Dialogue & Discourse 10(1) (2019) 87–135 doi: 10.5087/dad.2019.104

How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations

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

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

Fatemeh Torabi Asr FTORABIA@SFU.CA
Department of Linguistics,
Simon Fraser University

Editor: Barbara Di Eugenio

Submitted 07/18; Accepted 05/19; Published online 06/19

Abstract

Discourse-annotated corpora are an important resource for the community, but they are often annotated according to different frameworks. This makes joint usage of the annotations difficult, preventing researchers from searching the corpora in a unified way, or using all annotated data jointly to train computational systems. Several theoretical proposals have recently been made for mapping the relational labels of different frameworks to each other, but these proposals have so far not been validated against existing annotations. The two largest discourse relation annotated resources, the Penn Discourse Treebank and the Rhetorical Structure Theory Discourse Treebank, have however been annotated on the same texts, allowing for a direct comparison of the annotation layers. We propose a method for automatically aligning the discourse segments, and then evaluate existing mapping proposals by comparing the empirically observed against the proposed mappings. Our analysis highlights the influence of segmentation on subsequent discourse relation labelling, and shows that while agreement between frameworks is reasonable for explicit relations, agreement on implicit relations is low. We identify several sources of systematic discrepancies between the two annotation schemes and discuss consequences for future annotation and for usage of the existing resources.

Keywords: Coherence relations, discourse annotation, mapping,

1. Introduction

Several sizeable discourse annotated resources have been created – most notably the Penn Discourse Treebank (PDTB; Prasad et al., 2008) and the RST Treebank (RST-DT; Carlson et al., 2003) for English – and more discourse annotation projects are currently under way in various languages (see, for example, Oza et al., 2009; Stede & Neumann, 2014a). However, there is as of yet no consensus on a common discourse relation labelling scheme. Existing discourse frameworks share

©2019 Vera Demberg, Merel C.J. Scholman and Fatemeh Torabi Asr
This is an open-access article distributed under the terms of a Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

basic notions of what a coherence relation is, and many of them make relation sense distinctions that are based on similar underlying ideas, but frameworks differ in how they define discourse relational arguments, in terms of constraints on resulting discourse structure (e.g., whether it has to be a tree), and in whether they are "lexically grounded" (like PDTB-style annotations), semantically driven (like SDRT), or contain a combination of semantic and intentional relations (RST-DT). This makes it difficult to study discourse relations across resources annotated according to different schemes, or across languages. In automatic discourse relation classifiers, it also limits the extent to which all available resources can be used effectively for training classifiers. This situation has long been recognized as a problem: in the early nineties, Hovy & Maier (1995) taxonomized the more than 400 relations that have been proposed in different frameworks as a hierarchy of roughly 70 discourse relations.

More recently, several large initiatives have addressed this issue as well: the COST initiative TextLink[1] was aimed at organizing the properties of discourse relations and encouraging the use of a single taxonomy for subsequent discourse annotation, as well as for searches in existing corpora. In this context, some concrete proposals have been made for how discourse relations may be mapped onto one another (Bunt & Prasad, 2016; Benamara & Taboada, 2015; Chiarcos, 2014; Sanders et al., 2018). Bunt & Prasad (2016) developed an ISO standard for coherence relations, in which they propose a new set of coherence relations that are central in many frameworks, and relate each of these to existing labels in other frameworks. The proposed set of labels can also be used for mapping labels to each other. In a similar line of work, Benamara & Taboada (2015) proposed a unified set of 26 discourse relations based on distinctions made by several styles of the RST and SDRT frameworks. They compared discourse relation labels between the frameworks based on their definitions and their overall frequencies of occurrence, but they did not have any data available that was annotated according to both frameworks. They therefore did not evaluate whether the actual annotations of the two frameworks would correspond to one another.

Benamara & Taboada (2015)'s work highlights 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. Such relations with no correspondence across taxonomies need more consideration when using an intermediate framework. A possible solution for this mapping issue is to create an intermediate representation for mapping between frameworks, rather than creating a new framework. Chiarcos (2014) was the first to attempt this: he developed an ontology to integrate RST-DT, PDTB and OntoNotes annotations within a higher-level, more general framework. In this framework, the RST-DT and PDTB labels are assigned new labels with respect to the more general relation senses in both schemata. As part of a deliverable for TextLink, Sanders et al. (2018) created a different version of an intermediate representation. They worked out a mapping for discourse relations from various frameworks to a set of properties, such that each coherence relation can be described in terms of its properties or "dimensions" (such as "causal" vs. "additive", "positive" vs. "negative", and "subjective" vs. "objective"). Through this intermediary representation in terms of properties, coherence relations can be mapped onto one another.

As a community, we now find ourselves in a situation where several alternative proposals have been made for mapping coherence relation labels onto one another, but they have not been evaluated. These mappings were mainly proposed based on relation definitions, and we do not know whether alternative proposals for mapping relations are equivalent. Moreover, we do not know how

1. http://textlink.ii.metu.edu.tr

well any of these proposals live up to the annotation of actual data. It is possible, for instance, that coherence relation labels should correspond to one another according to the annotation guidelines, but that differences in the operationalizations of frameworks (i.e., how annotators are asked to proceed for deciding on a relation label) lead to slightly different usages of these labels in actual annotations. Annotation manuals are also often (necessarily) incomplete – they list clear cases or prototypical examples of certain types of coherence relations, but annotators learn through training and discussion how to deal with various kinds of difficult cases. The effect of such implicit knowledge on annotation practice may also lead to discrepancies between actual annotations compared to theoretically posited correspondences. A difference between annotations in the PDTB vs. RST-DT frameworks that runs even deeper is that RST-DT annotation aims to reflect in its relational annotations the intention of the author, whereas PDTB annotations focus on the logical semantic relation between discourse segments. RST theory explicitly distinguishes between *intentional*, *semantic* and *textual* relations in their definition of coherence relations. Intentional relations express the writer's communicative intentions: annotations relate to the writer's goal or intended effect of each segment of a text with respect to the neighbouring segments, whereas semantic relations express information such as causality or temporal sequentiality. Textual relations are linear relations (such as LIST and DISJUNCTION). The distinction between semantic and intentional levels has a long history in the discourse coherence community, and are also referred to as informational vs. intentional (Moore & Pollack, 1992), subject matter vs. presentational (Mann & Thompson, 1988), propositional vs. illocutionary (Sanders & Spooren, 1999), and ideational vs. interpersonal relations (Hovy & Maier, 1995). In this article, we will use the terms *ideational* and *intentional* relations. When searching for a discourse phenomenon across several corpora, it may therefore not be sufficient to rely on the theoretically mapped labels. Instead, additional insights for which labels to consider or exclude can be gained from learning which labels correspond to one another empirically based on a large set of annotated instances. This article therefore also aims to elucidate the extent to which aspects of operationalization or training may affect labelling decisions during annotation.

The PDTB 2.0 (Prasad et al., 2008, 2014a) and RST-DT (Carlson et al., 2003) corpora represent an excellent opportunity for addressing these questions, as they have been partly annotated *on the same texts* (meaning that there is overlap between the annotations of the corpora). We will therefore focus on the PDTB to RST-DT label mappings that have been proposed by various researchers, and compare them against the correspondences between the annotations of actual instances found in the corpora. In the ideal case, we would expect to find (i) that the different mapping schemes are consistent with each other, i.e. they propose the same set of equivalences between relations and (ii) that the proposed theoretical equivalences also hold for actual annotated data, i.e. if a given text is annotated according to two different schemes, and there is a mapping between these schemes, the actual annotations should correspond to one another as specified by the theoretical mappings.

The current study thus extends previous work by mapping existing PDTB and RST-DT annotations onto one another, and by comparing them to theoretically posited correspondences between relations. This allows us to identify systematic differences between the annotations. Such discrepancies could be caused by differences in the respective operationalizations in discourse annotations, or differences in annotation goals (annotating semantic relationships between discourse segments vs. annotating the intended communicative function of a segment with respect to another one).

The empirical mapping approach taken here can also provide valuable insight for training future automatic discourse relation classifiers. Automatic discourse relation classification has seen an increase in attention in recent years, with discourse relation labels having been shown to help improve

on down-stream tasks such as machine translation (Meyer & Popescu-Belis, 2012; Popescu-Belis, 2016), question answering (Jansen et al., 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 such as the PDTB. A better understanding of how labels from different annotation schemes relate to one another may also help researchers to develop methods that can better exploit simultaneously the annotations from different corpora. First attempts at doing this using a multi-task setup have been proposed (e.g., Liu et al., 2016), but these approaches could potentially profit from insight in how the annotations relate to one another, and what types of discrepancies there are.

This article first provides background on the RST-DT and PDTB frameworks (Section 2), and then proceeds to laying out the proposed mappings between RST-DT and PDTB 2.0 relations according to the three recent approaches which specified such mappings (Chiarcos, 2014; Bunt & Prasad, 2016; Sanders et al., 2018) in Section 3. Section 4 discusses challenges due to differences in discourse segmentation between RST-DT and the PDTB, and describes the alignment algorithm for mapping RST-DT annotations to the Penn Discourse Treebank. Results of the discourse relation label mapping are discussed in Section 5, and compared to the theoretically posited correspondences. Section 3.5 discusses the results from our mapping to a previous approach which used a similar methodology, albeit in a simplified setting and much smaller scale (Rehbein et al., 2016). Finally, we discuss implications for annotation as well as automatic discourse processing in Section 6.

Our article makes the following contributions:

- We propose a method for aligning RST-DT and PDTB 2.0 annotations.
- We evaluate how well existing proposals for mapping discourse relation labels correspond to the mapping between existing RST-DT and PDTB 2.0 annotations.
- We analyse how compatible RST-DT and PDTB 2.0 annotations are.
- We identify sources of systematic discrepancies between annotations according to the two 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.
- We provide an aligned discourse corpus where both PDTB 2.0 and RST-DT annotations can be queried simultaneously.

2. Background

In this section, we describe the notions underlying the two discourse relation annotation frameworks and their corresponding corpora that are mapped in this article, namely the PDTB 2.0 and the RST-DT. This background will provide the necessary information for understanding the reasons behind differences in segmentation and discourse relation sense labelling that we find in our study.

2.1 Rhetorical Structure Theory Discourse Treebank (RST-DT)

The framework that is used to annotate the RST-DT (Carlson & Marcu, 2001) is based on the Rhetorical Structure Theory (RST) as proposed by Mann & Thompson (1988). There are different implementations of RST annotation, including for example the Basque RST TreeBank (Iruskieta et al., 2013), the Potsdam Commentary Corpus (Stede & Neumann, 2014b), and the CSTNews Corpus (Cardoso et al., 2011). These corpora all follow the overall style of RST annotation, but may differ in how exactly they define discourse segments, what exact set of relation labels is chosen, and how nuclearity is interpreted or operationalized (c.f. Stede, 2008). For the current study, we focus on the RST-DT style of RST.

Basic premises RST-DT is a descriptive theory of discourse relations, originally developed to guide computational text generation (Taboada & Mann, 2006). Relations in RST-DT can be of semantic, intentional, or textual nature (Carlson et al., 2003), and are explicitly classified into these three categories.

A fundamental constraint on RST annotation is that each part of a text has to be included into the overall discourse structure, and that the discourse structure has to be arranged into a tree structure. A second essential characteristic of RST-DT is the assignment of *nuclearity*: texts spans are characterized as nuclei or satellites (every relation has to consist of at least one nucleus). The nucleus is the more central part of a relation in the text (with respect to its intentional discourse structure), while the satellite is supportive of the nucleus (see Example (1-a)). Some relations have symmetrically important arguments by definition. These relations consist of two nuclei rather than a nucleus and a satellite (see Example (1-b)), and are referred to as *multinuclear* relations.

(1) a. [But even on the federal bench, specialization is creeping in,]$_{nucleus}$ [and it has become a subject of sharp controversy on the newest federal appeals court.]$_{satellite}$
— ELABORATION-ADDITIONAL, wsj_0601

b. [That isn't much compared with what Bill Cosby makes, or even Connie Chung for that matter (...)]$_{nucleus}$ [But the money isn't peanuts either, particularly for a news program.]$_{nucleus}$
— CONTRAST, wsj_0633

Segmentation and annotation process In order to annotate a text in RST-style, each document is first decomposed into non-overlapping sequential text spans, called Elementary Discourse Units (EDUs). EDUs generally consist of clauses, but attributions, relative clauses, nominal postmodifiers, and phrases that begin with a strong discourse marker are also considered EDUs in RST-DT (see Carlson & Marcu, 2001, p.3). After determining the EDUs of a text, nuclearity is assigned to the nodes and adjacent spans are linked together via rhetorical relations. In order to assign nuclearity, annotators should consider the writer's intentions (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 determined simultaneously with the assignment of a discourse relation (Carlson et al., 2003). The discourse relations are linked recursively, thereby creating a hierarchical tree structure.

RST-DT relations can hold between two (or more) non-overlapping text spans. The tree structure in RST annotations does not allow crossing or embedded EDUs. In order to deal with these limitations and the restriction that EDUs cannot overlap, a SAME UNIT tag was introduced in RST-DT, which allows annotators to express that an EDU is discontinuous. To illustrate this, consider

Example (2) below. The clause *when implemented* (unit 2) is embedded in another clause. As a result, unit 1a cannot be connected to its other half, unit 1b, because unit 2 cannot be skipped. The tag SAME UNIT can be applied in this situation to express that the units 1a and 1b in fact make up one discontinuous segment.

(2) ... [that it will,]$_{\text{unit 1a}}$ [when implemented,]$_{\text{unit 2}}$ [provide significant reduction in the level of debt and debt service owed by Costa Rica.]$_{\text{unit 1b}}$
— SAME-UNIT, wsj_0624

Annotators are instructed to annotate the writer's goal of each segment of a text with respect to the neighbouring segments and the resulting hierarchical structure of the entire document.

Discourse Structure and relational inventory Carlson et al. (2003) distinguish 78 relation labels, partitioned into 16 classes that share some type of rhetorical meaning (see Appendix A for a list of RST-DT's relational inventory, and see the manual (Carlson & Marcu, 2001) for the definitions). The inventory is data-driven, based on analysis of the RST-DT corpus. RST-DT's relational inventory can be divided into ideational and intentional relation labels: relations such as CAUSE or RESTATEMENT belong to the ideational group of relations, EVALUATION or PURPOSE belong to the intentional group (Hovy & Maier, 1995).

Some RST-DT classes contain relations that are not considered to be coherence relations in other approaches such as PDTB 2.0; examples include the ATTRIBUTION relations (which is also annotated in PDTB, but not considered a coherence relation) and cohesion relations (cases where coherence between discourse segments is not achieved through a specific coherence relation, but rather through cohesion). Such cases are annotated as ELABORATION relations in RST-DT. PDTB treats them as a separate type of discourse relations (namely ENTREL).

Relational definitions of RST-DT's classes are based on functional and semantic criteria, and not on signals, because the creators argued that no unambiguous signal for any relation was found (Taboada & Mann, 2006); e.g., a connective such as *but* can mark different types of relations.

2.2 Penn Discourse Treebank (PDTB)-style Annotation

The PDTB 2.0 corpus (Prasad et al., 2008) is the largest manually annotated discourse relation corpus available at the moment. The framework that was used to annotate the corpus is referred to as PDTB as well. The framework has also been used to create new corpora in other languages and genres, such as Arabic (Al-Saif & Markert, 2010), Italian (Tonelli et al., 2010), and Chinese (Zhou & Xue, 2015). This has resulted in different styles of PDTB annotation, but they can be considered to be interoperable (cf. Prasad et al. 2014b). Here, we use the term PDTB to refer to the framework that corresponds to the original PDTB-style annotation.

The PDTB research group is set to release an enriched, enlarged and modified version of the corpus (PDTB 3.0, Prasad et al. 2018), for which they have also adapted the framework. Compared to the PDTB 2.0, PDTB's 3.0 relational hierarchy has been simplified and extended. For example, in PDTB 3.0, the types CONJUNCTION and LIST are merged, subtypes of CONDITION and CONTRAST are removed, and relation types such as MANNER, PURPOSE and NEGATIVE CONDITION are added. The Level-3 (subtype) senses are now restricted to differences in directionality. Regarding the annotation procedure, the PDTB 3.0 takes a more systematic approach to annotating multiple labels for a single relation (Webber et al., 2016).

