Mapping discourse annotations

How consistent are our discourse annotations?
Insights from mapping RST-DT and PDTB annotations

Vera Demberg*,**, Saarland University
Fatemeh Torabi Asr†, Indiana University
Merel Scholman‡, Saarland University

Discourse-annotated corpora are an important resource for the community. However, these corpora are often annotated according to different frameworks, making comparison of the annotations difficult. This is unfortunate, since mapping the existing annotations would result in more (training) data for researchers in automatic discourse relation processing and researchers in linguistics and psycholinguistics. In this article, we present an effort to map two large corpora onto each other: the Penn Discourse Treebank and the Rhetorical Structure Theory Discourse Treebank. We first propose a method for aligning the discourse segments, and then evaluate the observed against the expected mappings for explicit and implicit relations separately. We find that while agreement on explicit relations is reasonable, agreement between the frameworks on implicit relations is astonishingly low. We identify sources of systematic discrepancies between the two annotation schemes; many of the differences in annotation can be traced back to different operationalizations and goals of the PDTB and RST frameworks. We discuss the consequences of these discrepancies for future annotation, and the usability of the mapped data for theoretical studies and the training of automatic discourse relation labellers.

1. Introduction

In recent years, we have seen an increase in attention to discourse processing, in terms of discourse relation labelling as a task in automatic language processing, as well as
a feature for improving down-stream tasks such as machine translation (Meyer and Popescu-Belis 2012, Popescu-Belis 2016), question answering (Jansen, Surdeanu, and Clark 2014; Sharp et al. 2015) and sentiment analysis (Somasundaran et al. 2009; Zhou et al. 2011; Zirn et al. 2011). Progress on this topic has been made possible through the large-scale annotation of text corpora with discourse relation labels, most notably the Penn Discourse Treebank (PDTB; Prasad et al. 2008) and the RST Treebank (RST-DT; Carlson, Marcu, and Okurowski 2003) for English. There are also new resources for other languages, as well as numerous annotation efforts currently under way (see, for example, Oza et al. 2009; Stede and Neumann 2014).

However, as of yet there is no consensus on a single discourse relation labelling scheme. Existing discourse frameworks share basic notions of what a coherence relation is, and many of them make relation sense distinctions that have similar underlying ideas, but frameworks differ in how many and which types of relations they distinguish. Recently, however, a proposal has been developed for an ISO norm for discourse relation annotation (Bunt and Prasad 2016), in order to encourage future annotation efforts to follow a single scheme.

For resources that are already available, it would however also be very helpful to map existing annotations onto one another. This would allow researchers in automatic discourse relation processing to use larger amounts of training data, and adapt a discourse relation parser from one language to another more easily, even if the resources available for that language were annotated according to different annotation schemes. For researchers in linguistics, psycholinguistics or textual coherence, mapping relations would facilitate comparisons across corpora and languages, and enable them to identify more data related to the phenomena under investigation (Asr and Demberg 2015; Zufferey and Gygax 2016, among others).

The problem of differences in annotations has been recognized by the community, and several recent papers have suggested mapping schemes based on definitions of the different discourse relations according to the discourse annotation manuals (Hovy and Maier 1995; Chiarcos 2014; Benamara and Taboada 2015; Rehbein, Scholman, and Demberg 2016; Scheffler and Stede 2016). There is however to date no large-scale practical evaluation of whether expected correspondences between discourse relation labels of different schemes really bear out in practical annotations. For instance, it is possible that there is a gap between annotation manuals and implicit knowledge that annotators acquired during training, or biases introduced in annotation that help to increase annotator agreement. The goal of the present article is to empirically assess the agreement between the two most prominent discourse relation annotation schemes: PDTB and RST. The PDTB and RST-DT corpora happen to be annotated on the same text (several sections of the Wall Street Journal), thus allowing for a large-scale comparison of annotations. Furthermore, a theoretical mapping for discourse relations has recently become available for these relation frameworks (Sanders et al. Submitted). These allow us to compare predicted correspondences between labels with actual annotations.

A first challenge in mapping existing annotations is that PDTB and RST use different rules for segmenting a discourse into elementary units. In this article, we first propose a method for aligning discourse segments, and then evaluate observed against expected mappings for explicit vs. implicit discourse relations separately. This enables us to identify biases of the PDTB and RST schemes, and also gain insights into why automatic discourse relation labelling on implicit relations is so difficult (current state-of-the-art systems achieve an f-score of around 43% for 11-way classification of implicit discourse relations in English (Xue et al. 2016)).

Our article makes the following contributions:
We propose how coherence relations can be mapped onto one another even when segmentation differences are present.

We provide an aligned discourse corpus where both PDTB and RST annotations can be queried simultaneously.

Using contingency tables that express large amounts of annotated data, we show how well the annotations by different schemes correspond to one another.

We identify sources of systematic discrepancies between annotation schemes and discuss their consequences for future annotation, corpus search, and the training of automatic discourse relation labellers.

We identify coherence relations for which human annotation is informative and beneficial, as well as cases for which it is unclear whether manual annotation is sufficiently consistent to be useful.

The remainder of this paper is laid out as follows. First, we briefly summarize the most important aspects of PDTB, RST and the CCR-framework, which is the basis for the theoretical mapping of the PDTB and RST relations, in Section 2 and discuss related work in Section 3. We then proceed to describe the alignment process for the PDTB-RST mapping on the Penn Discourse Treebank in Section 4 and the results of the discourse relation label mapping in Section 5. Finally, we discuss implications for annotation as well as automatic discourse processing in Section 6.

2. Background

In this section, we describe the basic notions of the two frameworks that the annotations on the Wall Street Journal stem from, namely PDTB and RST-DT. This information provides the necessary background for understanding the reasons behind differences in segmentation and discourse relation sense labelling that we find in our study. We also present CCR, the framework underlying the theoretical mapping of PDTB and RST relations. Due to CCR’s representation of coherence relations in terms of their properties, CCR also functions as a useful lens to study the systematic differences and difficulties in discourse relation annotation.

2.1 Rhetorical Structure Theory (RST)

The framework that is used to annotate the RST-DT is based on the Rhetorical Structure Theory as proposed by Mann and Thompson (1988). RST was originally developed for computer-based text generation, and is intended to describe coherence in texts. Annotators are instructed to annotate the writer’s goal or intended effect of each segment of a text with respect to the neighbouring segments and the resulting hierarchical structure of the entire document.

