Mapping PDTB-style connective annotation to RST-style discourse annotation

Tatjana Scheffler **Manfred Stede**
UFS Cognitive Sciences
University of Potsdam, Germany
{tscheff|stede}@uni-potsdam.de

Abstract

Penn Discourse Treebank and Rhetorical Structure Theory annotation account for different aspects of discourse structure, but to some extent, their analyses also correspond to each other. For a corpus annotated with both types of information, we describe a procedure for mapping systematically from the first layer to the second. In this way, we can observe commonalities and differences in the annotations of discourse structure between the two approaches. The method also allows for a data-driven mapping of coherence relations from one taxonomy to another with a suitable independently-annotated corpus.

1 Introduction

Among the various approaches to discourse structure, Rhetorical Structure Theory (RST, (Mann and Thompson, 1988)) and the Penn Discourse Treebank (PDTB, (Prasad et al., 2008)) have inspired a range of annotation projects, so that a number of corpora are available for both, and can be compared to each other. For English, there is some overlap between the RST-DT (Carlson et al., 2003) and the PDTB texts, but to our knowledge the correspondences between the two layers have not been explored yet. In this paper, we describe our implementation of the mapping in the German *Potsdam Commentary Corpus* (Stede and Neumann, 2014), for which RST and PDTB-style connectives have previously been annotated independently.

Both RST and PDTB attempt to model discourse structure, particularly the coherence relations between abstract entities (propositions, etc.) in the text. However, there are well-known differences between the approaches, such as a global (RST) vs. local (PDTB) view on discourse structure, the grounding of coherence relations in the cognitive effect on the reader (RST) vs. the semantic relation between the relation's arguments (PDTB), etc. Still, there is considerable overlap in the inventory of coherence relations between the formalisms, and insights from one type of annotation can confirm or extend insights from the other. For this purpose, we have developed a procedure that maps corresponding parts of PDTB-style and RST annotations to each other. Besides the practical benefit of checking annotation consistency and quality, we see the mapping as potentially fruitful for further theory development:

- *Structural* decisions may differ: Annotators looking for individual relation–argument configurations in PDTB-style analysis, disregarding any notion of overall text structure, may assign different text spans to a relation than RST annotators do when they are forced to produce a well-formed overall tree. Are such disagreements merely accidental, or do they point to interesting cases of ambiguity? Do they yield evidence that the tree constraint of RST may be too strong?

- *Relation types* or *connective senses* in the two approaches overlap but are not identical. When relations are mapped, they can provide information on the granularity, ambiguity, or vagueness in the inventories of categories and their usage.

In the present paper, we do not address the relation types but focus on describing a procedure for the mapping of structures only.

In the following, Section 2 gives a brief description of the two approaches, and then Section 3 states the assumptions we make for the mapping procedure, which is outlined in Section 4. Then, Section 5 discusses our findings on the relationship between the two accounts of discourse structure in the corpus. Finally, Section 6 addresses related work, and Section 7 gives a summary.

2 Discourse annotation

2.1 Connectives: Penn Discourse TreeBank

In PDTB-style annotation, the primary goal is to identify connectives and to link them to their two arguments: 'Arg2' is the one that is syntactically integrated with the connective, and 'Arg1' is the "external" one. Usually, Arg1 and Arg2 are adjacent (or embedded), but occasionally, Arg1 can be non-adjacent. In addition to proper connectives, annotators are encouraged to also find "alternative lexicalizations" (such as productive phrasal expressions) that serve a connecting function. Furthermore, the PDTB also links adjacent sentences into a relation even when no connective lexicalization is present; these cases are called "implicit connectives". Any instance of a relation (signalled or not) receives a sense label, which is taken from a hierarchy of 43 senses.

A key point is that annotation decisions are made for each relation individually. Connective/argument triples are not being related to one another, so no global text structure is built. This is a deliberate decision of PDTB, which aims at taking "one step beyond sentence syntax" but not the leap toward a discourse representation whose construction would be more difficult to annotate and involve more subjective interpretation.