Basic premises PDTB-style annotation (Prasad et al., 2007, 2008, 2014a) is characterized by two basic premises. First, it has a theory-agnostic approach to annotation, thereby making no commitment to what kinds of high-level structures may be created from the low-level annotation of relations (Prasad et al., 2008). Second, PDTB follows a lexically-grounded approach to discourse relation representation, meaning that they focus on annotating lexical items that can signal discourse relations.

PDTB distinguishes between explicit and implicit discourse relations. Explicit relations are marked with a coordinating conjunction, subordinating conjunction or a discourse adverbial, which we will jointly refer to as *discourse connectives* in this article. Implicit relations, on the other hand, are not marked with a discourse connective. Instead, annotators are asked to insert a connective they think would best fit, and annotate the coherence relation with the inserted connective.

Three additional labels were employed for marking cases where an implicit connective could not be inserted. ALTLEX (alternative lexicalization) applies to coherence relations for which insertion of a connective leads to a perception of relation reduncancy (e.g., because including the connective *therefore* would sound odd and be redundant in a sentence starting with *for this important reason*). ENTREL is not a coherence relation as such, but rather marks cases which are connected only through cohesion rather than a specifiable discourse relation. NOREL is used when no discourse or entity relation holds. This label is necessary because of the way that annotations for implicit relations in PDTB was performed: annotators were asked to label all adjacent sentences (see also next section); sometimes adjacent sentences may however not stand in a direct relation to one another, because they belong to two different larger discourse segments. For these cases the NOREL label could be assigned.

Segmentation and annotation process Relations in the PDTB have two and only two arguments, referred to as Arg1 and Arg2. These arguments can be continuous or discontinuous. In the case of explicit relations, the argument that is syntactically bound to the connective is labeled as Arg2; the other argument is Arg1, and may be adjacent or non-adjacent with Arg2. Annotators were instructed to first identify explicit connectives based on a list of discourse cues; they then identified the discourse relational arguments. The selection of these arguments is restricted by the "minimality principle," according to which only as much material should be included in the argument as is *minimally required* and *sufficient* for the interpretation of the relation. Any material that is relevant but not "mininally necessary" for interpretating the relation is marked as supplementary information. Supplementary material is annotated for approximately 4% of PDTB 2.0 relations. After identifying explicit connectives and their arguments, a relation label is assigned, as in Example (3-a).

In PDTB 2.0, implicit discourse relations have only been annotated between adjacent sentences within paragraphs, as well as between complete clauses delimited by a semi-colon (";") or colon (":") (see also Prasad et al., 2017). In a first round of annotation, a connective was inserted, and the relation label was then assigned in a subsequent step, see Example (3-b). Because arguments of implicit relations have often been annotated as complete sentences or clauses of sentences with colons or semi-colons, the annotations have a slightly different pattern in segmentation for implicit relations compared to explicit relations (this observation will become important for the design of the mapping algorithm).

(3) a. Although [that may sound like an arcane maneuver of little interest outside Washington],$_{\text{Arg2}}$ [it would set off a political earthquake].$_{\text{Arg1}}$
— COMPARISON.CONCESSION.EXPECTATION, wsj_0609

b. [Mr. Carpenter denies the speculation]$_{Arg1}$ <implicit: meanwhile> [To answer the brokerage question, Kidder, in typical fashion, completed a task-force study].$_{Arg2}$
— TEMPORAL.SYNCHRONOUS, wsj_0604

Discourse Structure and relational inventory The framework distinguishes 43 relation labels (see Appendix A for the labels, and Prasad et al. 2007 for their definitions). These labels are organised in a hierarchy 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 (Prasad et al., 2008).

The PDTB taxonomy contains a few pragmatic labels (e.g., CONTINGENCY.PRAGMATIC CAUSE), but the focus of PDTB is on annotating ideational relations, rather than interpersonal relations. The frameworks differs in this respect from RST-DT.

Another important aspect distinguishing PDTB 2.0 annotations from RST-DT annotations is that PDTB annotators were allowed to assign several labels to the same relational arguments, when they found that multiple concurrent discourse relations held between the arguments. This is important for our evaluation of correspondences between assigned relation labels later on, as we decided to evaluate the correspondence in terms of the PDTB label that is most similar to the RST-DT label.

In terms of discourse structure, an important difference between PDTB-style annotation and RST-style annotation is that in PDTB-style annotation, it is not necessary for all parts of a text to be connected in an overall discourse structure. In fact, the minimality principle leads to many partial sentences not being part of any discourse relational argument. Furthermore, the restriction of only annotating implicit relations between adjacent sentences in PDTB 2.0 means that sentence-internal or cross-paragraph coherence relations may be missed (this was addressed in PDTB 3.0; also see a recent extension of PDTB annotation to VPs, Webber et al. 2016).

The bottom-up annotation of PDTB based on explicit connectives and adjacent sentences also entails that there is no guarantee that PDTB annotations follow a tree structure. In fact, Lee et al. (2006) report that discourse structure is a lot more variable than syntax, exhibiting nested, crossed and other "non-tree like" configurations. Lee et al. (2008) extends this analysis by focussing on cases where two different coherence relations share relational arguments of the form { X conn1 (Y } conn2 Z), where X, Y and Z are discourse relational arguments and conn1 and conn2 are connectives; they come to the conclusion that such non-tree structures are quite common in discourse.

3. Theory-based proposals for mapping RST-DT and PDTB 2.0 relations

A mapping of discourse relation labels between different frameworks can be achieved by determining the correspondence of relation labels from one framework to the other directly (e.g., PDTB 2.0 to RST-DT, RST-DT to SDRT, SDRT to PDTB 2.0). Another option is to map all frameworks to an intermediary representation, such that the mapping between any two frameworks can be obtained via this intermediary representation. This approach has the advantage of being more general in case many frameworks should be mapped onto one another: rather than creating a new mapping between the new framework and all other frameworks, researchers only have to create a single mapping from the new framework to the intermediary framework. Another advantage is that it produces a

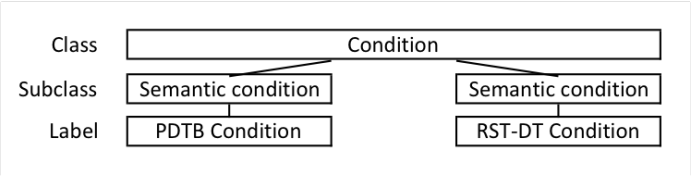

Figure 1: Simplified illustration of OLiA's hierarchical mapping approach.

candidate for a single set of relational labels potentially suitable for future annotation: the intermediary relation representation. All three of the mapping approaches discussed below propose such an intermediary representation.

The mapping approach by Benamara & Taboada (2015) could not be included here, because it focused on comparing RST and SDRT and has not yet completed the mapping to PDTB labels.[2]

3.1 Mapping according to the OLiA reference model

Chiarcos (2014) mapped the PDTB and RST-DT schemes onto each other as part of the Ontologies of Linguistic Annotation (OLiA,). OLiA provides a terminology repository that can be used to facilitate the conceptual interoperability of annotations (see Appendix B or the OLiA website[3] for more details). This is done using an intermediate level of representation that mediates between several existing frameworks. The intermediate representation is formalised as `subClassOf` descriptions. To illustrate this, consider Figure 1, which illustrates a hierarchical mapping of PDTB's CONDITION. In OLiA, this relation type is characterised as a subclass of *Semantic condition* relations, which is in turn a subclass of *Condition* relations. This can then be mapped onto RST-DT's class of *Condition*, which has the same superclasses. Chiarcos (2014) argues that ontologies are able to represent more fine-grained nuances of meaning, and to quantify the number of shared descriptions between annotations of different frameworks (Chiarcos, 2014). The Table in Appendix F shows the proposed correspondences (indicated by an 'o') between PDTB 2.0 and the RST-DT according to the proposal in OLiA.

3.2 Mapping via Unifying Dimensions

The Unifying Dimensions mapping (UniDim) was proposed by Sanders et al. (2018) with the goal of mapping labels from different frameworks onto each other using an interlingua (see Appendix C). In the UniDim proposal, relation labels are not mapped to intermediate labels; rather, they are described in terms of their characteristics, or values on certain dimensions. The set of unifying dimensions is an extended version of the dimensions originally proposed as the Cognitive approach to Coherence Relations (CCR; Sanders et al., 1992). The original CCR distinguishes four cognitive dimensions that apply to every relation, namely polarity, basic operation, source of coherence, and order of the segments. For example, a REASON relation would be represented as a relation with positive polarity, causal basic operation, objective source of coherence, and backward order of the segments. As PDTB and RST-DT make some distinctions which cannot be represented in terms of only these four dimensions, Sanders et al. (2018) extended CCR to account for more fine-grained properties of relations.

2. Personal communication November 2017
3. `http://www.acoli.informatik.uni-frankfurt.de/resources/discourse/`

The intermediate representation in terms of these dimensions allows for mapping relation labels from one framework into the representation as dimensions, and from this representation to the second framework. The method of describing relations in terms of their characteristics makes it easier to identify similarities and differences between relations. For example, similarities between relations can be described in terms of how many of their characteristics are identical: CAUSE and CONCESSION differ in one dimension (polarity) but they are both types of causal, objective, non-conditional relations (in the case of CONCESSIONS, the expected result has not occurred, or a result occurred even though the usual cause was not present). Appendix F shows the proposed correspondences (indicated by a 'u') according to the proposal by Sanders et al. (2018).

3.3 Mapping according to the ISO standard proposal

Bunt & Prasad (2016) provide a mapping of frameworks that is based on a different system. They proposed an international standard (ISO standard) for coherence relation annotation, which consists of a set of 20 core relations that are commonly found in some form in existing approaches (see Appendix D). They did not aim to provide a fixed and exhaustive set of coherence relations; rather, they aimed at providing an open, extensible set of relations. Bunt & Prasad (2016) propose that the ISO standard can be used for future annotation efforts, as well as for mapping between annotations using different frameworks. To this end, they provided mappings of these ISO relations to other frameworks, including PDTB and RST-DT. From these proposed correspondences to ISO standard candidate relations, we can infer how the PDTB 2.0 and RST-DT relations correspond to one another. The full table of hypothesized correspondences according to the ISO proposal are marked by 'i' in Appendix F.

3.4 Discussion of agreement and discrepancies between proposed mappings

Appendix F shows the grid of proposed correspondences between the schemes. While we can see many cells that include 'o' , 'u' and 'i' (OLiA, UniDim, and ISO, respectively), indicating that all approaches agree that these relation labels should correspond to one another, we can also see some relational labels that have only been linked by one of the proposals. We have identified three main reasons for these discrepancies, which we discuss below: differences in granularity of mapping schemes, differences in how concepts (in particular, order and subjectivity) are defined by the frameworks, and differences in the interpretation of definitions in the annotation guidelines.

Granularity of the proposals The first source of discrepancies is the granularity of the intermediary schemes. In principle, we would like to obtain a one-to-one mapping between the labels of one framework to another one. However, this is impossible if one framework makes more fine-grained relational distinctions than the other one, or if distinctions between relations don't correspond to one another. In this case, a one-to-many or many-to-many mapping will be necessary. From the perspective of desigining and intermediary mapping scheme, it is methodologically most preferable to have a scheme that does not "conflate" several relational labels, i.e. two different labels from a source scheme should never be mapped onto the same intermediary label. When designing a mapping scheme that interfaces with several frameworks at the same time, we necessarily obtain a one-to-many mapping from source framework to intermediary representation, and many-to-one mapping from intermediary representation to target framework. Mapping schemes like ISO, that

3. Mappings from the paper for certain labels were updated; personal communication January 2018.

was not designed for mapping but as a new set of relation labels, may already employ a many-to-many mapping from the source framework to the intermediary ISO representation, and hence only allow for a "coarser" mapping.

For instance, the causal relation labels in the ISO-based mapping are coarser than the distinctions in RST-DT and PDTB, because the proposed ISO standard doesn't differentiate between semantic and pragmatic causal relations (a distinction also known as objective vs. subjective or content vs. epistemic); i.e. causal relations where the link can be established based on the semantics of the two arguments and causal relations where the author posits a causal link. OLiA and UniDim do distinguish between semantic and pragmatic relations, and therefore do not map certain RST-DT labels that the manual mentions are semantic (e.g., REASON and EXPLANATION-ARGUMENTATIVE) to PDTB's pragmatic causal label JUSTIFICATION, whereas ISO does map semantic RST-DT labels to JUSTIFICATION.

An example of a relational category that is more fine-grained in the ISO proposal than in the other frameworks is EXCEPTION, which corresponds to PDTB's EXCEPTION, but has no equivalent mapping to an RST-DT relation. As a result, PDTB's EXCEPTION cannot be mapped to a corresponding RST-DT label based on the ISO proposal. In UniDim, however, EXCEPTION is mapped to CONTRAST, ANTITHESIS and PREFERENCE, i.e. a set of more general labels with similar characteristics to PDTB's EXCEPTION. The same goes for RST's MEANS relations.

Certain labels can also not be mapped by OLiA. For example, RST-DT's EVALUATION, COMMENT and DEFINITION relations are part of the superclass *Assessment*, which doesn't occur in the PDTB inventory; furthermore, the relations BACKGROUND and CIRCUMSTANCE are part of the superclass *Background*, which also doesn't occur in PDTB. OLiA also doesn't map any COMPARISON RST-DT relations. Labels could be mapped using their supersuperclass, but this would not be very informative (e.g., it would result in mapping the *Background* label to all PDTB EXPANSION labels).

In temporal relations marked with connectives such as *before* and *after*, PDTB assigns the label TEMP.ASYNC.SUCCESSION to relations marked with *after*, and the label TEMP.ASYNC.PRECEDENCE to relations marked with *before* as a subordinating conjunction. Similarly, RST uses the labels TEMPORAL-AFTER and TEMPORAL-BEFORE for these instances. UniDim and OLiA therefore proposed to map TEMP.ASYNC.SUCCESSION onto TEMPORAL-AFTER and TEMP.ASYNC.PRECEDENCE to TEMPORAL-BEFORE. ISO on the other hand more generally maps both of PDTB's TEMPORAL.ASYNC relations onto all asynchronous temporal relations in RST.

Discrepancies in granularity occur because of differences in the goals of the mapping frameworks, and different systems of mapping. UniDim was designed to be able to map labels, and, as a result, the interlingua can be used to map relational labels even when there is no direct equivalent. ISO, on the other hand, was proposed as a new set of relations; i.e. mapping was not its primary goal. The mapping that is provided in Bunt & Prasad (2016) mainly includes direct correspondences. OLiA also has the goal to identify all correspondences between frameworks, but because of the hierarchical structure of the interlingua, a relation label that is part of a unique superclass in one framework can often not be mapped properly to another framework.

Subjective vs. epistemic relations A somewhat difficult issue is the notion of subjectivity. A relation can be objective, subjective or epistemic. RST distinguishes between objective and subjective relations with labels such as CAUSE and RESULT on the one hand and REASON, EVIDENCE and EXPLANATION-ARGUMENTATIVE on the other hand, while PDTB distinguishes between non-

RST-DT label	Mapping to PDTB according to:		
	OLiA	**UniDim**	**ISO**
COMPARISON	Not mapped	CONJUNCTION	CONTRAST
ANTITHESIS	CONTRAST	CONCESSION & CONTRAST	CONCESSION
ELAB.-OBJECT-ATTRIBUTE	EXPANSION	SPECIFICATION & GENERALIZATION	ENTREL
BACKGROUND	Not mapped	CONJUNCTION & ASYNCHRONOUS	ENTREL
CIRCUMSTANCE	Not mapped	CONJUNCTION, SYNCHRONOUS & ASYNCHRONOUS	SYNCHRONOUS

Table 1: Overview of the differences between the OLiA, UniDim and ISO-based mappings caused by different interpretations of definitions.

epistemic relations CONTINGENCY.CAUSE.REASON and CONTINGENCY.CAUSE.RESULT and epistemic relations such as CONTENGENCY.PRAGMATIC CAUSE.JUSTIFICATION. OLiA proposed to map RST-DT's subjective labels such as EVIDENCE label only to PDTB's epistemic label PRAGMATIC CAUSE, which does however not do justice to PDTB's more restrictive notion of subjectivity as epistemic relations only. A similar problem occurs for RST-DT's MANNER, MEANS, PROBLEM-SOLUTION and ENABLEMENT.

Difference in interpretations of definitions The third source of discrepancies is rooted in different interpretations of relational definitions in the annotation manuals, and hence represents the theoretically most interesting case for comparison between proposals. In the remainder of this section, we will focus our discussion on these cases. Table 1 provides an overview of the differences between the OLiA, UniDim and ISO-based mapping according to this type of discrepancies.