Segmentation in RST. RST annotates full texts, building them up in a discourse tree structure. Each document is first decomposed into non-overlapping sequential text spans (called Elementary Discourse Units). EDUs are generally clauses, but complements of attribution verbs, relative clauses, nominal postmodifiers, and phrases that begin with a strong discourse marker are also considered EDUs (see Carlson and Marcu 2001, p.3). EDUs are connected and assigned a relation label. Example (1-a) illustrates such a re-
lation: the label ELABORATION-ADDITIONAL was chosen to represent the semantic link between the two segments in brackets. Next, these more complex combinations of EDUs are connected to one another by annotating new, higher-level discourse relations, until the entire text is connected and a tree structure is formed, as in Example [1-b]. The tree structure in RST annotations does not allow crossing or embedded arguments. In order to deal with these limitations and the restriction that EDUs can’t overlap, a SAME UNIT tag was introduced which allows annotators to express that an EDU is discontinuous. To illustrate this, consider [1-c]. The clause when implemented is embedded in another clause. As a result, the clause that it will cannot be connected to it’s other half because when implemented cannot be skipped. The tag SAME UNIT can be applied in this situation to express that the segments in fact make up one segment.

(1) a. [But even on the federal bench, specialization is creeping in,] [and it has become a subject of sharp controversy on the newest federal appeals court.] — ELABORATION-ADDITIONAL, wsj_0601
b. [In an age of specialization, the federal judiciary is one of the last bastions of the generalist.] [But even on the federal bench, specialization is creeping in, and it has become a subject of sharp controversy on the newest federal appeals court.] — CONTRAST, wsj_0601
c. ... [that it will,] [when implemented,] [provide significant reduction in the level of debt and debt service owed by Costa Rica.] — SAME-UNIT, wsj_0624

Discourse Structure in RST. RST works from the assumption that relations have at least two arguments, which can be nuclei or satellites. The nucleus is the central part of a text, and the satellite is supportive of the nucleus. Some relations have symmetrically important arguments by definition. These relations consist of two nuclei rather than a nucleus and a satellite. The writer’s intentions are important when assigning the nuclearity (i.e., what does the writer want to achieve?). Determining nuclearity can therefore rarely be done without taking the context of the relation into consideration. Nuclearity assignment is often determined simultaneously with the assignment of a discourse relation.

Carlson and Marcu distinguish 72 relation labels, partitioned into 16 classes that share some type of rhetorical meaning (Table 1 presents a few examples; for a full list, see Carlson, Marcu, and Okurowski 2003). Some of these classes contain relations that are generally not considered to be coherence relations in other schemes such as PDTB; examples include the ATTRIBUTION and QUESTION-ANSWER relations.

| Relation Name  | Nucleus                | Satellite/Other Nucleus                                      |
|----------------|------------------------|-------------------------------------------------------------|
| Preparation    | text to be presented   | text that prepares the reader                               |
| Elaboration    | basic information       | additional information                                      |
| Contrast       | one alternate          | the other alternate                                        |

Table 1
Examples of RST relations
2.2 Penn Discourse Treebank (PDTB) Annotation

PDTB focuses on low-level relations (within or between adjacent sentences), rather than on relations between relations. The framework has a connective-based approach: annotators were instructed to annotate explicit connectives, or insert a connective and annotate an implicit relation if no connective was present.

Segmentation in the PDTB. Relations in the PDTB have two and only two arguments, referred to as Arg1 and Arg2. The position of Arg2 depends on the position of the connective, since Arg2 always attaches to the connective. Annotators were instructed to first identify explicit connectives based on a list of discourse cues; they then identify the discourse relation arguments. The selection of these arguments is restricted by the “Minimality Principle,” according to which only as much content should be included in the argument as is minimally required and sufficient for the interpretation of the relation. Any other span that is considered to be relevant to the interpretation of the argument is annotated as supplementary information. After identifying explicit connectives and their arguments, a relation label is assigned to them, as in Example (2-a). For implicit discourse relation annotation, annotators were instructed to try to insert connectives between any two adjacent sentences that were not connected by a discourse marker in the original text, such as meanwhile in Example (2-b). Hence, PDTB does not annotate minimal argument spans for implicit relations; rather, the full sentences that the implicit relation connects are annotated as arguments.

(2) a.  [I’d like to see Kidder succeed] but [they have to attract good senior bankers who can bring in the business from day one].

— COMPARISON.PRAGMATIC CONTRAST, wsj_0604
Discourse Structure in the PDTB. The framework distinguishes a hierarchical set of relation labels, consisting of three levels: (i) class is the top level, which contains the four major semantic classes; (ii) type is the second level, which further refines the semantics of the class levels; and (iii) subtype is the most fine-grained level, which defines the semantic contribution of each argument. When an annotator was uncertain of the more fine-grained senses of subtype, s/he could choose the higher level type, which was also beneficial for inter-annotator agreement. If no suitable connective and coherence relation was identified for two consecutive sentences, the ENTREL and NOREL labels were used depending on shared entities or none, respectively. The resulting annotations are not always compatible with a tree structure (see Lee et al. 2006). This difference in argument span size is potentially magnified due to PDTB’s Minimality principle: It is possible that some text spans in a text are not part of a discourse relation in PDTB, while RST does not allow such gaps. For implicit relations on the other hand, PDTB does not follow minimal argument spans, but always annotates the two full adjacent sentences that the implicit relation connects, while RST might split these up into several EDUs.

2.3 The Cognitive approach to Coherence Relations

For the theoretical mapping of discourse relation labels, we make use of a proposal by (Sanders et al. Submitted), which has grown out of the COST action TextLink1. PDTB and RST-DT labels are mapped onto each other using an extended version of the dimensions originally proposed in the Cognitive Approach to Coherence relations (CCR, Sanders, Spooren, and Noordman 1992). In this framework, coherence relations are not assigned a single end label, but are described in terms of their characteristics. Original CCR distinguishes four cognitive dimensions that apply to every relation, namely polarity, basic operation, source of coherence, and order of the segments. As PDTB and RST make some distinctions which cannot be represented in terms of only these four dimensions, extended CCR to account for more fine-grained properties of relations. They mapped relation labels from one framework to another by translating them to their respective values on the dimensions, rather than by assigning a one-to-one mapping. The translation method makes it easier to identify similarities and differences between relations (other frameworks describe similarities by arranging relations in hierarchies (e.g., a COMPARISON or CONTINGENCY class).