2.2 Rhetorical trees: RST

In RST, coherence relations are being assigned to adjacent "minimal discourse segments", and recursively to larger spans. The original proposal of (Mann and Thompson, 1988) suggested some 25 relations, but different inventories have been used (notably the one for the aforementioned RST-DT, comprising 78 relations). Connectives can make this decision easier, but they are not the subject of annotation. For most relations, one segment is marked as central for the author's purposes ('nucleus') and the other as merely supportive ('satellite'). A few relations are multinuclear: these may contain two or more nuclei. Importantly, the relation assignment is recursively applied to larger spans as well, so that a tree structure results eventually, which spans the complete text and thereby serves as a model of its coherence. Crossing edges are not allowed according to Mann and Thompson, nor can there be any "gaps" in the analysis: The text is a contiguous sequence of minimal units.

Since many relation definitions involve speaker intentions, an RST analysis amounts to reconstructing the author's "plan", and for non-trivial texts this requires quite a bit of subjective interpretation.

In sum, the PDTB and RST analyses start out from quite different, and to a good extent complementary, goals. At the same time, they obviously have some overlap: Often a connective and its arguments will directly correspond to an RST relation and its segments. In research on RST, the role of signalling devices such as connectives has been discussed prominently (Taboada and Das, 2013). As stated earlier, one goal of our work is to be able to systematically study and quantify this overlap.

3 Constraints on the mapping

In our multi-layer annotation scenario, for the connective-argument layer we use a variant of the PDTB approach. We restrict our discussion in this paper to only explicit connectives (in the sense of (Fraser, 1999) or (Pasch et al., 2003)), excluding free phrasal expressions and non-signalled relations (although the method could be extended to these cases in future work). A connective can consist of multiple tokens, which can be continuous (e.g., *in particular*) or discontinuous (e.g., *either. . . or*), in which case there are exactly two parts. Our annotation does not currently include sense relations on this layer.

The RST layer follows the structural constraints defined by Mann and Thompson, and uses a relation set that is a slight adaptation of the original set. In contrast to the RST-DT, relations with centrally embedded segments are not annotated in our corpus.

Both layers (henceforth: *co* and *rr*) have been manually annotated with dedicated tools that support this process; details are given in the corpus description (Stede and Neumann, 2014). The annotators proceeded independently without consulting the other annotation layer.

In this setting, we make the following assumptions for the mapping from *co* to *rr*:

- In principle: If there is a connective, it corresponds to a relation. I.e., the mapping from *co* to *rr* should be total. The exception results from the non-annotated embedded relations; in a case like "The building, even though it is small, is quite comfortable" the *co*-annotated *even though* could not be mapped to a relation in *rr*.

- A *co* cannot signal more than one *rr*, i.e., the mapping is a function.
- Not every *rr* is signalled by an explicit connective, i.e., the mapping is not surjective.
- It is possible (if rare) that two different *co*'s (not a single, discontinuous *co*!) signal the same relation. I.e., the function is not injective.

4 The mapping algorithm

For matching the overt connectives to a corresponding discourse relation, we first converted the data to a common representation: a list of token offsets. Both a *co* and a *rr* annotation consist of two (or more, for multinuclear *rr*) segments/arguments that can be represented as token offset boundaries.

Our mapping algorithm proceeds heuristically on these token offset lists and identifies different structural categories of *co-rr* correspondences.

central We identify centrally embedded *co*'s, for which Arg2 is located within the boundaries of Arg1. As stated above, these cannot be accounted for by our RST trees.

internal Those *co*'s whose two arguments are both part of the same smallest possible *rr* segment are called *internal*. They cannot be matched to an *rr*. For example: *"It cannot be the case that expensive model projects are funded, but basic needs not met."*

exact Next we test for the existence of an *rr* whose two segments exactly match the two *co* arguments. Here we map the *co* to the corresponding *rr*.

boundary If no exact match is found, we differentiate between *local co*'s (the two arguments are adjacent to each other) and *long-distance co*'s (arguments are nonadjacent, with some intervening material). In the local case, we identify the inner segment boundary between the two connective arguments. There should be exactly one *rr* that also shares this segment boundary between its nucleus and satellite. We match the connective to this RST relation.