RST-DT's COMPARISON First, the proposals differ in their mapping of RST-DT's COMPARISON relations. The manual states that the two segments of a COMPARISON relation are not in contrast with each other (Carlson & Marcu, 2001, p. 50). Based on this description, UniDim mapped COMPARISON to PDTB CONJUNCTION. ISO, however, mapped COMPARISON to PDTB's CONTRAST relational class. Finally, OLiA mapped RST-DT's COMPARISON to the superclass *Non-contrastive comparison*. However, none of the labels in the PDTB have a superclass *Non-contrastive comparison*, and therefore RST-DT's labels do not have a correspondence in PDTB. The mapped data will be able to indicate how the COMPARISON label was used in practice

RST-DT's ANTITHESIS The frameworks also disagree on the mapping to contrastive and concessive relations. The former is a relation of semantic opposition; the latter contains a denial of expectation. To illustrate the difference between the two types of relations, consider the following examples:

3. Differences that are not discussed in this section include for instance: UniDim's mapping of RST-DT HYPOTHETICAL to PDTB PRAGMATIC CONDITION (in addition to CONDITION); ISO's mapping of RST-DT EVALUATION to the general EXPANSION class; OLiA's mapping of RST-DT SUMMARY relations to EQUIVALENCE (in addition to SPECIFICATION and GENERALIZATION).

(4) Dylan used to live in Washington DC, but now he lives in Baltimore.

(5) Dylan lives in Baltimore, but he works in Washington DC.

Example (4) presents a simple contrast between where Dylan used to live and where he lives now. The relation could also have been expressed by the connective *whereas*, which is a typical marker of contrastive relations. In Example (5), the first segment informs you where Dylan lives now. This segment presupposes an implicit expectation that Dylan also works there, but the second segment denies this expectation: Dylan works in a different city. This is typical of CONCESSION relations: one argument creates an expectation of a cause or consequence, which is denied by the other argument. The relation could also have been expressed with typical markers of concessive relations, such as *even though* or *nevertheless*. The distinction between CONTRAST and CONCESSION is relatively difficult to make, even for trained annotators (see, for example, Robaldo & Miltsakaki, 2014; Zufferey & Degand, 2013).

The frameworks all disagree on the mapping of the RST-DT label ANTITHESIS to PDTB CONTRAST and CONCESSION. RST-DT's annotation manual states that ANTITHESIS is a contrastive relation, but some of the examples that are provided are concessive relations. In UniDim, ANTITHESIS is therefore mapped to both contrastive and concessive PDTB relational labels, whereas in ISO, ANTITHESIS is mapped only to concessive labels. In OLiA, ANTITHESIS is mapped onto CONTRAST only.

RST-DT's ELABORATION-OBJECT-ATTRIBUTE The proposals differ in their mapping of RST-DT's ELABORATION-OBJECT-ATTRIBUTE label: OLiA maps it to general EXPANSION relations, UniDim maps it to PDTB's SPECIFICATION and GENERALIZATION relations, while the ISO-based proposal maps ELABORATION-OBJECT-ATTRIBUTE to ENTREL.

RST-DT's BACKGROUND **and** CIRCUMSTANCE Finally, UniDim and ISO differ in their treatment of RST-DT's BACKGROUND and CIRCUMSTANCE. ISO maps BACKGROUND to PDTB ENTREL, whereas UniDim maps BACKGROUND to CONJUNCTION and TEMPORAL.ASYNCHRONOUS based on the description of BACKGROUND in the manual: "The satellite IS NOT the cause / reason / motivation of the situation presented in the nucleus ... the events represented in the nucleus and the satellite occur at distinctly different times" (Carlson & Marcu, 2001, p. 47). Regarding CIRCUMSTANCE relations, UniDim and ISO agree on mapping these to SYNCHRONOUS relations, but UniDim also maps to ASYNCHRONOUS and CONJUNCTION labels.

The disagreements that are attributable to different interpretations of the definitions in the manual will be systematically analysed with respect to actual annotations, and discussed in Section 5, in order to determine which of the proposed correspondences are justified by the actual data. These results can then also be used to clarify annotation guidelines for future research, and highlight which definitions are particularly susceptible to inconsistent annotation.

3.5 Previous work on empirical evaluation of mapping coherence relation annotations

We are aware of three other efforts (Rehbein et al., 2016; Scheffler & Stede, 2016a; Polakova et al., 2017) to systematically evaluate the mapping of discourse annotations from different frameworks on the same text. Polakova et al. analyse the same resource as us: the PDTB and RST-DT corpora. Their study focuses on the question of how implicit relations are signalled. To this end, they identify a subset of 472 implicit relations in PDTB that have same argument spans and matching labels in

the RST annotation, and analyse what types of additional signals are present in these relations based on the annotations of such signals in the RST Signalling Corpus (Das & Taboada, 2018). Polakova et al. (2017) find that a large proportion of the PDTB implicit relations are signalled by semantic signals expressed in specific lexical chains in the relational arguments. Another frequent pattern observed among these implicit relations were parallel syntactic constructions between relational arguments, as well as the "unsure" label. They conclude that implicit relations cannot be easily annotated automatically based on signals such as the ones annotated in the RST Signalling corpus, as they are hard to identify and the subset of semantic lexical chains often falls out of well-defined semantic relations such as synonymy, antonymy etc.

Scheffler & Stede (2016b) compared PDTB 3.0-style annotations and RST-style annotations on the German Potsdam Commentary Corpus (PCC; Stede & Neumann, 2014b). The PCC includes 1104 explicit connectives annotated in PDTB-style but lacks PDTB-style annotations for implicit relations. Scheffler & Stede (2016b) propose a simple method for mapping the PDTB 3.0-style and RST-style annotations for explicitly marked relations in the Potsdam Commentary Corpus (PCC) onto one another, in order to empirically observe commonalities and differences in the annotations of discourse structure between the two approaches. They do not, however, compare relation labels for the mapped relations. The alignment algorithm used in Scheffler & Stede (2016b) compares the spans of the relation argument annotations, and distinguishes different segmentation constellations. They observe that the majority (84%) of instances in their corpus consists of cases that are easy to map, including exact match of discourse relational arguments (41%) and "boundary match" (39%), where a relation is annotated between two adjacent text spans, and the boundary between the two arguments is identical. They however also report difficult cases of non-local relations where the segment boundaries differ (13%) (as in the example of the contrast relation in Figure 3 below), and cases in which no match is possible (3%). As we will discuss in more detail in Section 4, we also find segmentation and alignment to be a challenging first step in aligning the annotations of the PDTB and RST-DT corpora.

Rehbein et al. created an English corpus of spoken discourse containing PDTB 3.0 and CCR annotations for every relation. They segmented the texts first, and then proceeded to assign sense labels according to both schemes for that given segmentation. This procedure thus avoided challenges related to differences in segmentation between frameworks. After annotation, the relation labels were mapped onto one another directly, and the correspondence between PDTB and CCR annotations was evaluated.

Rehbein et al. (2016) reported three systematic biases introduced in the operationalizations of PDTB 3.0 and CCR, which lead to differences in annotations in some areas: A first observation holds that PDTB's additive relations EXPANSION.INSTANTIATION, EXPANSION.RESTATEMENT.SPECIFICATION and EXPANSION.RESTATEMENT.EQUIVALENCE were quite often (30% of relations) annotated as causals in CCR. This finding is consistent with an observation we made in the present analysis, where we found that PDTB EXPANSION.INSTANTIATION and EXPANSION.RESTATEMENT are often annotated with a causal label in RST, see Section 5.3. As noted by Blakemore (1997) and Carston (1993), and shown in Scholman & Demberg (2017), these types of discourse relations are often ambiguous, as examples can at the same time also serve as evidence for a claim.

The second category of systematic disagreements concerns COMPARISON.CONTRAST and 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 con-

stitutes a CONCESSION, but note that distinguishing between contrastive and concessive discourse relations is a well-attested difficulty (see, for example, Robaldo & Miltsakaki, 2014; Zufferey & Degand, 2013). Again, the same difficulty is obvious in the mapping between RST-DT and PDTB, as discussed in Section 5.1.

As a third pattern of disagreements, Rehbein et al. (2016) report effects of operationalization of annotation procedures: Some instances marked by *but* were annotated as positive polarity relations in PDTB, but as negative in CCR (including instances marked with *but also*). These discrepancies were systematic and due to an annotation instruction – as a rule, all relations that can be marked with *but* are annotated as negative polarity relations in CCR. While this specific pattern is not relevant for the mapping between PDTB and RST-DT, we note that annotation operationalizations by different frameworks might have substantial effects on the annotations.

4. Data, Segmentation 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 et al., 2003).[4] The PDTB 2.0 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).

Our investigation will be based on the intersection of the PDTB and the RST-DT, with annotations from both frameworks included as different annotation layers.

4.1 Segmentation

Comparing annotations of the two corpora is not a trivial task, because annotations not only differ in the label sets that were used, but also in terms of segmentation. Firstly, there are discrepancies in what is considered an "elementary discourse unit" in RST-DT vs. what is considered a discourse relational argument or an attribution in PDTB 2.0. Note that in the remainder of this paper, we use the term "segment" to refer to the text elements that are part of a relation; that is, the *arguments* in PDTB and the *EDUS*, *nuclei* and *satellites* in RST-DT. There are also differences in the discourse structure: RST annotates discourse trees spanning the whole document, while PDTB 2.0 only annotates relations between adjacent sentences and relations marked by an explicit connective. This results in a considerably lower number of PDTB relations than RST-DT relations for the same text. We therefore use 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.

PDTB's minimality principle (cf. Section 2.2) and RST's tree structure (cf. Section 2.1) influence the result of the segmentation and annotation steps. An annotation alignment process must therefore take into account systematic differences arising from the respective segmentation procedures. In the automatic alignment step, our goal is to map as many discourse relation labels as possible in order to get a maximally complete picture regarding how well the annotations correspond

4. These numbers are the result of averaging over the inter-annotator agreement scores reported for every two annotators in Table 2 of Carlson et al. (2003).

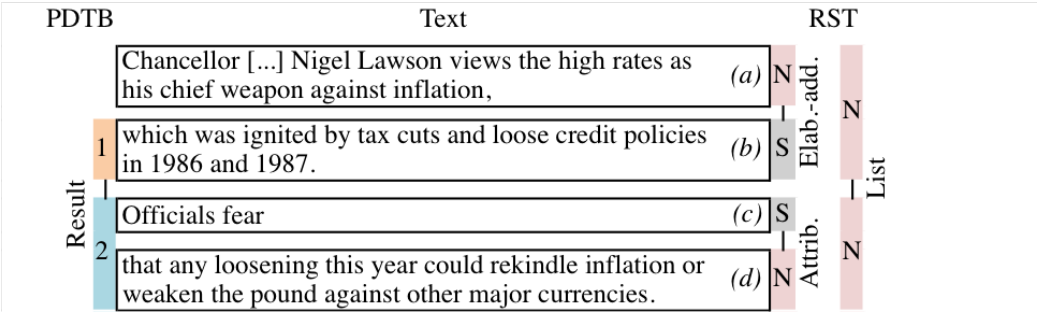

Figure 2: PDTB and RST-DT annotations for a paragraph of wsj_1172. *1* refers to Arg1 in PDTB; *2* refers to Arg2. *N* refers to Nucleus in RST; *S* refers to Satellite. (a-d) refer to RST-DT's EDUs.

to one another. At the same time, we must only map those labels where annotators inferred the same relation – if the RST-DT annotators annotated a relation holding between two text segments, and the PDTB annotators marked a relation between two different segments, these labels should *not* be recorded as valid alignments, and labels hence shouldn't be compared. To illustrate this, consider Example 2, which presents the PDTB (left) and RST-DT (right) annotations for a fragment of a Wall Street Journal article. PDTB segmented Arg1 differently than RST-DT, leading to a difference in interpretation. PDTB considers segments (c-d) as the result of the event in segment (b), whereas RST-DT focused on the different opinions expressed in segments (a-b) and (c-d). The disagreement between the two labels (RESULT vs. LIST, respectively) does not stem from annotator disagreement regarding the label, but from a more fundamental difference in segmentation. Such cases should therefore not be included in the evaluation of mapped labels.

The alignment algorithm proposed in this article (see Section 4.2 below) is more complex than the one proposed in Scheffler & Stede (2016b), in order to better address those cases for which there are differences in segmentation between annotation layers. The core idea of how valid alignments can be identified even in the face of mismatches between relational arguments builds on the Strong Nuclearity hypothesis (Marcu, 2000), which was used for RST-DT annotation. The Strong Nuclearity hypothesis 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. Note that the notion of "nuclearity" has had several slightly different interpretations throughout the conception and further development of RST, as laid out in Stede (2008). Independent of the theoretical discussion about the intentions behind nuclearity as such, the specific notion used in RST-DT annotation is helpful for determining relation alignment. RST-DT's nuclearity was assigned in an instance-by-instance decision for identifying those parts of a discourse relation which are crucial for that relation to hold; it was not used as a general property of relation types, as in some other variants of RST annotation. In that sense, the strong nuclearity annotation guideline from Marcu (2000) is related to the minimality principle used in PDTB 2.0 annotations: both help to indicate which segments of the text are central to establishing the discourse relation. To illustrate this, consider the CONTRAST relation in Figure 3; the Strong Nuclearity hypothesis means that if the relation holds between (6-7) and (8-9), it should also hold between (6) and (8), but not between (7) and (8) or (7) and (9). In the following, we will use the expression *nucleus path* to refer to the path between a complex argument of a high-level relation, and the single EDU which one ends up with if always following the path down the segments annotated as nucleus.

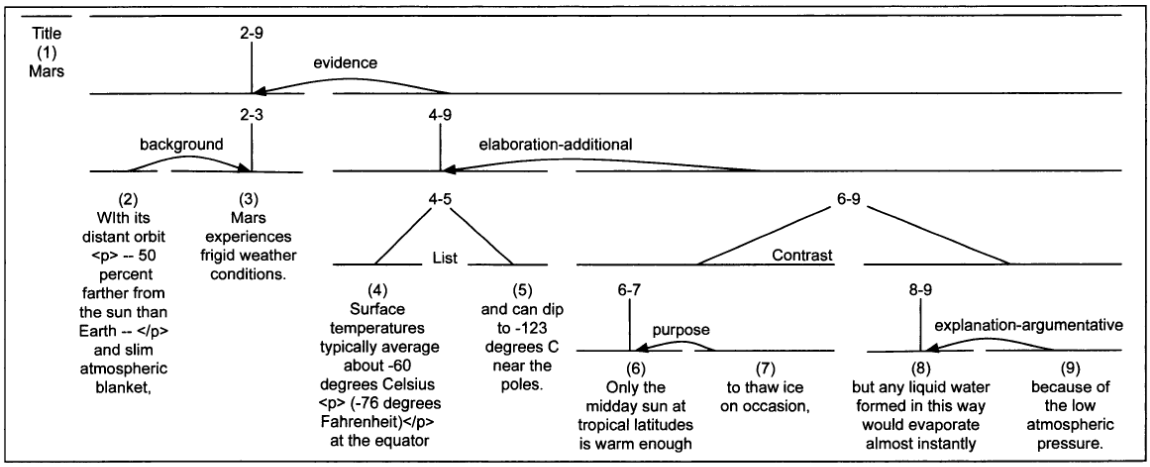

Figure 3: RST discourse structure, Figure 1.1 from Marcu (2000).

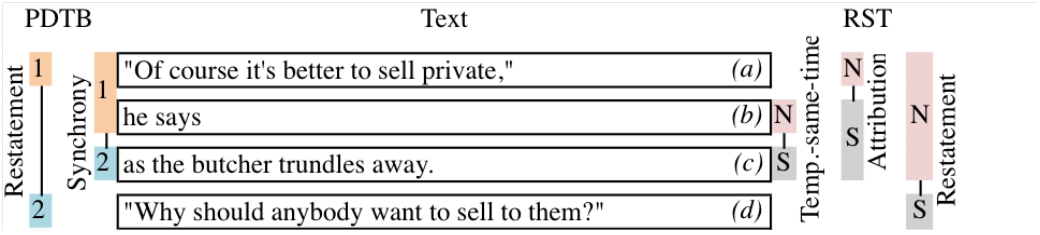

Figure 4: PDTB and RST annotations for a section of wsj_1146.