We will now introduce the dimensions. The polarity of a relation refers to the positive or negative character of a segment. A relation is positive if the propositions P and Q, expressed in the two discourse segments S1 and S2, are linked directly without any contrast or violation of expectations. A relation is negative if the negative counterpart of either P or Q functions in the relation, i.e. if P or Q expresses a contrast with the other argument or a contrast with an expectation raised by the other argument. The basic operation distinguishes between causal and additive relations. A relation is causal if an implication relation (P → Q) can be deduced between the two segments. A relation is

1 http://www.textlink.ii.metu.edu.tr; the proposal has been presented to and discussed with the community on several occasions (e.g., Sanders et al. 2016).
additive if the segments are connected as a conjunction (P & Q). The Source of Coherence distinguishes between objective and subjective relations. Objective relations consist of segments that describe situations that occur in the real world. Subjective relations, on the other hand, express the speaker’s opinion, argument, claim or conclusion. Hence, the author or speaker is actively engaged in the construction of subjective relations. The order of the segments corresponds to the surface order of the segments in causal relations. In a coherence relation with a basic order, the antecedent (P) is S1, followed by the consequent (Q) as S2. In a relation with a non-basic order, P maps onto S2 and Q onto S1. Additional features used for the mapping of RST and PDTB relations are temporality, list, specificity and alternative to distinguish between different types of additive relations, and the features conditionality and goal-orientedness to distinguish between different causal relations (Sanders et al. Submitted). These features can help to account for the more fine-grained differences in relation labels from different frameworks that CCR did not capture. Temporality distinguishes temporal relations from non-temporal relations. The list feature distinguishes relations whose arguments can be listed from non-list relations. Specificity distinguishes relations that are characterized by the specificity of one segment relative to the other segment, such as PDTB’s I NSTANTIATION or R ESTATEMENT relations. Alternative distinguishes additive relations in which the two segments are presented as alternatives (such as D ISJUNCTION relations) from additive relations in which this is not the case. Conditionality distinguishes conditional causal relations from non-conditional causal relations. Finally, goal-orientedness distinguishes relations for which one of the segments concerns an intentional action by an agent, such as RST’s P URPOSE and M EANS relations.

The current study presents a mapping effort for data annotated by both PDTB and RST-DT. Where relevant, we will refer to the “predicted” or “expected” mapping. These predictions are based on the decomposed values of labels according to the unifying dimensions. We will return to this issue in Section 5. First, we discuss related work.

3. Related Work

The challenge of working with a multitude of different discourse relation frameworks and their corresponding resources has been discussed in two main lines of work. One line of research has focused on creating a unified framework. For example, Bunt and Prasad (2016) proposed a set of relations that are central in many frameworks, and Hovy and Maier (1995) taxonomized the more than 400 relations that have been proposed in different frameworks into a hierarchy of roughly 70 discourse relations.

Another line of research has focused on comparing existing discourse annotations of two frameworks. Most similar to the present article is the work described in Rehbein, Scholman, and Demberg (2016). Rehbein et al. created an English corpus of spoken discourse containing PDTB and CCR annotations for every relation. After annotation, it was hence possible to map the relation labels onto one another directly, and verify in this way the proposed mapping from PDTB to CCR relations. Rehbein, Scholman, and Demberg (2016) reported three systematic biases introduced in the operationalizations of PDTB and CCR, which lead to differences in annotations in some areas: A first observation holds that PDTB’s additive relations Expansion.Instantiation, Expansion.Specification and Expansion.Equivalence were quite often (30% of relations) annotated as causals in CCR. It turns out that these types of discourse relations can often be ambiguous, as examples can at the same time also serve as evidence for a claim. For a more detailed discussion, see Scholman and Demberg (2017).
The second category of systematic disagreements concerns \textsc{Comparison.Contrast} and \textsc{Comparison.Concession} relations: among the negative relations, annotators often disagreed on the causal vs. additive basic operation. This was partly due to a slightly different definition of what constitutes a \textsc{Concession}, but note that distinguishing between contrastive and concessive discourse relations is a well-attested difficulty (see, for example, Robaldo and Miltsakaki 2014; Zufferey and Degand 2013). A third pattern of disagreements regarded the positive vs. negative polarity of relations: Rehbein, Scholman, and Demberg (2016) found that some instances marked by \textit{but} were annotated as positive polarity relations in PDTB, but as negative in CCR (this for example includes instances marked with \textit{but also}). This was due to an annotation instruction in CCR: as a rule, all relations that can be marked with \textit{but} are annotated as negative polarity relations. Crucially, this study therefore shows that disagreements in mapped data can only in part be attributed to typical annotator disagreement, a second important source are systematic differences in operationalization of the annotation schemes, i.e. differences in how exactly the annotators are supposed to decide between discourse relations. It is likely that this also holds for the mapping presented in the current study.

Other related work did not investigate the actual correspondence between different frameworks based on double-annotated data, but their efforts are nevertheless relevant to the current study: Scheffler and Stede (2016) set out to map the PDTB 3.0-style and RST-style annotations for explicitly marked relations in the Potsdam Commentary Corpus (PCC) onto one another. Rather than addressing the mapping of the relation labels, they focus on the mapping of structures: mapping explicit connectives to their corresponding label in either framework. Their work therefore represents a first step toward a comparison of the annotations according to the RST and PDTB frameworks for German, and provides useful test data for future use.

Benamara and Taboada (2015) proposed a unified set of 26 discourse relations based on distinctions made by the RST and SDRT frameworks. They then mapped the set to annotations in three corpora (RST-DT, SDRT Annodis and RST-ST), but they did not have any data available that was annotated according to both frameworks. They therefore were not in a position to evaluate whether the actual annotations of the two frameworks would be consistent with one another. Benamara and Taboada (2015)'s work provides interesting insights into how one could go about mapping between frameworks. Additionally, they highlighted differences in granularity between frameworks by identifying certain labels that exist in one framework but do not have a corresponding label in the other framework.

Chiarcos (2014) proposed an ontology to integrate RST, PDTB and OntoNote annotations within a higher-level, more general framework. In this framework, the RST and PDTB labels are assigned new labels with respect to the more general relation senses in both schemata. Chiarcos (2014)'s work provides a promising implementation of computing a mapping. However, this implementation was not applied to actual text data to analyze the correspondence between these frameworks in practice.

The current study extends this previous work by mapping existing PDTB and RST-DT annotations on the same text. This allows us to look at correspondences between relation labels of the two frameworks, and identify systematic differences between the annotations that could be caused by differences in the respective operationalizations and implicit biases of the frameworks. In the next section, we discuss the data that was used for this mapping. Section 4 discusses how the relations were mapped onto one another, focusing on resolving segmentation differences. In Section 5, the results of the mapping is presented, focusing on the correspondences between the relation labels.
4. Data and Automatic alignment

PDTB 2.0 and RST-DT annotations overlap for 385 newspaper articles in sections 6, 11, 13, 19 and 23 of the Wall Street Journal corpus. The annotation of the RST-DT involved more than a dozen of people and several phases of revision. The average inter-annotator agreement (final results for 6 taggers) on span detection, nuclearity assignment and relation sense annotation was 86.8%, 80.7%, and 72%, respectively (Carlson, Marcu, and Okurowski 2003). The PDTB reports an inter-annotator agreement of 94%, 84% and 80% for the class, type and subtype levels respectively, and PDTB’s discourse segments were identified with an agreement (exact string match) of 90.2% for explicit relations and 85.1% for implicit relations (Prasad et al. 2008).

The number of discourse relations that PDTB version of the corpus marks is substantially lower than the overall number of RST relations for the same text. This is due to RST building up complete trees for the whole document, while PDTB only asserts relations when connectives are present, and for adjacent sentences. We therefore used PDTB relations as a starting point in alignment, with the goal of identifying for each PDTB relation the corresponding relation label in the RST annotation.

