions in discourse relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, ahead of print:1–71.

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

Ted J.M. Sanders, Wilbert P.M.S. Spooren, and Leo G.M. Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35.

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

Wilbert Spooren. 1997. The processing of underspecified coherence relations. *Discourse Processes*, 24(1):149–168.

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

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. Ted multilingual discourse bank (TED-MDB): a parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, pages 1–27.