4.1.1 SEGMENTATION CONSTELLATIONS

We will now go through the different segmentation and alignment constellations using examples, to explain where challenges in alignment lie, and how these are dealt with by our alignment procedure. For ease of reference, we will here adopt the PDTB distinction between explicitly marked and implicit relations, even when referring to RST-DT annotations.

PDTB relations with adjacent arguments The simplest case is an exact match between the discourse relational arguments for the two annotation layers. Additionally, there can be cases where the argument spans largely overlap but differ in their exact boundaries. Consider Figure 4: in PDTB 2.0, segments (a-b) and (c) are connected in a TEMPORAL.SYNCHRONY relation. In RST-DT, a TEMPORAL-SAME-TIME relation was annotated between segments (b) and (c). Even though the spans differ in whether (a) is included, they clearly correspond to each other. As neither of the discourse relational arguments of the SYNCHRONY relation in PDTB nor the TEMPORAL-SAME-TIME relation in RST-DT is complex (i.e., no other relations are embedded under either of their arguments), it is straightforward to decide which relation labels should correspond to each other.

PDTB relations with non-adjacent arguments We also frequently encounter more complex cases, where the PDTB arguments are not directly adjacent to one another. This can happen both for explicitly marked relations and for implicit relations (when the sentences are adjacent but the chosen spans do not cover the complete sentence). Whenever the PDTB arguments are not adjacent, we will either have a mismatch between the size of the discourse segments in that the RST-DT EDU

is larger than the PDTB argument, or in that the RST-DT argument is *complex*, i.e. it consists of other relations. In Figure 4, this is the case for the RESTATEMENT relation: in PDTB, this relation holds between segments (a) and (d), whereas in RST-DT, the relation holds between segments (a-c) and (d). For deciding whether the PDTB EXPANSION.RESTATEMENT and RST-DT RESTATEMENT relations should be aligned, we rely on the Strong Nuclearity hypothesis. It says that the complex relation between (a-c) and (d) should also hold between the nucleus of (a-c), hence (a) and (d). We can then infer an exact match between discourse relational arguments (a) and (d) between the two annotation layers, and map the labels onto one another.

Such cases also occur among explicitly marked relations. Since RST-DT relations are annotated in a hierarchical tree structure, relations connecting non-adjacent sentences will have large discourse relational arguments. Consider Example 3 again: RST-DT annotates a CONTRAST relation with segments (6-7) as one nucleus and (8-9) as the other nucleus. In the PDTB annotation, because of the minimality principle of marking discourse relational arguments, the relation marked by *but* has segment (8) as its ARG2 and segment (6) as its ARG1. Similarly, the EVIDENCE relation between segments (2-3) and (4-9) would differ in terms of its argument boundaries in PDTB annotation, as segments (6-9) would typically not be included in the ARG2 of the implicit relation. Nevertheless, an alignment of relations is possible given nuclearity annotation, and labels from such cases are included in the mapping.

Relations with inconsistent nuclearity There are however also cases for which labels should not be mapped due to discrepancies in what relation the annotators intended to label. These cases can typically be identified through inconsistencies between the discourse relational arguments annotated by PDTB and the nuclearity assignment annotated in RST-DT.

To illustrate this, consider the passage in Figure 5. In this example, one would have to map PDTB relation CONTRAST to RST-DT's CONSEQUENCE, if one were to only take into account maximal overlap of discourse segments, but ignore nuclearity: the difference in span size (PDTB's annotation excludes segments (a) and (e)) affects the interpretation of the relation. In PDTB's annotation, the relation holds between the state's action and the farmers' actions. In RST-DT's annotation, on the other hand, the annotated relation holds between the state's action and the consequences of that action. The relation labels therefore correspond to different interpretations. Such cases are flagged automatically because PDTB's Arg2 cannot be safely traced to the nucleus of the satellite of RST-DT's relation, due to an intervening multinuclear relation.

In order to investigate how often relations with intervening multinuclear relations (i.e., relations in which one of the segments consists of a larger tree branch including a multinuclear relation) occur in the data and to what extent they pose a problem for automatic alignment, we extracted all instances that contain an intervening multinuclear relation. In total, 892 relations (13% of the data) have one or more intervening multinuclear relations. However, not all of these pose a potential risk to alignment – cases where the multi-nuclear relation is the relation to be compared, and multi-nuclear relations that are not on the nucleus path, are safe to map. We found that 295 multinuclear cases were flagged as potentially violating strong nuclearity. We manually checked 50 of these flagged instances and found that almost all of them indeed do not represent valid alignments, and should therefore be excluded from the mapping analysis.

Internal relations The segmentation granularity between the two frameworks can differ, which can lead to a specific instance from the finer-grained framework not being mapped to a label in the coarser framework. For example, two RST-DT EDUs could occur internally within a sentence,

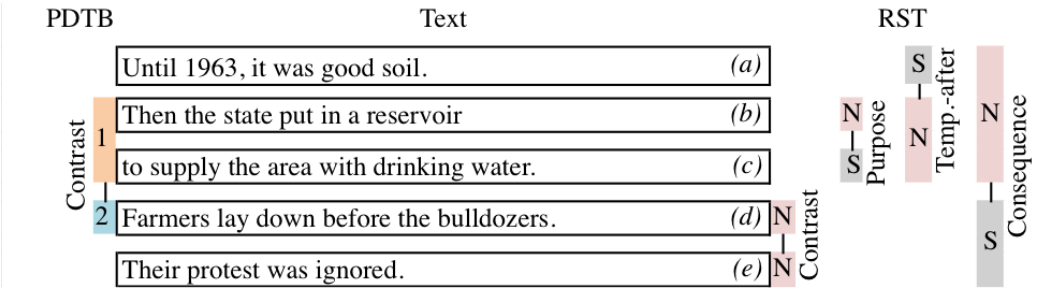

Figure 5: PDTB and RST-DT annotations for a paragraph of wsj_1146. Note: Only the PDTB annotation that is relevant for this example is included.

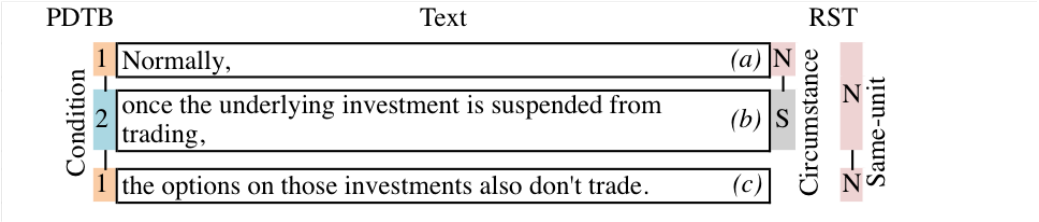

Figure 6: PDTB and RST-DT annotations for a paragraph of wsj_1962.

without this relation being annotated in the PDTB annotation layer. Figure 3 shows an example where the BACKGROUND relation between segments (2) and (3) has no corresponding PDTB annotation. Similarly, there are also cases where a connective is annotated as an explicit marker for a relation in the PDTB 2.0 annotation, but both arguments of this relation are part of the same RST EDU, as in Example (6): PDTB annotated the connective *because* whereas in RST-DT, the entire sentence was considered as one EDU. Internal relations can inherently not be aligned, and hence were excluded from our mapping analysis.

(6) [That's] because [municipal-bond interest is exempt from federal income tax – and from state and local taxes too, for in-state investors.]
— CONTINGENCY.CAUSE.REASON, wsj_0689

RST-DT's SAME-UNIT Centrally embedded relations (where the ARG1 is discontinuous and ARG2 is located inside the span of ARG1) can be successfully mapped by our algorithm if the RST-DT annotation contains a SAME-UNIT annotation that links the discontinuous parts of the RST segments mapping to the PDTB ARG1. This is illustrated in Figure 6. If a SAME-UNIT annotation in RST-DT is identified, the label of the corresponding PDTB relation is mapped to the label of the relation below the SAME-UNIT relation (96 instances). We manually verified a subset of mapped relations to make sure that this heuristic is valid in practice (see Section 4.3). For SAME-UNIT relations where both of the RST-DT segments contain multiple EDUs (44 instances), the corresponding relation could not be determined automatically in a reliable way. These cases are flagged and excluded from the mapping analysis.

Cohesion The PDTB relation label ENTREL is used to mark cohesion when no specific coherence relation can be identified. As cohesion is not defined as a type of coherence relation, the

ENTREL label is not included in the UniDim mapping. However, RST-DT contains several labels (ELABORATION-ADDITIONAL, DEFINITION, BACKGROUND) that tend to be annotated when there is cohesion but no easily identifyable coherence relation. In our analysis, we therefore decided to include the ENTREL label, to check whether it actually coincides with RST-DT's labels signalling cohesion.

4.2 Alignment Algorithm

Our procedure for mapping annotations provided by the two frameworks takes the PDTB relations as a starting point, because there are fewer relations annotated in the PDTB 2.0 than in the RST-DT. The alignment procedure is aimed at determining the optimal mapping of each PDTB relation to an RST relation. Thereby, our goal is to identify as many valid correspondences between annotations as possible, but at the same time minimise mapping "noise" by also identifying those cases in which we cannot be sure that the annotators identified the same underlying relation.[5] Our mapping algorithm involves two major steps:

1. Identifying for every PDTB discourse relation those RST-DT segments (EDUs or sub-trees containing more than one EDU) that best correspond to the PDTB segments Arg1 and Arg2 separately.

2. Identifying the RST-DT relation label that describes the relation between the Arg1-equivalent and Arg2-equivalent spans.

In the first step, for each PDTB argument, we iterate over all RST EDUs in the source file and select the one with maximum overlap (common characters) and minimum margin (extra characters). We then determine whether a PDTB argument should be aligned to more than a single RST EDU by iterating over all RST relation annotations (sub-trees spanning over several EDUs) using the same criteria. Having identified the closest matching RST-annotated text spans for both arguments of the PDTB relation (Arg1-equivalent and Arg2-equivalent RST spans), we move to step 2 to find the lowest RST relation within the discourse tree that contains the two RST spans obtained in the previous step in different arguments. In this step, several error flags are set for manual investigation of possibly invalid alignments, which we briefly introduced in the previous section and will discuss in more detail below.

To illustrate the alignment procedure, consider the example shown in Figure 7. The first step of the alignment algorithm is to identify the corresponding text segment for PDTB's Arg2 (here, segment a) and Arg1 (here, segments c-d). In order to do that, the algorithm would first compare all EDUs separately to the PDTB argument spans to identify the ones with most character overlap, and then move on to comparing larger spans. This would mean that it would first identify (c) as a matching EDU for PDTB's Arg1, and then replace this by the better-matching combination of the span comprising both EDUs (c) and (d).

We can see that in this example, segments (c-d) are conjoined in a CONDITION relation in RST-DT, and are part of an ATTRIBUTION relation with segment (b), as well as a LIST relation with segments (e-g). Finally, there is a relation connecting segment (a) to segments (b-g). PDTB's

5. The alignments will be made available.

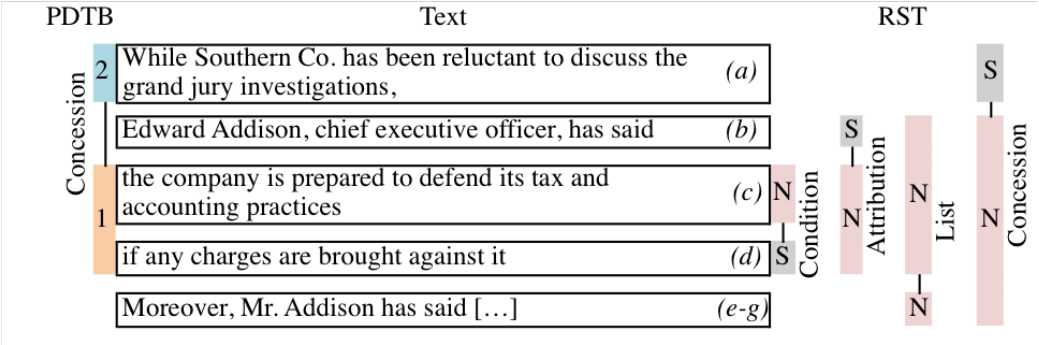

Figure 7: PDTB and RST-DT annotations for a paragraph of wsj_0619. Note: Only the PDTB annotation that is relevant for this example is included. Segment (e-g) is simplified, it actually consists of multiple RST-DT relations.

Arg1 is hence embedded in several relations as part of a tree branch in RST-DT, including one multinuclear relation (the LIST relation).

As part of its second step, the automatic algorithm proposes an alignment of the PDTB CONCESSION label to the RST CONCESSION label, because RST's CONCESSION relation is the lowest relation which includes the PDTB Arg2-equivalent span (a) and the Arg1-equivalent span (c-d) in separate arguments. However, it would also automatically flag this instance due to the intervening multi-nuclear LIST relation: the multi-nuclear relation prevents it from unambiguously identifying the nucleus path from the nucleus of RST's CONCESSION relation to the relevant textspan that corresponds to PDTB's Arg1.

The algorithm flags all instances for which the mapping was potentially problematic, based on various criteria. First, relations for which PDTB arguments are discontinuous (e.g., containing text spans which are marked as not belonging to the relation, or centrally embedding the other PDTB argument, or overlapping with it) are flagged to allow for further manual checking. Second, relation mappings that are inconsistent with the Strong Nuclearity hypothesis (Marcu, 2000) were flagged automatically for exclusion from further analysis. For instance, imagine that PDTB had annotated a relation between segments (a) and (d) in the Example in Figure 5. Such a constellation would be detected as violating strong nuclearity because (a) is the satellite of the relational argument that forms a relation with a span including segment (d). The labels for such a pair of relations should then not be compared. Additionally, we added flags for relations that contain intervening multi-nuclear relations, for relations that were originally labelled as SAME-UNIT, and for relations that contain an intervening RST-DT ATTRIBUTION relation.

4.3 Results of the alignment procedure

In total, we were able to include 76% of PDTB relations from the joint corpus into our mapping analysis (a total of 5141 relations). 52% of these relations (a total of 2662) have directly corresponding argument spans, for which argument spans are exactly identical or differ only with respect to punctuation or inclusion/exclusion of connective in a segment.

The remaining 48% (2489 instances) of the data included in the mapping analysis consists of relations for which the RST-DT tree is more complex than the PDTB relation. In other words, at least one of the PDTB arguments mapped onto an RST-DT relation that consists of multiple

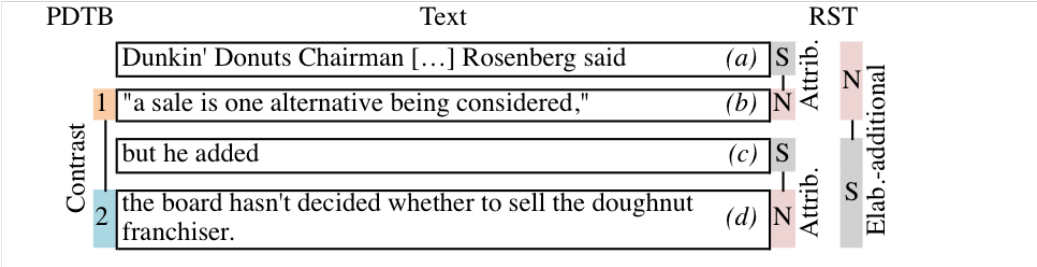

Figure 8: PDTB and RST-DT annotations for a paragraph of wsj_1176. Note: Only the PDTB annotation that is relevant for this example is included.

RST-DT EDUs. In order to investigate whether these more complex relations are aligned correctly, we randomly selected 100 instances and evaluated whether the algorithm was justified in mapping PDTB and RST-DT labels for these instances. We found that 95 relations were mapped successfully, while 5 instances were unjustified. In these five cases, the nucleus of a larger RST-DT span matched PDTB's argument, but the annotators did not evaluate the same type of relation. This was largely due to one of the segments (usually Arg2) being part of a larger branch in RST-DT. Even though this segment was the nucleus of that branch, it still blocked a stronger interpretation.

To illustrate this, consider the example in Figure 8. PDTB annotated a CONTRAST relation between segments (b) and (d) (with segments (a) and (c) included as attribution), whereas RST-DT annotated an ELABORATION-ADDITIONAL relation between segments (a-b) and (c-d). The inclusion of the two ATTRIBUTION relations in RST-DT changes the interpretation of the two segments: the relation is focused on the speaker saying something and then adding to that (expressed by an ELAB.-ADDITIONAL relation), rather than on the content of what the speaker is saying (expressed by a CONTRAST relation). Although both the PDTB and the RST-DT annotators selected the same arguments for these relations, the labels might not always match because ATTRIBUTION can (in a subset of cases) "block" an interpretation. To quantify the risk related to intervening RST attributions, we counted the occurrence of attribution relations in the otherwise successfully mapped relations and found that 595 PDTB relations (12%) in total have at least one RST attribution relation in either of the matched segments; 49 (less than 1% of the data) have two or more intervening attributions (like the example given above). Note that out of these 49 instances, only some exhibit the problem of attribution leading to different annotated labels between the two corpora; we therefore decided to include these instances in our analysis.