no match For local *co*'s, if there is no *rr* which shares the *co*'s segment boundary, we conclude a no match. This indicates a segmentation difference.

relaxed In the long-distance case, we try to find a corresponding *rr* for a *co* by matching only the left segment boundary of the (linearly) second segment. Long-distance relations are typical for backward-referring adverbials (e.g. *instead* or *therefore*), which will be captured with this heuristic. In this relaxed setting, we also allow for a one-token difference between the segment boundary of RST and the connectives, to account for possible idiosyncrasies in the connective annotation, where the *co* itself may be included or excluded from Arg2.

non-adjacent Finally, if no match is found for long-distance *co*'s, these are marked as *non-adjacent*.

5 Experiments

5.1 Data: Potsdam Commentary Corpus (PCC)

We have applied our mapping algorithm to the PCC, which consists of 175 documents taken from the editorials page of a local newspaper. The typical text length is 8 to 10 sentences, with 15.8 words on average and 1.8 verbs per sentence; the total number of tokens is roughly 32,000. This collection contains 1104 annotated connectives and 2536 RST relations.

5.2 Results

The results of the mapping algorithm, sorted by category, are shown in Table 1. Altogether, 84.4%

452	**exact match**
431	**boundary match**
49	**relaxed match**
54	central
89	internal
18	no match
11	non-adjacent
1104	connectives

Table 1: Results of the mapping process

of *co*'s could be matched to a corresponding RST relation (the bold rows in the table). This includes 48 times that two *co*'s were matched to the same *rr*. Usually these are combinations of a conjunction and an adverbial (*aber dann* 'but then') etc. The remaining 16.6% could not be matched, mostly due to design differences between the two kinds of annotations: As noted earlier, centrally embedded segments (4.9%) are not annotated in the RST

trees. In addition, 89 (8%) *co*'s were included in the PDTB-style annotation that were not accounted for in RST (the "internal" case). This group consists in large part of coordinating conjunctions that relate phrases smaller than full finite clauses (e.g., VPs or infinitives). It also includes examples where one connective argument is elliptical and very short, such as *Furchtbar, wenn* ('[It's] Terrible, when').

The few remaining failures to match (no match or non-adjacent, 2.6% in total) point to difficult cases such as two-part adverbial connectives (*zwar...aber* 'admittedly... but') or true long-distance relations. The latter relate to the distinction made by (Webber, 2006), who points out that RST analysis corresponds to a constituency structure in syntax, and PDTB analysis also accounts for dependency structure (with a corresponding distinction between 'structural' and 'anaphoric' connectives). All cases in these groups bear future study.

The majority of *co*'s matches exactly one *rr*. The 12 most common *co*'s and the *rr*'s they frequently map to are shown in Table 2.

The results in the main confirm our basic assumptions (as presented in Section 3). The vast majority of connectives match exactly one RST relation. Mismatches are due to the different segment definition in the two annotation layers and to differences in the treatment of long-distance relations between the two approaches to discourse structure (local/lexicalized vs. global). On the other hand, of the 2536 RST relations, only 932 were marked by an explicit connective, showing that the majority of rhetorical relations in our corpus is unsignalled (63%), at least by connectives in the traditional sense. This number corresponds closely to previously reported signalling ratios (Stede, 2011, p. 110). Double marking of the same *rr* was rare (48 instances, less than 2%).

Connective	RST Relation
aber (74) 'but'	concession : 21 antithesis : 18 background : 6 list : 6 joint : 5 interpretation : 4
auch (29) 'also'	list : 13 background : 3 elaboration : 3 joint : 3
dann (35) 'then'	condition : 5 result : 5 sequence : 4 reason : 3
denn (50) 'since'	reason : 32 evidence : 8 cause : 4 interpretation : 4
deshalb (22) 'therefore'	reason : 13 cause : 2 interpretation : 2
doch (83) 'however'	concession : 30 antithesis : 20 contrast : 8 interpretation : 5 reason : 5
oder (25) 'or'	list : 6 disjunction : 6
so (25) 'thus'	reason : 7 condition : 4 evidence : 4
sondern (22) 'but instead'	antithesis : 16 conjunction : 2
und (243) 'and'	conjunction : 94 list : 22 joint : 15 cause : 9 elaboration : 8 ... and 13 further relations
weil (22) 'because'	cause : 16 reason : 3
wenn (75) 'if, when'	condition : 38 circumstance : 13 interpretation : 5