Discourse relation arguments in the two frameworks can be of very different size: Since RST relations are annotated in a hierarchical tree structure, the relations higher up in the discourse relation tree can have very large argument spans, which only partially overlap with a corresponding PDTB argument. This might be true even if the annotators of both schemes had the same interpretation and same relation in mind. Furthermore, the difference in argument span size is potentially magnified due to the PDTB annotation instructions of marking only “minimal” argument spans for the annotation of explicit relations. Hence, there can be text spans in PDTB that are not part of any discourse relation, while RST does not allow such gaps. Additionally, relational segments in RST can also sometimes be smaller than those in PDTB due to different segmentation principles: PDTB annotators as a rule marked the complete sentences they were connecting when annotating implicit relations, while RST uses the same principle for identifying elementary discourse units in explicit and implicit relations.

Our procedure for determining the optimal mapping of RST relations for every PDTB relation involves two major steps:

1. Identifying for every PDTB discourse relation those RST segments (EDUs) that best correspond to the PDTB segments (Arg1 and Arg2): Given a PDTB argument, we first iterated over all RST segments in the source file and selected the one with maximum overlap (common characters) and minimum margin (extra characters). Overlap and margin were calculated with respect to the character offsets from the file onset (begin and end of the annotation). We then tried to determine whether a PDTB argument should be aligned to more than a single RST segment by iterating over all RST relation annotations (sub-trees rather than single segments) using the same criteria. Having the best candidates for both arguments of the PDTB relation (let’s call them Arg1-equivalent and Arg2-equivalent RST spans), we moved to the next step.

2 These numbers are the result of averaging over the inter-annotator agreement scores reported for every two annotators in Table 2 of Carlson, Marcu, and Okurowski (2003).
3 The alignments will be made available.
2. **Identifying the RST relation label that describes the relation between the Arg1-equivalent and Arg2-equivalent spans:**

The aim of this step was to find the lowest RST relation within the text tree that connects the two RST spans obtained in the previous step.

During the procedure described above, several relations were flagged as instances for which the mapping was potentially problematic, based on two criteria. First, relations for which a PDTB argument contained larger text spans (e.g., intervening RST segments) or a complicated structure (overlapping/embedded arguments) were flagged as suspicious. Second, relations that did not adhere to the Strong Nuclearity hypothesis (Marcu 2000) were flagged as suspicious. The Strong Nuclearity hypothesis, which was used for RST-DT annotations, states that when a relation is postulated to hold between two spans of text, it should also hold between the nuclei of these two spans. For the analyses presented in this paper, we only focus on those relations that can be mapped onto one another with high confidence, i.e. that were not flagged as potentially problematic mappings. Finally, given that some relations carried two PDTB relation labels (in case the annotators thought that both relations held), we chose the relation that was most similar to the corresponding RST relation label.

### 4.1 Successful alignments

In total, we were able to successfully and confidently map 74% of PDTB relations from the joint corpus to corresponding RST relations (a total of 4987 relations). 53% of these relations have directly corresponding spans, for which the arguments map directly onto one another and are not complex (although spans may differ slightly). To illustrate this, consider Figure 2 which presents the PDTB (left) and RST (right) annotations for a fragment of a Wall Street Journal article. In PDTB, segments (a-b) and (c) are connected in a **TEMPORAL.SYNCHRONY** relation. In RST, a similar **TEMPORAL-SAME-TIME** relation is annotated between segments (b) and (c). Even though the spans differ slightly, they do map onto each other and none of the arguments contains multiple relations.

The remaining 47% of the successfully mapped data consists of relations for which the RST tree is more complex than the PDTB relation. In other words, at least one of the PDTB arguments maps onto an RST argument that consists of multiple RST relations. In Figure 2 this is the case for the **RESTATEMENT** relations: in PDTB, this relation holds between segments (a) and (d), whereas in RST, this relation holds between segments (a-c) and (d). The segments (a-c) contain other relations; namely **TEMPORAL-SAME-TIME** and **ATTRIBUTION**. Even though the RST nucleus is more complex than PDTB’s Arg1, the relation still adheres to the Strong Nuclearity hypothesis: the nucleus of the RST **RESTATEMENT** relation can be traced to segment (a), which corresponds to Arg1 of the PDTB **RESTATEMENT** relation.

### 4.2 Data that was excluded because of alignment

26% (1714 instances) of PDTB relations could not be mapped with high confidence to corresponding RST relations. That is, these relations were in fact mapped, but there is a higher chance that the labels do not apply to the same relation. They therefore need to be checked manually to ensure that the alignment is correct before they can be included in the analyses. These instances are relations for which one of the RST arguments is larger than the corresponding PDTB argument, or one of the PDTB arguments spans over the two RST arguments. The labels therefore possibly do not apply to the same relation. To
illustrate this, consider the the passage in Figure 3. In this example, the PDTB relation CONTRAST is mapped onto RST’s CONSEQUENCE because the segments are somewhat similar, even though PDTB’s segments comprise a smaller span (PDTB excludes (a) and (e) from the relation). However, this difference in span size affects the interpretation of the relation. In PDTB’s annotation, the relation holds between the state’s action and the farmers’ actions. RST’s annotation, on the other hand, holds between the state’s action and the consequences of that action. The labels therefore do not map onto the same relation. The Strong Nuclearity Hypothesis is also violated in this example: the nucleus of the satellite cannot be traced to PDTB’s Arg2, since RST’s CONTRAST relation is multinuclear.

However, not all of these instances that are not included in the following analyses are in fact incorrectly aligned. Consider the passage in Figure 4. The PDTB relation TEMPORAL SYNCHRONY was marked as possibly problematic, because it was not certain that the RST relation CIRCUMSTANCE correctly maps onto this relation; i.e., that the two relations are marked for the same arguments. However, looking at the tree it seems that the mapping is in fact correct. The uncertainty comes from the two intervening multinuclear relations, LIST and CONSEQUENCE. Due to these multinuclear relations, it is not clear what the nucleus of the satellite of CIRCUMSTANCE is. That is, the nucleus path cannot be traced to segment (c-d), which corresponds to PDTB’s Arg2. When interpreting the relation, it seems however that the PDTB and RST relation labels correspond to each other. Relations such as these would have to be filtered out by manual correction, which is outside the scope of the present article.

Finally, two types of PDTB labels were excluded from the following results, namely ENTREL and NOREL. ENTREL is used to mark cohesion when no specific coherence
relation can be identified. The label *NOREL* is used when there is neither an identifiable coherence relation nor cohesion in terms of shared entities between two adjacent sentences. Even though these labels were excluded from analysis, the alignment did uncover interesting patterns. First, the alignment showed that *ENTREL* “relations” are often annotated as RST *ELABORATION-ADDITIONAL*, which seems to be a reasonably consistent mapping. This results suggests that in practical terms, it might be viable to nevertheless include a mapping for *ENTREL*.