4.3.1 INSTANCES THAT COULD NOT BE SUCCESSFULLY ALIGNED

24% (1621 instances) of PDTB relations were flagged by at least one of our flags indicating difficult cases. That is, these relations were automatically aligned, but these instances exhibit e.g. violation of the strong nuclearity principle, such that there is a higher chance that the annotated relation labels do not correspond to one another (i.e., annotators had different discourse structures or interpretations in mind).

To provide a more quantitative idea of the cases that were excluded from the mapping, we again randomly selected and manually evaluated 100 instances. We found that 76 items were correctly flagged as unjustified mappings. Out of these, 15 instances consisted of ENTREL labels. Finding many ENTREL labels among the flagged instances is expected, since ENTREL is annotated between

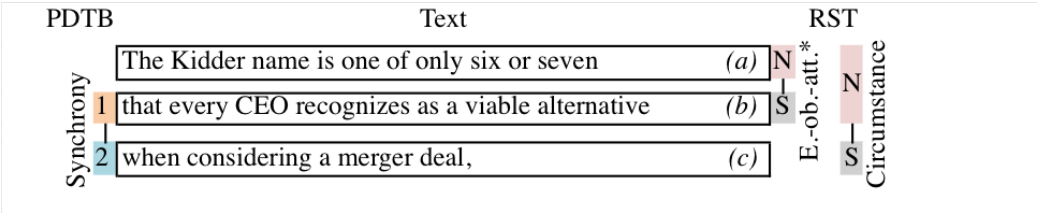

Figure 9: PDTB and RST-DT annotations for a paragraph of wsj_0604. Note: Only the PDTB annotation that is relevant for this example is included. *The full label is ELABORATION-OBJECT-ATTRIBUTE.

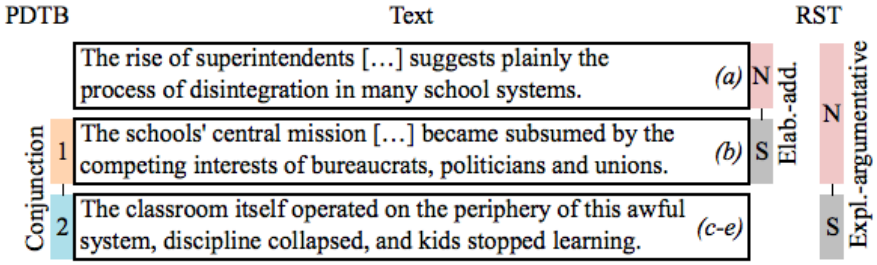

Figure 10: PDTB and RST-DT annotations for a paragraph of wsj_1315. Segment (c-e) is simplified.

adjacent segments that do not have a stronger reading. Often, these occur on the boundary of larger RST-DT spans (that is, discourse segments that function at higher RST-DT tree levels). Looking at the total set of flagged mappings, we find that ENTREL makes up a large portion of this set: it occurs 333 times, which amounts to 19% of all flagged mappings. Additionally, PDTB's NOREL occurs 32 times in the flagged mappings (NOREL is used for adjacent arguments between which no discourse relation holds).

The 76 correctly flagged relations also included four cases of SAME-UNIT relations which could not be mapped automatically because both segments of these SAME-UNIT relations consisted of multiple EDUs. An additional four SAME-UNIT relations were correctly mapped by choosing the relation label below the SAME-UNIT relation. After inspection of additional samples of SAME-UNIT relations for which the label could be unambiguously resolved automatically, we decided to include all of the unambiguously resolvable SAME-UNIT relations into our analyses, while SAME-UNIT relations where both segments have multiple EDUs are excluded.

The remainder of 24 relations do in fact represent valid mappings. This illustrates that our algorithm prefers high reliability of the mapping over full coverage. In more than half of these instances (13 cases), the RST-DT annotation seems to be inconsistent with the Strong Nuclearity principle[6]. Figure 9 illustrates this: PDTB annotated a SYNCHRONY relation between segment (b) and (c). RST-DT annotated a CIRCUMSTANCE relation between segment (a-b) and (c), but segment (a) is in fact annotated as the nucleus. In other words, the nucleus path could not be traced back to PDTB's Arg1, and therefore the relation was flagged. However, the CIRCUMSTANCE relation does not actually hold between *The Kidder name is one of only six or seven* and *when considering a merger deal*; it holds between what PDTB marks as Arg1 and Arg2.

6. Note that difficulties in consistently annotating nuclearity have been pointed out before (see e.g., Stede, 2008)

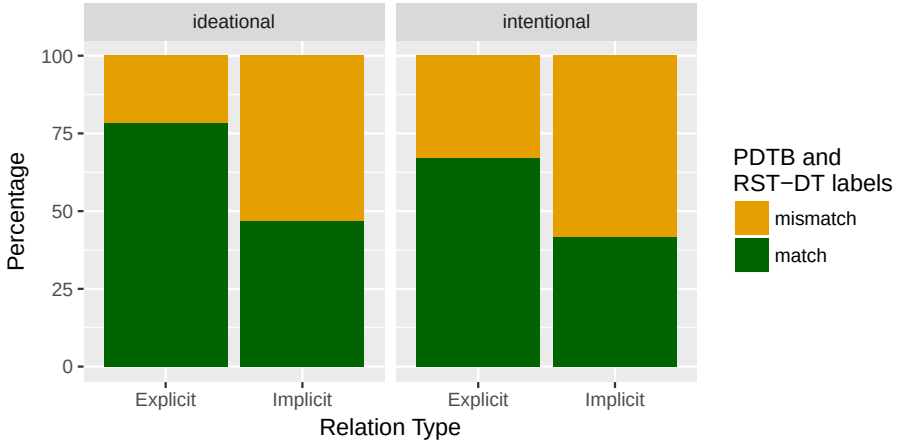

Figure 11: Agreement between theoretically posited and practically found label mappings. The analysis is split out by explicitly marked vs. implicit relations and intentional vs. ideational relations.

Another typical case for automatic flagging by our algorithm occurred in cases where PDTB's annotation constraint for annotating only adjacent implicit relations caused a mismapping: this can happen when two adjacent sentences convey a similar message and are followed by a third sentence which is Arg2. In those cases, the second but not the first sentence has to be annotated as Arg1 in PDTB. Figure 10 presents an example of this: segments (a) and (b) present similar content. PDTB selects segment (b) as Arg1 because of their segmentation principles for implicit relations. RST-DT, however, selected segment (a) as the nucleus and (b) as the satellite. With this structure, the nuclearity path cannot be followed to (b) and the relation is flagged.

We conclude that our manual inspection of a representative sample of instances confirms that our algorithm finds the right alignment between corpora in the majority of cases and thus provides us with reliable data for cross-corpora analysis of relation annotation.

5. Correspondence between mapped relation labels

Our analysis of correspondences between mapped labels is based on a total of 5141 PDTB labels that could be mapped automatically with high confidence (see Section 4.3). For any relation that carried two PDTB labels (in case the annotators thought that both relations held), we selected the label that is most similar to the corresponding RST-DT relation.

A first overview of the results is given in Figure 11, which summarizes the levels of agreement between actual annotations in the discourse corpora with the theoretically posited mapping (the present graph is based on the UniDim mapping; specific differences between mapping proposals are however discussed in more detail in Section 5.1). The figure shows that agreement between the empirically observed and theoretically predicted label correspondences is far from perfect. Second, it shows that there is a big difference in agreement between explicit and implicit relations: for explicitly marked relations, empirical data is consistent with theoretical mappings in more than 70% of cases, while it is consistent with theoretical predictions for implicits in less than 50% of cases. We also show separately here the correspondence in mappings for ideational and intentional relations. Following Hovy & Maier (1995), we classified each RST-DT label as ideational vs. intentional. This

		Temp.		Cont.		Comp.			Expansion							
RST-DT label	PDTB label	Synch.	Asynch.	Cause	Condition	Contrast	Pragm. contr.	Concession	Conjunction	Chosen alt.	Exception	Instantiation	Specification	List	EntRel	Total
Comparison		2	1	1	1	52		2	24					2	2	87
Antithesis		1	3			186	2	37	15	5	2	1			2	254
Elab.-object-attr.			6						7				1		1	15
Background			10	13		13		2	19			2	4		21	84
Circumstance		90	83	39	22	15		4	21				5		15	294

Table 2: Alignment of RST-DT relations for those labels that were identified as theoretically interesting cases of discrepancies between the three mapping proposals.

split is theoretically motivated by the observation that PDTB focuses more on ideational relations; hence we might expect a higher agreement on ideational relations compared to intentional ones. We indeed find such a pattern, especially for explicit relations; however, a closer look at these cases shows that the effect is mostly driven by the fact that RST's CONCESSION relations (which are classified as intentional relations) are often annotated as CONTRAST relations in PDTB. We will get back to a more detailed analysis of the different types of discrepancies and the reasons for why they occur below.

In the following sections, we will first discuss results concerning the discrepancies in theoretical mappings between frameworks as introduced in Section 3, and then analyse in more detail the cases where theoretical predictions differ from empirical observations of labels assigned by the two frameworks in the corpus. As results differ strongly between explicit and implicit relations, we will discuss these relations separately: explicit relations are discussed in Section 5.2 and implicit relations in Section 5.3.

5.1 Analysis of theoretically interesting discrepancies in expected mappings

Section 3 laid out theoretically interesting discrepancies between proposed mappings, relating to highlighting the RST-DT relations COMPARISON, ANTITHESIS, ELABORATION-OBJECT-ATTRIBUTE, BACKGROUND and CIRCUMSTANCE. Table 2 shows the mapping of relations corresponding to these labels. We find that the empirical data confirms the usage of the label ANTITHESIS primarily for contrast relations rather than concessives; we would suggest that any new annotation efforts using the ANTITHESIS label should clarify its intended use in the annotation guidelines; the same holds for the label COMPARISON, which would need to be more clearly defined for future use to avoid misinterpretation. The mapping of relations BACKGROUND and CIRCUMSTANCE to PDTB is difficult, because these labels are more general, used when additional information needs to be provided to allow the reader/listener to understand. Therefore, these segments have a function within the overall goal of the text, but do not seem to stand in a consistent semantic relation to the segment that they are supposed to provide additional context for. We next discuss each of the cases in relation to the proposed mappings in more detail.

Regarding RST-DT's COMPARISON, PDTB does not have a directly corresponding label. Sanders et al. (2018) mapped this label to CONJUNCTION, because the RST-DT annotation manual ex-

plicitly defines COMPARISONS as non-contrastive. However, Bunt & Prasad (2016) mapped it to CONTRAST. Looking at the empirical results (see Table 2), we find that approximately one third of RST-DT COMPARISON instances are annotated as CONJUNCTION, while two thirds (63%) are annotated as CONTRAST in PDTB; we specifically note a high proportion of PDTB CONTRAST.JUXTAPOSITION labels for these instances. This distribution is similar for explicit and implicit instances of COMPARISON. Typical markers for explicit instances are *while*, *but* and *however*. We therefore conclude that RST-DT's COMPARISON should be mapped to both PDTB CONJUNCTION and CONTRAST.

A second area of disagreement was the label ANTITHESIS, which UniDim proposed to map to CONTRAST or CONCESSION, but ISO only to CONCESSION, and OLiA only to CONTRAST. The empirical data clearly shows that the vast majority of ANTITHESIS relations (73%) map onto PDTB's CONTRAST relations, while 15% map to CONCESSION relations. These results are hence best in line with the OLiA and UniDim proposals.

Next, we consider RST-DT's ELABORATION-OBJECT-ATTRIBUTE. OLiA proposed a correspondence with general EXPANSION labels, UniDim proposed a mapping to SPECIFICATION and GENERALIZATION, and ISO proposed a mapping to PDTB's ENTREL. The data only contained 15 instances of ELABORATION-OBJECT-ATTRIBUTE, and the majority of these (seven instances) were mapped to CONJUNCTION, confirming OLiA's mapping. Only one instance was annotated as SPECIFICATION. Six cases were mapped to PDTB's ASYNCHRONOUS label, which none of the frameworks predicted. Closer inspection of these cases revealed that the modifying clause (which makes RST-DT annotate ELABORATION-OBJECT-ATTRIBUTE) often has a temporal aspect to it, as in Example (7), thereby making PDTB more likely to annotate a temporal relation. The closest corresponding labels for ELABORATION-OBJECT-ATTRIBUTE therefore seem to be the more general CONJUNCTION label, as well as the temporal ASYNCHRONOUS type.

(7) Lawmakers would avoid [putting many spending projects into legislation in the first place for fear of the embarrassment] [of having them singled out for a line-item veto] <u>later</u>.
— RST-DT ELAB.-OBJECT-ATTRIBUTE, PDTB ASYNCHRONOUS, wsj_0609

Finally, we consider the labels BACKGROUND and CIRCUMSTANCE. UniDim mapped BACKGROUND to PDTB CONJUNCTION and ASYNCHRONOUS, whereas ISO mapped it to ENTREL. As shown in Table 2, BACKGROUND relations are annotated as a variety of relations: 23% are annotated as PDTB CONJUNCTION, 12% as ASYNCHRONOUS, and 25% as ENTREL, but CAUSE and CONTRAST (both 15%) also occur frequently.

Regarding CIRCUMSTANCE, UniDim mapped it to PDTB CONJUNCTION, SYNCHRONOUS and ASYNCHRONOUS, and ISO only mapped it to SYNCHRONOUS. Table 2 shows that CIRCUMSTANCE is also annotated as a wide range of PDTB relations, indeed including CONJUNCTION (7%), SYNCHRONOUS (31%) and ASYNCHRONOUS (28%), but also CAUSE (13%) and CONDITION (7%). RST-DT's BACKGROUND and CIRCUMSTANCE therefore seem to be more general labels than both UniDim and ISO predicted. Specifically, the RST-DT manual states that these labels are not causal, but in reality, they can be ambiguous. Such cases are often marked with 'as' and 'when' (as in Example (8)), which are known to be ambiguous (see e.g. Asr & Demberg, 2013). It is thus not surprising that RST-DT might annotate such relations as BACKGROUND or CIRCUMSTANCE, whereas the PDTB would annotate these as CAUSE. This could also in part be due to the lack of corresponding label in PDTB.

(8) [But even that niche is under attack] as [several Wall Street firms pulled back from program trading last week under pressure from big investors.]
— RST-DT CIRCUMSTANCE, PDTB CAUSE.REASON, wsj_604

We conclude that the empirical mapping can provide insights for correcting or adjusting the theoretical proposals, for instance in the case of the ELABORATION-OBJECT-ATTRIBUTE relation. For other relations, we find that several proposals capture some of what actually happens in practical annotations, reinforcing the idea that oftentimes, the best mapping we can get between existing relation labels is not a one-to-one mapping, and that some labels, like BACKGROUND and CIRCUMSTANCE capture a textual function which cannot be easily described as a specific semantic relation. We now turn to the findings for the mapping of other labels to see if the two frameworks' annotations are compatible.

5.2 Mapping for explicitly marked relations

Table 3 displays the mapping of PDTB annotations onto RST-DT annotations for explicitly marked discourse relations that occurred more than 30 times in total. In the table, relation mappings which were suggested by all three proposals are considered "expected" mappings and are marked in the table by underlined, bold numbers. Mappings of labels for which at least two of three proposals agree are indicated by underlined numbers. Note that for some labels, several entries in a row or column are marked as "expected". This is because a one-to-one mapping is not always possible, as some labels have multiple matching candidate labels in another framework due to differences in the granularity of distinctions between frameworks (see Section 3).

Generally, we find that deviations from expected mappings are often related to the relation begin marked with an ambiguous connective (e.g., *as*, *when*, *but*, *while*). These ambiguous connectives are in some instances interpreted differently in the two frameworks. Furthermore, we find systematic differences in the operationalization between frameworks, which lead to mismatches in annotations for relations such as LIST. We will next take a closer look at the correspondence between theoretical mappings and empirical matches of labels for each of the major discourse relation classes (temporals, causals, contrastives, additives).

Temporals The results show that most (81%) of the explicitly marked relations that were classified as SYNCHRONOUS by PDTB were tagged as RST-DT TEMPORAL-SAME-TIME or CIRCUMSTANCE. There are however also some cases where annotations deviated from expected mappings for temporal PDTB labels. These include cases where one of RST-DT's causal labels (specifically, EXPLANATION-ARGUMENTATIVE or CONSEQUENCE) was annotated. Closer inspection revealed that frequent connectives in these relations, which did not receive a temporal sense label in RST-DT, were *as* and *when*. These connectives are known to be ambiguous markers (see e.g. Asr & Demberg, 2013), and are frequent among temporal relations such as CIRCUMSTANCE and TEMPORAL-SAME-TIME in RST-DT. We hence find that there are some instances containing these ambiguous connectives which could not be consistently disambiguated between frameworks, with one framework labelling these instances as temporal and the other as causal. We will analyse the annotation of ambiguous connectives in more detail in Section 5.2.1

PDTB's temporal ASYNCHRONOUS relations generally also map well to their corresponding RST-DT classes (79%). 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

		Temp.		Cont.		Comp.		Expansion			
RST-DT label	PDTB label	Synch.	Asynch.	Cause	Condition	Contrast	Concession	Conjunction	Instantiation	List	Total
Temp.-same-time		**63**	1	1	1	4		5			75
Temporal-after			**48**			1		2			54
Sequence		2	**29**	1		1		19			52
Circumstance		89	79	29	21	7	4	10			259
Result		1	3	**30**		2		5			41
Consequence		5	6	**45**		4	1	16		1	95
Explanation-arg.		6		**39**		7	2	1			57
Reason			2	**67**							72
Condition		5	13	1	**104**	2	1	1			182
Contrast		1	1			**160**	23	17			208
Concession		5	4		3	101	**53**	4			182
Antithesis		1	3			170	37	10			243
Comparison		2	1			26	2	**9**			41
Elaboration-add.		4	3	1		30	8	122	3	3	192
Example		1				1		3	**29**		35
List		2	4	1		17	1	303		**47**	377
Total		187	197	215	129	533	132	527	32	51	

Table 3: Alignment of explicit discourse relation classes for which n >30. Numbers indicate how many instances occurred in our high-confidence mapping; values are represented as colors. Underlined, bold numbers indicate the mapping predicted to all three proposals; numbers that are only underlined (not bold) are predicted by two of three proposals.

inconsistencies in disambiguation of this marker, or could be more systematically related to RST-DT annotating the intention in subjective relations, while PDTB annotations stay closer to the semantic relation (Scholman & Demberg, 2017).

For RST-DT's SEQUENCE class, we find that a substantial portion is annotated as PDTB's CONJUNCTION (37%), i.e. annotations by the different frameworks disagree for these instances whether the relation is temporal or not. This was not predicted by any of the mapping proposals. A closer look at these instances reveals that they are mostly marked by the underspecified marker *and*, and can indeed have a somewhat temporal aspect to them, as Example (9) illustrates.

(9) [The farmer leaves] <u>and</u> [the naczelnik shuts his door.]
— RST-DT SEQUENCE, PDTB CONJUNCTION, wsj_1146

Causals Explicit causal and conditional PDTB relation labels generally map well onto causal and conditional RST-DT labels; a finding that can be attributed to the relative definiteness of causal markers (we will see a different results for implicit relations in Section 5.3). RST-DT distinguishes more types of causal relations, which results in causal PDTB relations being distributed among the various causal RST-DT classes. Unexpected mappings – e.g., PDTB causals and conditionals annotated as RST-DT's CIRCUMSTANCE (12% and 16%, respectively) – were found to occur again for instances marked by the ambiguous connectives *as* and *when*, respectively.

Contrastives The majority of PDTB's CONTRAST relations was mapped to RST-DT's CONTRAST and ANTITHESIS relations (62%), as expected. However, we also found that a substantial portion (19%) of PDTB's CONTRAST relations are annotated as RST-DT's CONCESSION; and some also as RST-DT's ELABORATION-ADDITIONAL (6%). These cases were often marked by the connective *but*, which is an ambiguous connective.

A closer look at the subtypes of PDTB CONCESSION reveals that relations annotated as PDTB's CONCESSION.EXPECTATION map quite well (54%) onto RST-DT's CONCESSION relations, while CONCESSION.CONTRA-EXPECTATION relations are often annotated as CONTRAST in RST, especially when marked with the connective *but*. We note that the distinction between concession and contrast relations is known to be difficult in discourse relation annotation (Robaldo & Miltsakaki, 2014). It is possible that the observed differences stem from differences in interpretation between annotators, and slight biases in the frameworks. Overall, PDTB has a stronger bias than RST-DT towards assigning the CONTRAST label: the majority of three RST-DT relational labels, namely CONTRAST, ANTITHESIS and CONCESSION, are mapped to PDTB's CONTRAST.

Additives Finally, looking at PDTB's EXPANSION relations, we find that a majority of relations annotated as PDTB's CONJUNCTION is annotated as RST-DT'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-DT LIST stems from the fact that PDTB annotation guidelines say that lists have to be "defined in the prior discourse" (Prasad et al., 2008, p.37), as *stocks are up* does in Example (10). Unannounced lists cannot be annotated as LIST relations in PDTB. Such a criterion is however not applied in RST-DT. We find that PDTB's LIST and INSTANTIATION relations map well onto RST-DT's LIST and EXAMPLE relations respectively.

(10) Those stocks are up because [their earnings are up] <u>and</u> [their dividends are up.]
— RST-DT LIST, PDTB LIST, wsj_0681

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 contain the connectives *but* or *while*, again indicating that connective disambiguation may not always be consistent between frameworks.

A final interesting observation regards the annotation of the connective *unless*: these instances are annotated as as ALTERNATIVE.DISJUNCTIVE relations in PDTB, but as CONDITION in RST-DT (note that RST-DT does not have a label corresponding to PDTB's DISJUNCTIVE). Although these relation types might seem to differ greatly, in reality they can be compatible. Consider Example (11), which was annotated as PDTB ALTERNATIVE.DISJUNCTIVE and RST-DT CONDITION. This is a negative conditional relation, but also can be considered disjunctive: two possibilities are evoked but only one of them can hold.

(11) [Essentially, he can't make any hostile moves,] <u>unless</u> [he makes a tender offer at least $300 a share.]
— RST-DT CONDITION, PDTB ALTERNATIVE.DISJUNCTIVE, wsj_1305

More generally, the finding that there are differences in the interpretation of ambiguous connectives may seem somewhat surprising, since disambiguation of these connectives is one of the prime motivation for annotating explicitly marked discourse relations in the first place. We will therefore next analyse the annotation of connectives in more detail, before turning to the mapping of implicit relations.

5.2.1 ANALYSIS OF ANNOTATION FOR AMBIGUOUS CONNECTIVES

Given that the disagreements between the PDTB and RST-DT can be related to different interpretations of specific ambiguous connectives (as discussed in Section 5.2), we studied the agreement on annotating ambiguous connectives in more detail. After all, it is not surprising that annotators can reach high agreement on relations with an unambiguous connective such as *if* for CONDITION; given the empirical findings from previous annotation work in English, this could be done automatically in the future. Rather, the value of additional manual annotation comes from disambiguating between relations when the connective can mark different relations (or when a relation is not explicitly marked, see Section 5.3).

Methods To investigate the agreement on connectives, we tested the independence of PDTB and RST-DT annotations by calculating a separate χ^2 test for each connective. For connectives where the distribution violated the assumptions of the χ^2 test (less than 5 expected observations in a cell), we instead used the non-parametric Fisher's exact test. These analyses reveal whether the annotations agreed more with each other than can be expected based on the distribution of the connectives in the data. The results will provide insight into whether the content of the discourse relation arguments had been taken into account for the actual annotations of the relation instances, or whether the distinction is potentially too difficult or too subtle for human annotators to make reliably. Note that this analysis is more strict than the usual kappa for inter-annotator agreement, because we use 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 are not normally known.

We report some representative results for connectives where the null hypothesis of PDTB and RST-DT annotations for a connective being independent from one another. Here, independence would mean that the content of the actual content of an instance is not predictive of which label was chosen by the one vs. the other framework; in the ideal case, we would expect strong non-independence: given a pair of segments, the labels chosen by one and the other framework should correspond to one another and hence be deterministic not random.

Results The connective *while* is an example of a case where labels corresponded well to one another, i.e., where the null hypothesis of labels being independent could be rejected with high confidence: we find that annotators could reliably distinguish between the TEMPORAL.SYNCHRONOUS vs. CONTRAST / COMPARISON reading of *while*; the annotations from the two frameworks almost always agreed on the reading ($p < 0.0001$).

There are however also connectives for which we find that the observed distributions are similar to random distributions. An example for this are connectives that are ambiguous between more similar discourse relations (CONTRAST vs. CONCESSION), such as *but*, *although* and *however*. According to a χ^2 test for *but* and Fisher's exact tests for *although* and *however*, the meaning distributions did not significantly differ from a random distribution of these sense labels (given the marginals). Calculating κ 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) corresponds to $\kappa < 0.1$ for these relation label distinctions between frameworks.

To summarize, we find that there are some ambiguous connectives for which manual annotation from the two frameworks reliably agreed on how the connective should be disambiguated and hence provide valuable additional information. This was mostly the case for when the alternative readings of the connective strongly differ from one another. However, we found no conclusive evidence

that most subtle distinctions could be made reliably – on these cases, annotations from the two frameworks often don't agree with one another more than would be expected by random assignment of labels that a connective can occur with (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.

5.3 Mapping of implicit discourse relations

The overall picture of annotation agreement between the two frameworks looks a lot more problematic for implicit than for explicit relations. To get an idea of what underlying causes these differences stem from, we decided to provide two perspectives on the results: once from the PDTB view, which provides an overview of how RST labels are distributed for a given PDTB relation (Table 4a) and once from the RST view, showing how the PDTB labels are distributed for a given RST relation (Table 4b). To keep tables readable, we only include those labels that occurred more than twenty times in the data.

Mapping of relation annotations for implicit relations, as seen from the PDTB perspective, is shown in Table 4a. Expected correspondences in the table are again indicated by underlined numbers. The colours in the table indicate the percentage of correspondence to RST-DT labels, with darker shades indicating a higher proportion of instances with a certain PDTB label falling into that RST-DT category. For example, 7 instances of PDTB TEMPORAL.ASYNCHRONOUS relations are annotated as RST-DT BACKGROUND. As these 7 instances represent more than 10% of the data on TEMPORAL ASYNCHRONOUS relations, the number is shaded in light green. On the other hand, the 7 counts of COMPARISON.CONTRAST relations labelled as RST's CONSEQUENCE represent less than 10% of the 219 PDTB COMPARISON.CONTRAST relations; the entry is therefore not shaded.

A first striking observation is that the agreement between frameworks is a lot worse than for explicit relations. While we saw a diagonal line of shaded cells that largely overlapped with theoretically expected correspondences, we can see that a substantial proportion of instances from almost all PDTB classes were annotated as RST-DT's ELABORATION-ADDITIONAL.

Table 4b shows the alignment of implicit relations from the RST-DT perspective. Many cells here are shaded green, which indicates that the annotations for many of the RST-DT relations have a wide variety of PDTB labels.

Generally, a stronger relation (biasing away from annotating simple additive labels) tended to be chosen by PDTB annotators; this bias can likely be linked to the PDTB annotation instructions of first inserting an explicit connective that fits the relation, and then assigning a label. Wherever the RST-DT annotators chose a label other than ELABORATION-ADDITIONAL (the predominant label assigned to the implicit cases), the annotations matched with their PDTB equivalents relatively well. We will next analyse in more detail the mapping results for each of the four major relation classes.

Temporals Temporal relations were not consistently identified between frameworks. Only roughly one third of PDTB's ASYNCHRONOUS.PRECEDENCE relations had the expected RST label SEQUENCE, and most of PDTB's ASYNCHRONOUS.SUCCESSION relations were labelled as ELABORATION-ADDITIONAL in RST-DT. The RST-DT TEMPORAL-AFTER label is very rarely annotated among implicit relations.

Causals For PDTB's CAUSE.REASON relations, less than 40% of instances were annotated as one of the expected causal classes by RST-DT annotators (expected classes were EXPLANATION-ARGUMENTATIVE, REASON, EVIDENCE or INTERPRETATION), and only a very small percentage of PDTB's CAUSE.RESULT were annotated as CONSEQUENCE, RESULT or CAUSE-RESULT relations in RST. Instead, most of these relations are annotated as ELABORATION.ADDITIONAL, showing that PDTB annotators tended to choose a "stronger" label than RST annotators for these instances.

RST-DT's causal relations CONSEQUENCE and REASON map relatively well onto PDTB's causal relations. However, other causal RST-DT 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-DT focuses more on the intentional level. Scholman & Demberg (2017) show that often two functions can be identified in these specific relations, both illustrated in Example (12): a segment can provide an example or a specification of something mentioned in the first argument (in this case, the track record), as well as providing evidence for a previously stated claim (e.g., that the track record stands out). These double functions are reflected in the mapping.

(12) [In the cornucopia of go-go apples, the Fuji's track record stands out.] [During the past 15 years, it has gone from almost zilch to some 50% of Japan's market]
— RST-DT EVIDENCE, PDTB SPECIFICATION, wsj_1128

Contrastives Only a minority of PDTB CONTRAST relations was also labelled as a contrastive relation in RST-DT (17% for RST-DT CONTRAST, 10% for COMPARISON). Instead, the majority maps to ELABORATION-ADDITIONAL and even LIST labels.

From the perspective of RST-DT'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).

Additives One PDTB implicit relation that matches well with RST-DT annotation is LIST, for which 84% of instances were annotated as RST-DT's LIST relation. For RST-DT'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.

Taking a look at PDTB ENTREL labels, we can see that these are predominantly annotated as ELABORATION-ADDITIONAL and LIST in RST-DT. Nevertheless, we also observe a number of different annotations on the RST side. We analysed a randomly selected subset of these instances and found that a common reason for the richer RST-DT labels is that these EntRel relations are in fact high-level relations in RST-DT. The stronger interpretation is in those cases often due to the effect of additional content (outside the nucleus path), which facilitates the stronger interpretations. In our view, it is arguable whether EntRel labels should really be considered to be mapped to RST-DT labels, as the task given to the annotators differed markedly for these instances: in the PDTB case, the annotator task was to label the relation holding between two adjacent sentences, whereas the task of the RST-DT annotators was to join high-level discourse segments and describe their relation.

RST-DT's COMMENT relation shows an almost uniform distribution across PDTB labels, consistent with other communicative functions that are not represented in the PDTB annotation scheme, such as BACKGROUND and CIRCUMSTANCE, which have been discussed in Section 5.1.

Finally, we observe that RST-DT's ELABORATION-GENERAL-SPECIFIC relations are often labelled as PDTB's RESTATEMENT (55%). This correspondence was predicted according to the expected mappings. 21% of instances were labelled as INSTANTIATION, which can be attributed to the subtlety of the distinction between these two labels.

Discussion on correspondence of annotations for implicit relations We find that the level of agreement in labels for implicit relations is a lot lower than for explicit ones. 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-DT 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.

6. Discussion

In the current paper, we evaluated how well theoretical proposals for mapping discourse relational labels correspond to the mapping between existing RST-DT and PDTB 2.0 annotations. Some of the most important findings include: (i) RST-DT and PDTB agree more on explicit than implicit relations. The differences in annotation procedure and operationalization contribute greatly to this. (ii) Among explicit relations, the ambiguity of connectives is a major source of disagreement. (iii) Certain relations, such as CONCESSION and CONTRAST, are inherently difficult to tease apart; the PDTB uses the CONTRAST label more often while RST-DT more frequently uses the CONCESSION label. These differing label usages then lead to discrepancies in the mapping.

The study consisted of three main steps: alignment of relations, evaluation of theoretically interesting discrepancies between the proposals, and evaluation of the mapped labels for explicit and implicit relations. We will discuss each one in turn.

Automatic alignment In order to evaluate existing annotations, we proposed an automatic mapping algorithm for PDTB and RST-DT discourse relation annotations and applied it to a segment of the WSJ corpus that contains annotations from both frameworks. The algorithm proposed in this article allowed us to align 76% of PDTB discourse relation annotations to corresponding RST annotations. Our manual error analysis shows that the alignment algorithm is highly accurate; it also correctly identifies instances where annotations cannot be aligned due to more fundamental differences in how the discourse is analysed.