Second, many of PDTB’s *NOREL* instances were excluded by our criteria for good-quality mappings. This reflects that the exclusion procedure was successful, since *NOREL* instances are not considered relations by PDTB. The mapping tells us that in fact RST annotators agree that there is no direct relation between these two sentences, but that instead at least one of these sentences is part of another more complex relation of which it is not the nucleus. Hence annotations for *NOREL* that were excluded from analysis can actually be considered to be consistent with RST annotation.

5. Correspondence between mapped relation labels

Our analysis of correspondences between mapped labels is based on a total of 4987 PDTB labels that could be mapped with high confidence. These results will be represented in tables with lighter and darker colours, also known as heat maps. The colours represent the percentage agreement: darker colours mean higher agreement between the two frameworks for a specific label. To keep tables readable, we only include those labels that occurred more than ten times in the data. We first discuss explicit and then implicit discourse relations. We also consider both mapping directions separately, i.e. mapping PDTB labels onto RST labels, vs. mapping RST labels onto PDTB labels (the results are different due to different granularities in distinctions between discourse relation classes). With this mapping, we can take stock of whether our hypothesized correspondences are consistent with the actual annotations, and can identify areas which might need additional careful consideration.

Throughout this section, we will refer to the “expected” mapping. This expectation is based on the mapping proposed in Sanders et al. (Submitted). Note that a one-to-one mapping is not always possible, as some labels have multiple matching candidate labels in another framework. This cannot always be avoided, since the granularity
Table 2
Alignment of explicit discourse relation classes (PDTB level 2 instead of level 3) for which \( n > 30 \).
Numbers indicate how many instances occurred in our high-confidence mapping. Underlined, bold numbers indicate the predicted mapping. Colours encode percentage agreement from PDTB perspective: darker colours show that most instances of a PDTB relation type occurred in that specific RST class.

5.1 Mapping PDTB annotations onto RST annotations for explicitly marked relations

Table 2 displays the mapping of PDTB annotations onto RST annotations for explicit discourse relations. The colours in the table indicate the percentage of correspondence to RST labels, with darker shades indicating a higher proportion of instances with a certain PDTB label falling into that RST category. For example, 34% of PDTB TEMPORAL.SYNCHRONOUS relations are annotated as RST TEMPORAL-SAME-TIME, and its count of 61 instances is hence shaded in a light green tone. Expected correspondences in the table are indicated by underlined numbers and follow roughly the diagonal of the table.

When looking at mapping the explicitly marked relations, we can distinguish a number of clusters, which correspond quite well to our expected mappings. In this section, we will discuss the mapping for every PDTB class separately, starting with temporal relations.

The results show that most (82%) of the explicitly marked relations that were classified as SYNCHRONOUS by PDTB were tagged as RST TEMPORAL-SAME-TIME or
CIRCUMSTANCE, which is consistent with definitions in the annotation manual. There are however also some cases where annotations deviated from expected mappings for temporal PDTB labels. These include cases where one of RST’s causal labels (specifically, EXPLANATION-ARGUMENTATIVE, or CONSEQUENCE) was annotated. Closer inspection revealed that frequent connectives in these relations, which did not get a temporal sense label in RST, were as and when. These connectives, which are known to be ambiguous markers (see e.g. [Asr and Demberg 2013], are also frequent among temporal relations such as CIRCUMSTANCE and TEMPORAL-SAME-TIME in RST. We hence find that there are some instances of these ambiguous connectives which could not be consistently disambiguated between frameworks through manual annotation, with one framework labelling these instances as temporal and the other as causal. Temporal ASYNCHRONOUS relations generally also map well to their corresponding RST classes (78%). The most notable unexpected pattern consists of PDTB temporal relations marked with until often being classified as Condition in RST. This mismatch could be indicative of inconsistencies in disambiguation of this marker, or could be more systematically related to RST annotating the intention in subjective relations, while PDTB annotations stay closer to the semantic relation (Scholman and Demberg 2017).

Causal and conditional PDTB relation labels generally map well onto causal and conditional RST labels well (> 82% agreement), as these relations are usually strongly marked. RST distinguishes more types of causal relations, which results in causal PDTB relations being distributed among the various causal RST classes. Unexpected mappings (PDTB causals annotated as RST’s CIRCUMSTANCE (14%) and PDTB conditionals annotated as RST CIRCUMSTANCE relations (13%)) were found to occur again with the ambiguous connectives as and when, respectively.

Next, we take a look at PDTB’s COMPARISON relations. For PDTB’s CONTRAST relations, we find that the majority was mapped to RST’s CONTRAST and ANTITHESIS relations (62%). ANTITHESIS is defined such that it contains both contrastive and concessive relations, so both correspondences are marked as expected. A substantial portion (20%) of PDTB’s CONTRAST relations are annotated as RST’s CONCESSION; and some also as RST’s ELABORATION-ADDITIONAL (6%). We found that these cases were often marked by the connective but. Relations annotated as PDTB’s CONCESSION EXPECTATION map quite well (54%) onto RST’s CONCESSION relations, while CONCESSION CONTRA-EXPECTATION relations are often annotated as CONTRAST in RST, especially when marked with the connective but. The distinction between concession and contrast relations is known to be difficult in discourse relation annotation. It is possible that the observed differences stem from differences in interpretation between annotators, and slight biases in the frameworks to prefer the one label over the other in ambiguous cases.

Finally, looking at PDTB’s EXPANSION relations, we find that a majority of relations annotated as PDTB’s CONJUNCTION is annotated as RST’s LIST (57%), which was not expected based on the theoretical definitions of these relations. Closer inspection shows that the high number of cases annotated as RST LIST stems from the fact that PDTB annotation guidelines say that lists have to be “announced”. Unannounced lists cannot be annotated as LIST relations in PDTB. We also observe a substantial amount of noise, i.e. 26% of CONJUNCTION relations have a temporal, causal or contrastive label in RST; some of these cases probably occur due to the presence of explicit markers such as but or while. We find that PDTB’s LIST and INSTANTIATION relations map well onto RST’s LIST and EXAMPLE relations respectively.

To summarize, we can say that deviations from expected mappings are mostly related to ambiguous connectives (e.g., as, when, but, while) which are in some instances
resolved differently in the two frameworks, as well as differences in operationalization between frameworks which leads to mismatches in annotations for relations such as LIST. We will next view the table from the RST perspective.

### 5.2 Mapping RST annotations onto PDTB annotations for explicitly marked relations

Table 3 shows the same data as Table 2 but with the colours indicating the percentage of correspondences relative to the expected PDTB class. We will here only point out findings which were not obvious when looking at the table from the PDTB perspective.

For RST’s SEQUENCE class, we find that a substantial portion is annotated as PDTB’s CONJUNCTION (36%), i.e. annotations by the different frameworks disagree in these instances as to whether the relation is temporal or not. The RST relation CIRCUMSTANCE is a frequent label in RST, but it does not have a directly corresponding label in PDTB. A large variety of PDTB labels were assigned to RST’s CIRCUMSTANCE relations. For example, while most instances of CIRCUMSTANCE are labelled as temporal, we also find them annotated as CONJUNCTION, CONDITION, REASON or CONTRAST relations in PDTB.