Our study highlights the importance of discourse segmentation. Segmentation has a strong effect on determining the scope and argument structure of a discourse relation (see also Hoek et al., 2017). The differences in segmentation may hold interesting insights about effects of operationalization of discourse segmentation on discourse annotation, which could be explored in future work to refine annotation processes both for manual annotation and automatic processing.

0 10 50% 100

RST-DT label \ PDTB label	Temp. Asynch.	Cont. Cause	Comp. Contrast	Expansion Conjunction	Alternative	Instantiation	Restatement	List	EntRel
Sequence	**16**	1	2	4					3
Background	7	12	10	16		2	4		21
Circumstance	**1**	10	8	11			6		15
Consequence	4	**16**	7	11		2	2	1	4
Evidence		**12**	2	12	2	29	28		6
Explanation-arg.	2	**114**	16	16	2	35	51	1	22
Reason	1	**20**	1	3		1	2		3
Result	1	**12**	1	4			3	1	5
Evaluation		7	9	11		2	1		12
Interpretation		**16**	3	9		2	8		8
Contrast	1	2	**35**	2				2	3
Comparison		1	24	14	4			2	2
Antithesis			15	5	2	1			2
Elaboration-add.	29	168	73	221	10	36	151	7	266
Example		9	1	6		**64**	21		2
Elab.-gen.-spec.		8		11		17	**44**		18
List	6	24	29	120	2	6	13	**74**	30
Restatement		5		2	1	2	**12**		1
Comment		16	9	8			4		8
Total	75	499	277	511	26	202	406	88	463

(a) 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-DT class.

RST-DT label \ PDTB label	Temp. Asynch.	Cont. Cause	Comp. Contrast	Expansion Conj.	Altern.	Instant.	Restat.	List	EntRel	Total
Sequence	**16**	1	2	4					3	27
Background	7	12	10	16		2	4		21	72
Circumstance	**1**	10	8	11			6		15	51
Consequence	4	**16**	7	11		2	2	1	4	49
Evidence		**12**	2	12	2	29	28		6	92
Explanation-arg.	2	**114**	16	16		35	51	1	22	267
Reason	1	**20**	1	3		1	2		3	31
Result	1	**12**	1	4			3	1	5	28
Evaluation		7	9	11		2	1		12	46
Interpretation		**16**	3	9		2	8		8	49
Contrast	1	2	**35**	2				2	3	56
Comparison		1	24	14	4			2	2	44
Antithesis			15	5	2	1			2	26
Elaboration-add.	29	168	73	221	10	36	151	7	266	991
Example	1	9	1	6		**64**	21		2	106
Elab.-gen.-spec.		8		11		17	**44**		18	99
List	6	24	29	120	2	6	13	**74**	30	311
Restatement		5		2	1	2	**12**		1	25
Comment		16	9	8			4		8	49

(b) Colours encode percentage agreement from RST-DT perspective, i.e. darker colours show that most instances of a RST-DT relation type occurred in that specific PDTB class.

Table 4: Alignment of implicit discourse relation classes for which n >20. Numbers indicate how many instances occurred in our high-confidence mapping.

Evaluation of mapping proposals As a result of the annotation alignment, 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 that were determined based on three recent proposals for mapping discourse relations onto one another. Our aim in evaluating three different mapping proposals was not to identify the "best" proposal; rather, the comparison between proposed correspondences and empirical coocurrences of annotated labels is helpful for achieving a deeper understanding of how certain definitions in the annotation guidelines were applied in practical annotation, and can help to decide between alternative proposals for mappings. The observed mismatches between the proposals can furthermore be used for clarifying annotation guidelines in future annotation projects; for example, the three proposals differed in their interpretation of RST-DT's ANTITHESIS label, which indicates that the definition of this label could be expanded on.

We found high numbers of disagreement between observed and expected annotations for those relations that did not have a direct correspondence in the other scheme, including labels that seem to be often used as blankets for ambiguous cases. Examples of such relations include RST's BACKGROUND and CIRCUMSTANCE relations. The three proposals treated these relations differently from each other based on the definitions in the annotation manual, but generally, they were mapped to temporal or additive labels. However, the empirical mapping showed that many of these relations were also annotated as causals or even contrastives in PDTB. We see two possible explanations for what causes this: 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 that is not reflected in the PDTB scheme, and should be considered to be added to relation schemes, again possibly by annotating both the ideational and intentional functions separately.

The current work addressed efforts in the community to create a standard typology of relations (consider frameworks such as the PDTB and RST-DT, but also international standards and intermediary approaches, such as those proposed by Benamara & Taboada, 2015; Bunt & Prasad, 2016; Chiarcos, 2014; Hovy & Maier, 1995). This paper has highlighted that there are significant differences in the predicted correspondences of these proposals, and that we are still far from the goal of creating a standard relational inventory. The European COST initiative TextLink, which ran from April 2014 until April 2018, was aimed at unifying the numerous, scattered linguistic resources on discourse structure. The deliverables of this COST Action include a database of discourse-annotated corpora, a database of lexicons, and a web portal that provides access to resources and a small suite of search, visualization and dissemination capabilities. One of TextLink's original goals was to create a single taxonomy that could be used in future annotation efforts. This proved to be inconceivable. Instead, the Unifying Dimensions approach (Sanders et al., 2018) was created as a final deliverable.The results from the current study show that the UniDim approach was relatively successful in mapping between PDTB and RST-DT, but our empirical analysis also shows that framework-specific guidelines and operationalizations can cause mismappings that likely no intermediate approach can successfully deal with – the annotations done in the different frameworks simply do not always correspond to one another. The present article provides a detailed analysis that can help researchers to be aware of what types of relations they might find if searching the interoperable corpus originally annotated in one framework using a specific label from another framework. It also points out ways in which the research community can design the annotation guidelines in future annotation to reduce such discrepancies.

Evaluation of mapped explicit and implicit relations After evaluating the theoretical discrepancies between the proposals, we looked at the mapped data for explicit and implicit relations separately. The most striking observations were a lower than expected level of agreement on annotations for implicit relations, and low agreement on more fine-grained distinctions for explicitly marked relations. In order to get more insight into the issue of difference in agreement between implicit and explicit relations, we recommend that future annotation efforts (corpus annotation as well as other tasks) report agreement on implicit and explicit relations separately. While the rate of non-correspondences does seem problematic, we were able to identify several patterns that lead to these observed disagreements. Specifically, many of the differences can be traced back to different operationalizations employed during annotation, as well as to the different goals of PDTB 2.0 vs. RST-DT annotation.

First, the operationalization of discourse annotations may have a strong effect on the resulting annotations: in the PDTB annotation process, annotators were asked to annotate implicit relations by first identifying a discourse connective that would fit the relation, and then in a second step annotate the relation sense. It seems that this practice encourages annotators to assign more specific relation labels to implicit relations than RST-DT's annotation procedure does. We therefore find that most implicit relations receive the RST-DT label ELABORATION-ADDITIONAL and a more specific PDTB label. For future annotation efforts, it is important that this consequence of annotation operationalization is taken into account. We cannot determine whether RST-DT annotators relied too heavily on the absence of a connective, which may have biased them to annotate the ELABORATION-ADDITIONAL relation more often, or whether the insertion of connectives as a task made some of the interpretations of relations stronger than they were without the connective, i.e. whether the operationalization to insert a connective may have changed the inferred relation in some cases. To answer this question, we suggest systematic annotation experiments for measuring effect of annotation instructions.

Second, the framework-specific goals have led to a focus on different levels of analysis of discourse relations, namely the *ideational* and the *intentional* level. Ideational relations describe the semantic relation between the information conveyed in the consecutive elements of a coherent discourse (cf. Moore & Pollack, 1992). *Intentional relations* on the other hand involve the writer's attempts to affect the addressee's beliefs, attitudes, desires etc. by means of language (cf. Hovy & Maier 1995; see also Crible & Degand 2017; Redeker 1990). This distinction is relevant for the data used in the current study, because the goals of RST-DT annotations and PDTB 2.0 annotations differ with respect to these functions. While annotators in RST-DT were instructed to annotate the writer's goal or intended effect of each segment of a text with respect to the neighbouring segments, PDTB annotators were asked to assess the relation between relational arguments, with a strong focus on the role of connectives (lexically-driven approach).

The analysed empirical mappings provide support for the idea to annotate several levels of discourse relational arguments when appropriate. For instance, a segment can be an example for something that was said in the other segment, but it may at the same time serve as evidence for a claim (see also Carston, 1993; Blakemore, 1997). Some of these patterns are systematic and go beyond the PDTB approach of allowing to annotate multiple labels for the same relation: PDTB annotators were not asked to systematically try to annotate all relations that hold. We conclude that future research should explore in more depth whether it is possible to devise an annotation procedure that allows researchers to identify both functions in a systematic way (see also Crible & Degand, 2017).

Finally, regarding the annotation of explicit relations, we found that disagreements are often related to ambiguous connectives such as *as*, *but* and *while*. We analysed whether the annotations of ambiguous connectives agreed more with each other than can be expected based on the distribution of connectives in the data. We found that the PDTB and RST corpus annotations agreed well for relations marked by connectives that can mark very different types of relations (e.g., *while* can mark a causal or a temporal relation), but they disagreed often on annotation of connectives that mark similar types of relations (e.g., *but* can mark a contrastive or a concessive relation). We believe that these cases warrant further study in order to better understand why the disagreements occur, and what the implications should be (e.g., a less fine-grained distinction among discourse relations if the present distinction cannot be made reliably, or an improved operationalization of the annotation process in order to achieve more agreement?).

Future directions The mapped annotations will be made available online so that other researchers can profit from the aligned corpus. 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 cases that could not be aligned due to differences in structuring the discourse, and cases where different labels were chosen, 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.

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 defining future annotation guidelines. 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). Our detailed results can also inform the ongoing discussion on identifying a set of labels for discourse relation annotation, which has been a long-lasting issue causing a lot of controversy in the literature.

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. The classification of implicit discourse relations has received the bulk of the attention and work, given that classification of explicit relations was found to be relatively easy and accurate (Pitler et al., 2008). Implicit discourse relation classification has recently also been 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? The alignment data can also be used directly to select easy vs. difficult relation instances for training and evaluation of automatic relation identification systems.

Finally, we would like to emphasize that some of the methodological decisions for the present mapping are specific to the two exact frameworks we worked with, RST-DT and PDTB2.0. Other instances of RST-style annotation may treat nuclearity differently, in which case some of the assumptions of our alignment algorithm would not necessarily generalize to those annotations. Furthermore, PDTB-style annotation also differs between languages; different PDTB resources may use different relation inventories; nevertheless, we would expect that some of the fundamental observations we made (such as the effect of operationalization like the usage of implicit connectives during annotation) would transfer to those other PDTB-style resources.

References

Al-Saif, A., & Markert, K. (2010). The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)* (pp. 2046–2053). Valletta, Malta.

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

Benamara, F., & Taboada, M. (2015). Mapping different rhetorical relation annotations: A proposal. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics,* SEM* (pp. 147–152).

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

Bunt, H., & Prasad, R. (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)* (pp. 45–54).

Cardoso, P. C., Maziero, E. G., Jorge, M. L., Seno, E. M., Di Felippo, A., Rino, L. H., Nunes, M. G., & Pardo, T. A. (2011). CSTnews – a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting* (pp. 88–105).

Carlson, L., & Marcu, D. (2001). *Discourse tagging reference manual*.

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

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

Chiarcos, C. (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)* (pp. 4569–4577).

Crible, L., & Degand, L. (2017). Reliability vs. granularity in discourse annotation: What is the trade-off? *Corpus Linguistics and Linguistic Theory*, .

Das, D., & Taboada, M. (2018). Rst signalling corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, *52*, 149–184.

Hoek, J., Evers-Vermeul, J., & Sanders, T. J. (2017). Segmenting discourse: Incorporating interpretation into segmentation? *Corpus Linguistics and Linguistic Theory*, *Advance online publication*.

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

Iruskieta, M., Aranzabe, M. J., de Ilarraza, A. D., Gonzalez, I., Lersundi, M., & de Lacalle, O. L. (2013). The RST Basque TreeBank: an online search interface to check rhetorical relations. In *4th workshop RST and Discourse Studies* (pp. 40–49).

Jansen, P., Surdeanu, M., & Clark, P. (2014). Discourse complements lexical semantics for non-factoid answer reranking. In *ACL (1)* (pp. 977–986).

Lee, A., Prasad, R., Joshi, A., Dinesh, N., & Webber, B. (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* (pp. 79–90).

Lee, A., Prasad, R., Joshi, A., & Webber, B. (2008). Departures from tree structures in discourse: Shared arguments in the penn discourse treebank. In *Proceedings of the Constraints in Discourse III Workshop* (pp. 61–68).

Liu, Y., Li, S., Zhang, X., & Sui, Z. (2016). Implicit discourse relation classification via multi-task neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)* (pp. 2750–2756).

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

Marcu, D. (2000). *The theory and practice of discourse parsing and summarization*. MIT press.

Meyer, T., & Popescu-Belis, A. (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)* (pp. 129–138). Association for Computational Linguistics.

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

Oza, U., Prasad, R., Kolachina, S., Sharma, D. M., & Joshi, A. (2009). The Hindi discourse relation bank. In *Proceedings of the third Linguistic Annotation Workshop (LAW)* (pp. 158–161). Association for Computational Linguistics.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., & Joshi, A. K. (2008). Easily identifiable discourse relations. Technical report.

Polakova, L., Mirovsky, J., & Synkova, P. (2017). Signalling implicit relations: A pdtb-rst comparison. *Dialogue & Discourse*, *8*, 225–248.

Popescu-Belis, A. (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* (p. 16).

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

Prasad, R., Forbes-Riley, K., & Lee, A. (2017). Towards full text shallow discourse relation annotation: Experiments with cross-paragraph implicit relations in the PDTB. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue* (pp. 7–16).

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

Prasad, R., Webber, B., & Joshi, A. (2014a). Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, .

Prasad, R., Webber, B., & Joshi, A. (2014b). Reflections on the Penn Discourse TreeBank, Comparable Corpora, and Complementary Annotation. *Computational Linguistics*, *40*, 921–950.

Prasad, R., Webber, B., & Lee, A. (2018). Discourse annotation in the PDTB: The next generation. In *Proceedings 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation* (pp. 87–97).

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

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

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

Sanders, T. J. M., Demberg, V., Hoek, J., Scholman, M. C. J., Torabi Asr, F., Zufferey, S., & Evers-Vermeul, J. (2018). Unifying dimensions in discourse relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, *ahead of print*, 1–71.

Sanders, T. J. M., & Spooren, W. (1999). Communicative intentions and coherence relations. In W. Bublits, U. Lenk, & E. Ventola (Eds.), *Coherence in Spoken and Written Discourse* (pp. 235–250). John Benjamins.

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

Scheffler, T., & Stede, M. (2016a). Adding semantic relations to a large-coverage connective lexicon of german. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Scheffler, T., & Stede, M. (2016b). Mapping PDTB-style connective annotation to RST-style discourse annotation. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*.

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

Sharp, R., Jansen, P., Surdeanu, M., & Clark, P. (2015). Spinning straw into gold: Using free text to train monolingual alignment models for non-factoid question answering. In *HLT-NAACL* (pp. 231–237).

Somasundaran, S., Namata, G., Wiebe, J., & Getoor, L. (2009). Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1* (pp. 170–179). Association for Computational Linguistics.

Stede, M. (2008). RST revisited: Disentangling nuclearity. *Subordinationversus Coordinationin Sentence and Text*, (pp. 33–59).

Stede, M., & Neumann, A. (2014a). Potsdam Commentary Corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 925–929).

Stede, M., & Neumann, A. (2014b). Potsdam Commentary Corpus 2.0: Annotation for discourse research. In *LREC* (pp. 925–929).

Taboada, M., & Mann, W. C. (2006). Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies*, *8*, 423–459.

Tonelli, S., Riccardi, G., Prasad, R., & Joshi, A. K. (2010). Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)* (pp. 2084–2090). Marrakesh, Marocco.

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

Xue, N., Ng, H. T., Pradhan, S., Prasad, R., Bryant, C., & Rutherford, A. (2015). The CoNLL-2015 shared task on shallow discourse parsing. In *CoNLL Shared Task* (pp. 1–16).

Xue, N., Ng, H. T., Rutherford, A., Webber, B., Wang, C., & Wang, H. (2016). CoNLL 2016 shared task on multilingual shallow discourse parsing. *Proceedings of the CoNLL-16 shared task*, (pp. 1–19).

Zhou, L., Li, B., Gao, W., Wei, Z., & Wong, K.-F. (2011). Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 162–171). Association for Computational Linguistics.

Zhou, Y., & Xue, N. (2015). The Chinese Discourse TreeBank: A Chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, *49*, 397–431.

Zirn, C., Niepert, M., Stuckenschmidt, H., & Strube, M. (2011). Fine-grained sentiment analysis with structural features. In *IJCNLP* (pp. 336–344).

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

Appendix A. PDTB 2.0 and RST-DT tagsets

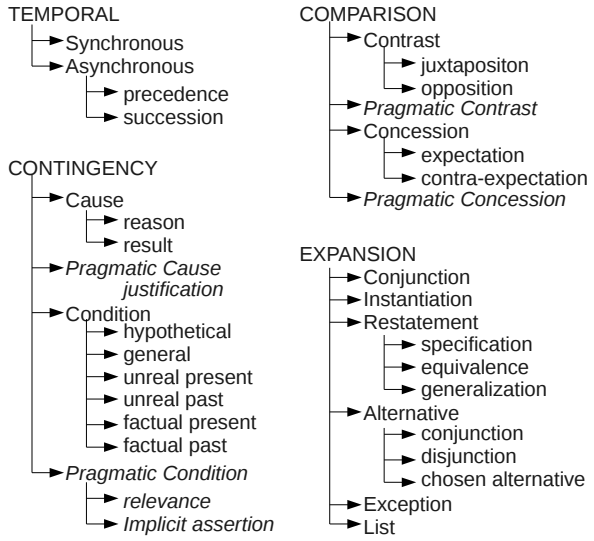

Figure 12: Hierarchy of relation senses in PDTB (Prasad et al., 2008).

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

Figure 13: Tagset of relation senses in RST-DT (Carlson & Marcu, 2001).

Appendix B. Information on the OLiA repository

The OLiA format is a machine-readable format that uses several intermediary hierarchies of relations. As a result of its format, it cannot be easily represented compactly as a tagset or something similar. The OLiA website[7] does, however, provide two figures to illustrate the approach. These figures are replicated here. Figure 14 presents an example mapping of several instances annotated by both the PDTB and RST-DT. Figure 15 presents a more detailed example mapping of the CONDITION relation labels in the PDTB and RST-DT.

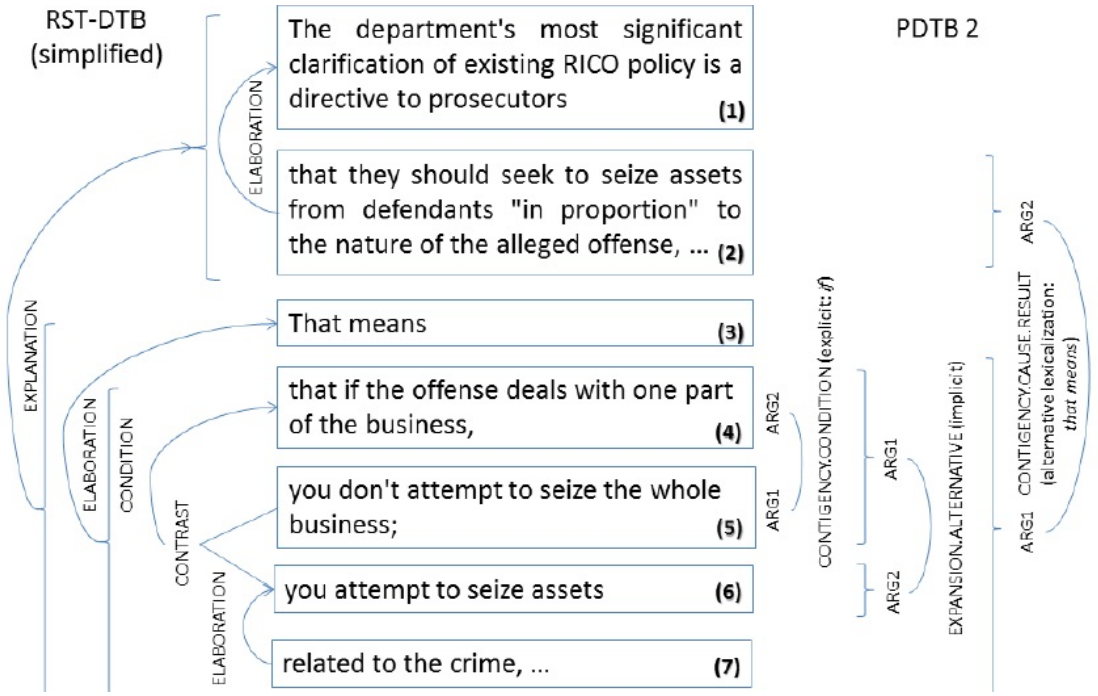

Figure 14: Example mapping of the relations in a text fragment carrying RST-DT and PDTB 2.0 annotations.

PDTB Annotation & Linking	pdtb:contingency.condition.general a pdtb:GeneralCondition
	pdtb:GeneralCondition ⊑ olia:GeneralCondition_PDTB
OLiA Reference Model	olia:GeneralCondition_PDTB ⊑ **olia:SemanticCondition**
	olia:SemanticCondition ⊑ olia:Condition
	olia:Condition ⊑ olia:Contingency
	olia:Contingency ⊑ olia:CoherenceRelation
	olia:CoherenceRelation ⊑ olia:DiscourseRelation
	olia:ParatacticDiscourseRelation ⊑ olia:DiscourseStructuralPattern
	olia:DiscourseStructuralPattern ⊑ olia:DiscourseRelation
RST Linking & Annotation	rst:Condition ⊑ **olia:SemanticCondition** ⊓ olia:ParatacticDiscourseRelation
	rst:condition a rst:Condition

Figure 15: Example mapping of the CONDITION relation labels in PDTB and RST-DT using OLiA.

7. http://www.acoli.informatik.uni-frankfurt.de/resources/discourse/

Appendix C. Unifying Dimension definitions

Dimension	Values	Meaning	Typical connectives
Polarity	Positive	Propositions P and Q are linked without one of them being negated.	*and, because*
	Negative	Negated version of either P or Q functions in the relation.	*but, although*
Basic operation	Causal	An implication relation (P --> Q) can be deduced between the two segments.	*because, although*
	Additive	The segments are connected in a conjunction (P & Q).	*and, whereas*
Source of coherence	Objective	Both segments describe situations in the real world; the speaker is not actively involved.	none in English
	Subjective	The speaker is actively engaged in the construction of the relation (e.g., through reasoning).	none in English
Order of the segments	Basic	The segment presenting the antecedent (e.g., cause) precedes the segment presenting the consequent.	*as a result, so*
	Non-basic	The segment presenting the antecedent (e.g., cause) follows the segment presenting the consequent.	*because, therefore*
Further distinctions within the class of additive relations:			
Specificity		The specificity of the content of the segments plays a role in the definition of the relation.	*more specifically, in other words*
List		The segments of the relation can be listed.	*First, secondly*
Alternative		The segments are presented as alternatives.	*or, instead*
Further distinctions within the class of causal relations:			
Condition		The cause presented in a causal relation is not yet realized.	*if, otherwise*
Goal-oriented		One of the segments concers an intentional, goal-directed action by an agent.	*in order to, so that*

Figure 16: The unifying dimensions and features, and their values (Sanders et al., 2018)

Appendix D. ISO tagset

	ISO DRel	Symmetry	Relation and Argument-Role Definitions
1.	Cause	Asymmetric	Arg1 provides a reason for Arg2 to come about or occur.
2.	Condition	Asymmetric	Arg1 is an unrealized situation which, when realized, would lead to Arg2.
3.	Negative Condition	Asymmetric	Arg1 is an unrealized situation which, when not realized, would lead to Arg2.
4.	Purpose	Asymmetric	Arg1 enables Arg2.
5.	Manner	Asymmetric	Arg1 is a way in which Arg2 comes about or occurs.
6.	Concession	Asymmetric	An expected causal relation between Arg1 and Arg2, where Arg1 is expected to cause Arg2, is cancelled or denied by Arg2.
7.	Contrast	Symmetric	One or more differences between Arg1 and Arg2 are highlighted with respect to what each predicates as a whole or to some entities they mention.
8.	Exception	Asymmetric	Arg1 evokes a set of circumstances in which the described situation holds, while Arg2 indicates one or more instances where it doesn't.
9.	Similarity	Symmetric	One or more similarities between Arg1 and Arg2 are highlighted with respect to what each predicates as a whole or to some entities they mention.
10.	Substitution	Asymmetric	Arg1 and Arg2 are alternatives, with Arg2 being the favored or chosen alternative.
11.	Conjunction	Symmetric	Arg1 and Arg2 bear the same relation to some other situation evoked in the discourse. Their conjunction indicates that they are doing the same thing with respect to that situation, or are doing it together.
12.	Disjunction	Symmetric	Arg1 and Arg2 are alternatives, with either one or both holding.
13.	Exemplification	Asymmetric	Arg1 describes a set of situations; Arg2 is an element of that set.
14.	Elaboration	Asymmetric	Arg1 and Arg2 are the same situation, but Arg2 contains more detail.
15.	Restatement	Symmetric	Arg1 and Arg2 are the same situation, but described from different perspectives.
16.	Synchrony	Symmetric	Some degree of temporal overlap exists between Arg1 and Arg2. All forms of overlap are included.
17.	Asynchrony	Asymmetric	Arg1 temporally precedes Arg2.
18.	Expansion	Asymmetric	Arg2 provides further description about some entity or entities in Arg1, expanding the narrative forward of which Arg1 is a part, or expanding on the setting relevant for interpreting Arg1. The Arg1 and Arg2 situations are distinct.
19.	Functional dependence	Asymmetric	Arg2 is a dialogue act with a responsive communicative function; Arg1 is the dialogue act(s) that Arg2 responds to.
20.	Feedback dependence	Asymmetric	Arg2 is a feedback act that provides or elicits information about the understanding or evaluation by one of the dialogue participants of Arg1, a communicative event that occurred earlier in the discourse.

Figure 17: Tagset of relation senses in the ISO proposal (Bunt & Prasad, 2016)

Appendix E. Pseudocode of the PDTB-RST mapping algorithm

```
PROCEDURE findRelationMappings(file){
    FOR pdtbRelation IN file DO:
        // Find equivalent RST spans to arg1 and arg2
        arg1Equivalent := findSpanMatch(pdtbRelation.arg1);
        arg2Equivalent := findSpanMatch(pdtbRelation.arg2);
        bestRelationMatch := null;
        // Find the RST relation at the lowest (smallest) sub-tree spanning over the two RST spans
        FOR rstRelation IN file DO:
            IF rstRelaion.covers(arg1Equivalent, arg2Equivalent) AND
                (bestRelationMatch == null OR rstRelation.spanSize < bestRelationMatch.spanSize) THEN:
                    bestRelationMatch := rstRelation;
        // Checks for replacement and/or flagging:
        //#1: Check for same-unit relation
        same-unitCheck(arg1Equivalent, arg2Equivalent, bestRelationMatch);
        //#2: Check for intervening attribution relations
        attributionCheck(arg1Equivalent, arg2Equivalent, bestRelationMatch);
        //#3: Check for the Strong Nuclearity principle
        nuclearityCheck(arg1Equivalent, arg2Equivalent, bestRelationMatch);
        // After all checks are done, the mapping will be added to the pool
        allMappings.add(pdtbRelation, bestRelationMatch)
    Return allMappings;
}

PROCEDURE findSpanMatch(pdtbArg, file){ // A Span in RST can be a single EDU or a multi-EDU sub-tree
    argEquivalent := null;
    maxOverlap := 0;
    minMargin := 100000;
    FOR rstSpan IN file DO:
        IF (rstSpan.overlap(arg1) > maxOverlap) OR
            (rstSpan.overlap(arg1) == maxOverlap AND rstSpan.margin(arg1) < minMargin)  THEN:
            argEquivalent := rstSpan;
}
```

Figure 18: PDTB-RST mapping algorithm pseudocode

Appendix F. Theoretical mapping between RST-DT and PDTB 2.0 labels according to ISO, OLiA, and Unifying Dimensions

PDTB →	Temporal			Contingency						Comparison						
		Asynch.		Cause		Pragm. cause		Prag. cond.		Contrast				Concess.		
RST-DT ↓	Synchronous	Precedence	Succession	Reason	Result	Justification	Condition	Relevance	Impl. assert.	Contrast	Juxtapos.	Opposition	Pragm. contr.	Expectation	Contra-exp.	Pragm. conc.
Background		u	u													
Circumstance	u, i	u	u													
Cause				o, u, i	o, u, i	i										
Cause-result				o, u, i	o, u, i	i										
Result				o, u, i	o, u, i	i										
Consequence				o, u, i	o, u, i	i										
Comparison										i	i	i				
Preference										u	u	u	u	i	i	
Analogy																
Proportion							u	u	u							
Condition							o, u, i	o, u	o, u							
Hypothetical							o, u, i	u	u							
Contingency							o, u, i									
Otherwise							u, i									
Contrast										o, u, i	o, u	o, u	o, u			
Concession														o, u, i	o, u, i	o, u
Antithesis										o, u	o, u	o, u	o, u	u, i	u, i	u
El-additional																
El-gen.-spec.																
El-part-whole																
El-proc.-step																
El-object-attr.																
El-set-mem.																
Example																
Definition																
Purpose				o	o, u, i											
Enablement					u, i			o	o							
Evaluation				u		u										
Conclusion				o, u	o	o, u	o	o	o							
Comment																
Evidence				u, i	u, i	o, u, i										
Expl.-argum.				o, u, i	o, u, i	i										
Reason				o, u, i	o, u, i	i										
List																
Disjunction																
Summary																
Restatement																
Temp.-before		o, u, i	i													
Temp.-after		i	o, u, i													
T.-same-time	o, u, i															
Sequence		o, u, i	u, i													
Inverted-seq.		u, i	o, u, i													
Means				u	u			o	o							
Problem-sol.				u	u			o	o							
Not mapped								i	i				i			i

Table 5: Proposed mappings between RST-DT and PDTB 2.0 labels (EXPANSION labels on next page). 'i' indicates proposed correspondence according to the ISO proposal, 'o' corresponds to OLiA, and 'u' corresponds to the Unifying dimensions proposal.

PDTB →	Expansion												
				Restatement			Alternative						
RST-DT ↓	Expansion	Conjunction	Instantiation	Specification	Equivalence	Generaliz.	Conjunctive	Disjunctive	Chosen alt.	Exception	List	EntRel	Not mapped
Background		u										i	o
Circumstance		u											o
Cause													
Cause-result													
Result													
Consequence													
Comparison		u											o
Preference									u	u			o
Analogy		u, i											o
Proportion		u, i											o
Condition													
Hypothetical													
Contingency													
Otherwise							o	o	o				
Contrast										u			
Concession													
Antithesis									i	u			
El-additional	o	u, i										i	
El-gen.-spec.				o, u, i		o, u, i							
El-part-whole	o			u, i		u, i							
El-proc.-step	o			u, i		u, i							
El-object-attr.	o			u		u						i	
El-set-mem.	o		u, i										
Example			o, u, i										
Definition	i			u		u							o, i
Purpose													
Enablement													
Evaluation	i			u		u							o
Conclusion				i		i							
Comment	i	u		u		u							o
Evidence													
Expl.-argum.													
Reason													
List											o, u, i		
Disjunction							o, u, i	o, u, i	o, u				
Summary				o, u, i	o	o, u, i							
Restatement					o, u, i								
Temp.-before													
Temp.-after													
T.-same-time													
Sequence													
Inverted-seq.													
Means													i
Problem-sol.	i												
Not mapped										o, i		u	

Table 6: Proposed mappings between RST-DT and PDTB 2.0 EXPANSION labels, including a general EXPANSION category. 'i' indicates proposed correspondence according to the ISO proposal, 'o' corresponds to OLiA, and 'u' corresponds to the Unifying dimensions proposal.

Note: Labels that were excluded by at least two frameworks were not included in this table. These labels are RST-DT's QUESTION-ANSWER, STATEMENT-RESPONSE, TOPIC-COMMENT, COMMENT-TOPIC, RHETORICAL-QUESTION, TOPIC-SHIFT, TOPIC-DRIFT, ATTRIBUTION. General labels for the types CONDITION, CONTRAST, and EXPANSION were included as subtypes because the three proposals mapped to these more general labels.