PDTB’s classification of RST’s COMPARISON relations is interesting, because the RST annotation manual explicitly defines these relations as non-contrastive. The results show that the majority of these instances (63%) were annotated as CONTRAST in PDTB. The markers that were present in these relations are *while*, *but* and *however*. In the mapping view with RST as the source framework, it also becomes obvious that PDTB
has a stronger bias than RST towards assigning the CONTRAST label, as this is the predominant label for three other RST relations, namely CONTRAST, ANTITHESIS and CONCESSION.

Final interesting observations regards the annotation of the connective unless: these instances are annotated as CONDITION in RST but as ALTERNATIVE.DISJUNCTIVE relations in PDTB (note that RST does not have a corresponding label).

Overall, we can conclude that a large number of unexpected annotations happen in case that one of the frameworks does not provide a corresponding label.

5.3 Analysis of annotation for ambiguous connectives

Given the disagreements which can also sometimes be related to different interpretations of specific ambiguous connectives as discussed in sections 5.1, we studied the agreement on annotating ambiguous connectives in more detail. After all, it’s not surprising to reach high agreement between frameworks for labelling relations with an unambiguous connective like if as CONDITIONAL; this could be done automatically. Rather, the value of manual annotation comes through adding additional information by disambiguating between relations when the marker is ambiguous (or when a relation is not explicitly marked, see below section 5.4).

To this end, we tested for each connective, whether annotations by the two schemes were agreeing with each other significantly more than would be expected from a baseline of randomly assigning labels to the connectives given the proportions of their occurrence in the data. Note that this analysis is more strict than usual kappa for inter-annotator agreement, because we here put in the distribution of relations per connective, i.e. which relations a connective can mark, and how often it does so for each relation type in the text at hand, which we do not normally know. This analysis can tell us whether actual annotations of the relation instances were successfully taking into account the content of the discourse relation arguments, or whether the distinction is potentially too difficult or too subtle for human annotators to make reliably.

We tested the independence of RST and PDTB annotations by calculating a separate $\chi^2$ test for each connective, where large numbers of observations were present in our corpus, and Fisher’s exact test for those connectives where assumptions of the $\chi^2$ test (less than 5 expected observations in a cell) were violated. Here, we will report some representative results for connectives where the null hypothesis of RST and PDTB annotations for a connective being independent could be rejected with high confidence vs. cases where observed distributions were very similar to expected random distributions. The connective while is an example of the first case: we find that temporal synchronous vs. contrastive / comparison readings could be distinguished very well, the annotations from the two frameworks almost always agreed on this ($p < 0.0001$).

On the other hand, the distribution for sense labels of connectives which are ambiguous between more similar discourse relations (contrast vs. concession), such as but, although and however, did not significantly differ from a random distribution of these sense labels (given the marginals), according to a $\chi^2$ test for but and fisher’s exact tests for although and however. If we calculated $\kappa$ values in this strict reading (i.e. taking for granted that but cannot mark causals or temporals and only testing agreement on different subtypes of negative relations, this corresponds to $\kappa$ values < 0.1 for these relation label distinctions between frameworks.

To summarize, we find that there are only some connectives which are ambiguous between very different relation types for which manual annotation from two different frameworks actually agrees on how to annotate the readings. For more subtle distinc-
tions, we cannot find conclusive evidence that annotations from the two frameworks agree with one another more than would be expected by chance (given the distribution of these connectives). This lack of agreement between humans may also provide a partial explanation for why automatic discourse relation disambiguation is difficult – it is unclear whether the training data that these distinctions are trained on is fully consistent internally, and/or the distinction may be so subtle in a substantial number of real data cases that even humans find it hard to agree. We will next move on to the analysis of agreement between frameworks on implicit discourse relation annotation.

5.4 Mapping PDTB annotations onto RST annotations for implicit relations

Mapping of relation annotations for implicit relations, as seen from the PDTB perspective, is shown in Table 4. A first striking observation is the agreement between frameworks is a lot worse than for explicit relations (in the graph, there is no diagonal green line of corresponding relation clusters similar to those that could be identified for explicitly marked discourse relations). Instead, we find that a substantial proportion of instances from almost all PDTB classes were annotated as RST’s ELABORATION-ADDITIONAL. An exception is PDTB’s LIST relation, for which 85% of instances were annotated as RST’s LIST relation.

For PDTB’s CAUSE.REASON relations, less than 40% of instances were annotated as one of the expected causal classes by RST annotators (EXPLANATION-ARGUMENTATIVE, REASON, EVIDENCE or INTERPRETATION). For PDTB CONTRAST relations, we find that the majority of these instances were not annotated as CONTRAST (17%) or COMPARISON (10%) relations by RST annotators, as would have been expected based on the expected mapping of relation senses.

Temporals were also not consistently identified between frameworks, with most PDTB TEMPORAL. ASYNCHRONOUS relations being annotated as ELABORATION-ADDITIONAL in RST. A more fine-grained analysis shows that roughly half of PDTB’s ASYNCHRONOUS.PRECEDENCE relations were annotated as SEQUENCE by RST annotators, while PDTB ASYNCHRONOUS.SUCCESSION relations were annotated as ELABORATION-ADDITIONAL in RST.

These results raise the question of how these very substantial differences in annotations of implicit relations can be explained. We think that the discrepancy can be attributed largely to the differences in annotation guidelines and operationalizations for implicit discourse relations. PDTB’s connective-driven approach biases against annotating simple additive relations when a connective can be inserted and hence an additional stronger interpretation of the discourse relation is available. RST prescribes a different strategy: annotators are asked to annotate the writer’s intentions. The resulting low agreement between PDTB and RST-DT on implicit relations have implications for the reliability and validity of these annotations. We will expand on this point in the Discussion. Next, we will now take a look at the implicit relation labels from the perspective of RST.

---

4 The dataset contains 24 instances of implicit RST SEQUENCE relations, therefore this label is not included in the table.
Table 4
Alignment of implicit discourse relation classes for which \( n > 30 \). Numbers indicate how many instances occurred in our high-confidence mapping. Colours encode percentage agreement from PDTB perspective, i.e. darker colours show that most instances of a PDTB relation type occurred in that specific RST class.

5.5 Mapping RST annotations onto PDTB annotations for implicit relations

Table 5 shows the alignment of implicit relations from the RST perspective. Many cells are coloured green, which indicates that the annotations for many of the RST relations have a wide variety of PDTB labels. For RST’s BACKGROUND, CIRCUMSTANCE and COMMENT relations in particular, the PDTB labels are almost equally distributed among the major classes of relations. This may indicate that these RST labels classify relations according to communicative functions which are not represented in the PDTB annotation scheme.

RST’s causal relations CONSEQUENCE, REASON and RESULT map relatively well onto PDTB’s causal relations. However, other causal RST labels (EVIDENCE and EXPLANATION-ARGUMENTATIVE) are often mapped onto the additive PDTB labels INSTANTIATION and SPECIFICATION. This difference can be attributed to a fundamental difference between the approaches: PDTB annotates the lower-level ideational relations between arguments, while RST focuses more on the intentional level. Scholman and Demberg (2017) show that often two functions can be identified in these relations: a segment can provide an example or a specification of a set, as well as providing evidence for a previously stated claim. These double functions are reflected in the annotation mapping between PDTB and RST.

5 The dataset contains 28 instances of implicit RST REASON relations.
6 The dataset contains 23 instances of implicit RST RESULT relations.
Table 5
Alignment of implicit discourse relation classes for which \( n > 30 \). Numbers indicate how many instances occurred in our high-confidence mapping. Colours encode percentage agreement from RST perspective, i.e. darker colours show that most instances of a RST relation type occurred in that specific PDTB class.

We also can see that, as predicted according to the expected mapping, RST’s ELABORATION-GENERAL-SPECIFIC relations are often labelled as PDTB’s RESTATEMENT (55%). 21% of instances were labelled as INSTANTIATION, but note that the difference between these labels is rather subtle. For RST’s LIST relation, we see that the largest proportion of its instances (44%) are annotated as CONJUNCTION in PDTB; as mentioned earlier, this problem is partially due to the guideline in PDTB that lists have to be announced.

For RST’s CONTRAST relations, we observe that they were for the most part annotated as contrastive relations in PDTB, usually using the underspecified CONTRAST label (rather than one of its subtypes).

Our analysis of the implicit relations from both perspectives documents a disappointingly low level of agreement between labels assigned by RST and PDTB. Generally, a stronger relation (biasing away from annotating simple additive labels) tended to be chosen by PDTB annotators, which may be a direct consequence of the connective insertion strategy. Nevertheless, wherever the RST annotators chose a label other than ELABORATION-ADDITIONAL (the predominant label assigned to the implicit cases), the relations decently matched with their PDTB equivalents.

6. Discussion

We proposed an automatic mapping algorithm for PDTB and RST-DT discourse relation annotations for a segment of the WSJ corpus that contains annotations from both frameworks. As a result, we are able to offer a more complete picture of how annotations from the two frameworks relate to one another in practice. We compared actual annotations
to expected correspondences which were determined based on a recent proposal for mapping discourse relations onto one another.

Our most striking observations were a lower than expected level of agreement on annotations for implicit relations, and little agreement on more fine-grained distinctions even for explicitly marked relations. We have been able to identify some patterns that lead to these observed disagreements: many of the differences in annotation can be traced back to different operationalizations and goals of the PDTB and RST frameworks. We propose that some segments can stand in more than one relation to one another, i.e. a segment can be an example for something that was said in the other segment, but it may also at the same time serve as evidence for a claim, see also Scholman and Demberg (2017). The former type of relation is referred to as an ideational relation, which involves the relation between the information conveyed in the consecutive elements of a coherent discourse (cf. Moore and Pollack 1992), whereas the latter type of relation is known as an intentional relation, in which the writer attempts to affect the addressee’s beliefs, attitudes, desires etc. by means of language (cf. Hovy and Maier 1995). This view is conform with other work that separates ideational (i.e. semantic) from intentional levels of discourse relations, such as Crible and Degand (in press); Hovy and Maier (1995); Moore and Pollack (1992) and Redeker (1990). We believe that these functions are orthogonal and should therefore be separated in the annotation process. Of course, not all coherence relations necessarily have an intentional function. In order to deal with this, annotators can be asked to annotate whether there is an intentional relation or not.

We furthermore found high numbers of disagreement between observed and expected annotations for those relations that did not have a direct correspondence in the other scheme. Examples of such relations include RST’s BACKGROUND, CIRCUMSTANCE or COMMENT relations. While it was possible to suggest an expected corresponding mapping for those relations (temporal or conjunctive readings) based on the definitions in the annotation manual, actual mapping showed that many of these relations were also annotated as causals or even contrastives in PDTB. We see two possible explanations for why this happens: either the mapping scheme (and possibly the annotation manuals) would have to be revised to more clearly or exhaustively describe the relations, or there is a function to the relation which is not reflected in the PDTB scheme, and should be considered to be added to relation schemes, again possibly annotating both the ideational and intentional functions separately.

A third observation is that the operationalization in the PDTB framework for implicit relations (first annotate a discourse connective that would fit the relation, and then in a second step annotate the relation sense) encourages annotators to assign more specific relation labels than RST’s annotation procedure does. We therefore find that most implicit relations receive the RST label ELABORATION-ADDITIONAL but a more specific PDTB label. For future annotation efforts, it is important that this consequence of annotation operationalization is taken into account.

The low agreement on implicit relations also raises questions about the validity of discourse relation annotation, given that there is a well-defined correspondence of relations as seen for explicits. This difference in agreement between explicit and implicit relations might also be due to the multiple readings of a single relation. If one assumes that the two levels of annotation (i.e., ideational and intentional) indeed hold but only

---

7 The ideational and intentional levels are also referred to as informational vs. intentional (Moore and Pollack 1992), subject matter vs. presentational (Mann and Thompson 1987), and internal vs. interpersonal (Hovy and Maier 1995).
one level is annotated, it can be argued that these levels are in competition with each other. If a connective is present, annotators can then focus on that signal and annotate that level. For example, if a relation is marked by *for instance*, annotators can focus on the ideational level and ignore the other function. If no connective is present, annotators cannot rely on such a strong cue and the two levels are in competition again. Separating the two levels during discourse annotation can therefore possibly benefit agreement on implicit relations especially. In order to get more insight into this issue of difference in agreement, we also recommend that future annotation efforts (corpus annotation as well as other tasks) report agreement on implicit and explicit relations separately.

The mapped annotations will be made available online so that other researchers can profit from it. We see several possible directions of research for which this mapped data can be useful. First, for theoretical studies, the data can serve to further investigate the frameworks and the effects of their operationalizations on the annotations. Especially the mismatches are interesting from this viewpoint. Some of the mismatches that occur between the PDTB and RST-DT annotations are systematic; for example, certain causal labels in RST-DT are often annotated as additive labels in PDTB, and RST’s CONTRAST is often annotated in PDTB as CONCESSION. The mapping reveals these patterns and can therefore function as a starting point for other experiments that investigate these systematic mismatches. For example, Scholman and Demberg (2017) investigated the interpretations of INSTANTIATION and SPECIFICATION relations based on the findings in the current article and Rehbein, Scholman, and Demberg (2016).

Second, the mapping can prove to be useful for future annotations. The patterns of matches and mismatches that can be observed in the data can function as input for discussion when annotating new data. The mapped data reveals which relation types may be particularly relevant for carrying several functions, and hence displayed less agreement between frameworks. The labels and definitions agreed upon across frameworks can be considered well-established, but for other types of relations, our mapping indicates that definitions may need to be refined in future efforts (for example, PDTB’s and RST’s CONTRAST and CONCESSION, but also RST’s COMPARISON deserves more consideration).

Third, the mapped data can contribute towards automated discourse parsing efforts. Discourse relation annotations have been used as training data in all recent efforts in automatic discourse relation classification, with specific attention being given to implicit discourse relation classification, given that classification of explicit relations was found to be relatively easy and accurate (Pittler et al. 2008). Implicit discourse relation classification has been recently also the subject of two CoNLL shared tasks (Xue et al. 2015, 2016), with accuracies just over 40% F-score on implicit relation sense labelling for an 11-way classification. Important questions to be considered in the light of the mapping results in this article relate to how these classification results can be interpreted in the light of the difficulty of the implicit relation classification task. How can we make sure that consistency is improved for training automatic discourse relation classifiers? Can and should we train classifiers separately for ideational vs. intentional discourse relation levels? Should classifiers be evaluated by taking into account several possible labels for a relation, so that either the PDTB label or the corresponding RST label would be considered correct? Or should we weigh differently classification mismatch for categories that humans don’t commonly replace for one another versus those that are more interchangeable?

Finally, our study also highlights the importance of discourse segmentation, which has a strong effect on determining the scope and argument structure of a discourse...
relation. Using heuristics, we were only able to align part of the total number of annotations in the corpus (74%). The remaining annotations failed to yield reliable mappings because annotators indicated different clauses as the core arguments of a discourse relation (indicated in PDTB through the argument annotation, and in RST via the nuclearity principle). In future work, these cases could be manually checked to determine whether relation labels for these cases do correspond to one another, or whether the interpretation of the annotators differs more fundamentally. We expect that agreement will be worse on these cases than on the safer cases which we were able to map automatically. The differences in segmentation may hold interesting insights about effects of operationalization of discourse segmentation on discourse annotation, which could be explored to refine annotation processes both for manual annotation and automatic processing.

References
Asr, Fatemeh Torabi and Vera Demberg. 2013. On the information conveyed by discourse markers. In Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics, pages 84–93.
Asr, Fatemeh Torabi and Vera Demberg. 2015. Uniform information density at the level of discourse relations: Negation markers and discourse connective omission. In Proceedings of the International Conference on Computation Semantics, pages 118–128.
Benamara, Farah and Maite Taboada. 2015. Mapping different rhetorical relation annotations: A proposal. In Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics,*SEM, pages 147–152.
Bunt, Harry and Rashmi Prasad. 2016. Iso-dr-core (iso 24617-8): Core concepts for the annotation of discourse relations. In Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12), pages 45–54.
Carlson, Lynn and Daniel Marcu. 2001. Discourse tagging reference manual.
Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Current and new directions in discourse and dialogue. Springer, pages 85–112.
Chiarcos, Christian. 2014. Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), pages 4569–4577.
Crible, Ludivine and Liesbeth Degand. in press. Reliability vs. granularity in discourse annotation: What is the trade-off? Corpus Linguistics and Linguistic Theory.
Hovy, Eduard H and Elisabeth Maier. 1995. Parsimonious or profligate: How many and which discourse structure relations. Unpublished manuscript.
Jansen, Peter, Mihai Surdeanu, and Peter Clark. 2014. Discourse complements lexical semantics for non-factoid answer reranking. In ACL (1), pages 977–986.
Lee, Alan, Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, and Bonnie Webber. 2006. Complexity of dependencies in discourse: Are dependencies in discourse more complex than in syntax? In Proceedings of the 5th International Workshop on Treebanks and Linguistic Theories (TLT), Prague, Czech Republic, pages 79–90.
Mann, William C and Sandra A Thompson. 1987. Rhetorical Structure Theory: A theory of text organization. University of Southern California, Information Sciences Institute.
Mann, William C and Sandra A Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. Text-Interdisciplinary Journal for the Study of Discourse, 8(3):243–281.
Marcu, Daniel. 2000. The theory and practice of discourse parsing and summarization. MIT press.
Meyer, Thomas and Andrei Popescu-Belis. 2012. Using sense-labeled discourse connectives for statistical machine translation. In Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), pages 129–138, Association for Computational Linguistics.
Moore, Johanna D and Martha E Pollack. 1992. A problem for RST: The need for multi-level discourse analysis. Computational Linguistics, 18(4):537–544.
Demberg, Asr, and Scholman Insights from mapping RST-DT and PDTB annotations

Oza, Umangi, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma, and Aravind Joshi. 2009. The hindi discourse relation bank. In Proceedings of the third Linguistic Annotation Workshop (LAW), pages 158–161, Association for Computational Linguistics.

Pitler, Emily, Midhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K Joshi. 2008. Easily identifiable discourse relations. Technical report.

Popescu-Belis, Andrei. 2016. manual and automatic labeling of discourse connectives for machine translation. In TextLink–Structuring Discourse in Multilingual Europe Second Action Conference Károli Gáspár University of the Reformed Church in Hungary Budapest, 11–14 April, 2016, page 16.

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

Redeker, Gisela. 1990. Ideational and pragmatic markers of discourse structure. Journal of Pragmatics, 14(3):367–381.

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

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

Sanders, Ted J. M., Vera Demberg, Jet Hoek, Merel C. J. Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermuel. Submitted. Unifying dimensions in discourse relations: How various annotation frameworks are related. Corpus Linguistics and Linguistic Theory.

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

Sanders, Ted JM, Vera Demberg, Jet Hoek, Merel CJ Scholman, Sandrine Zufferey, and Jacqueline Evers-Vermuel. 2016. How can we relate various annotation schemes? unifying dimensions in discourse relations. In TextLink Second Action Conference, pages 110–112.

Scheffler, Tatjana and Manfred Stede. 2014. Potsdam Commentary Corpus 2.0: Annotation for discourse research. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), pages 925–929.

Xue, Nianwen, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. In CoNLL Shared Task, pages 1–16.

Xue, Nianwen, Hwee Tou Ng, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. Conll 2016 shared task on multilingual shallow discourse parsing. Proceedings of the CoNLL-16 shared task, pages 1–19.

Zhou, Lanjun, Binyang Li, Wei Gao, Zhongyu Wei, and Kam-Fai Wong. 2011. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 162–171, Association for Computational Linguistics.

Zirn, Cäcilia, Mathias Niepert, Heiner Stuckenschmidt, and Michael Strube. 2011. Fine-grained sentiment analysis with structural features. In IJCNLP, pages 336–344.

Zufferey, Sandrine and Liesbeth Degand. 2013. Annotating the meaning of discourse connectives in multilingual corpora. Corpus Linguistics and Linguistic Theory, 1:1–24.
Zufferey, Sandrine and Pascal M Gygax. 2016. The role of perspective shifts for processing and translating discourse relations. *Discourse Processes*, 53(7):532–555.