Table 2: Connectives and their signalled relations

6 Related Work

The general literature on coherence relations and their signals is vast but not the main issue of this paper. We mention here a recent study that realizes an annotation project somewhat similar to ours: (Taboada and Das, 2013) add a layer of signalling information to the existing RST annotations in the RST-DT. The authors emphasize that a very wide range of signals (syntactic constructions, layout, genre, etc.) "beyond connectives" is instrumental for coherence. Again, while the work shares our spirit of multi-layer annotation, the range of signals is a separate issue; we believe that connectives—used roughly in the sense as in PDTB—are the clearest class of signals and can be annotated with high reliability; and since both PDTB-style and RST-style annotation is used widely, we regard the task of mapping between the two as of general interest. Finally, we see it as important to correlate two annotations that arose independently; the goal of Taboada and Das is different in that they first inspect the RST relation and then, "assuming the relation annotation is correct" (p. 259) actively search for the signals of that particular relation.

Very recently, attention has centered on mapping different sense hierarchies characterizing discourse

relations to each other. In this line of research, (Rehbein et al., 2016) annotated the explicit and implicit connectives in a corpus of spoken dialogues with semantic relations from the PDTB schema and according to the Cognitive approach to Coherence Relations (CCR, (Sanders et al., 1992)). The authors then map the two sense hierarchies onto one another. Relatedly, (Lapshinova-Koltunski et al., 2015) take a multilingual view in annotating discourse relations and coherence devices across languages and genres, employing different annotation schemas. However, since both approaches use the same basic items as the carriers of discourse relations/structure for each of their annotations (namely, explicit or implicit connectives), structural differences cannot be identified using these methods, which operate on the level of semantic hierarchies.

7 Summary and Conclusion

We provided a procedure for mapping connective annotation (in PDTB style) to RST annotation on the same corpus. The underlying theories play somewhat different roles for discourse analysis, yet one would expect them to be in general compatible; therefore, a systematic comparison of annotated data can reveal points of ambiguity, lack of clarity, or simplification in one of the two conceptions. Here, we gave initial results on the connective-relation mapping in the Potsdam Commentary Corpus. Of particular interest for our future work are the cases where an argument of a connective is not present in the RST analysis, and where this is not due to a straightforward difference in grain size. We will inspect these cases in order to find out whether they are due to ambiguity (a relation can be read as involving a longer or a shorter argument span; both analyses are plausible) or to simplification on the part of RST (the relation perceived in the connective annotation is simply absent in the RST tree, because multiple relations are not allowed by the theory).

In addition to testing the theories, we pointed out that the technique can be useful for a mutual validation of the annotations – the mapping can be used to identify certain annotation errors or guideline inconsistencies.

The PCC data with the two annotation layers is available via our website[1].

[1] http://angcl.ling.uni-potsdam.de/resources/pcc.html

Acknowledgments

Part of the work reported in this paper was funded by Deutsche Forschungsgemeinschaft via SFB 632 *Information Structure*. We thank the anonymous reviewers for their constructive suggestions for improving the paper.

References

L. Carlson, D. Marcu and M.E. Okurowski. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In: J. van Kuppevelt and R. Smith, eds. *Current Directions in Discourse and Dialogue*. Kluwer, Dordrecht.

B. Fraser. 1999. What are discourse markers? *Journal of Pragmatics*, 31(7):931–952.

E. Lapshinova-Koltunski, A. Nedoluzhko and K.A. Kunz. 2015. Across Languages and Genres: Creating a Universal Annotation Scheme for Textual Relations. In *Proceedings of The 9th Linguistic Annotation Workshop*. Denver, Colorado, USA.

W. Mann and S. Thompson. 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *TEXT*, 8:243–281.

R. Pasch, U. Braußе, E. Breindl and U.H. Waßner. 2003. *Handbuch der deutschen Konnektoren*. Walter de Gruyter, Berlin/New York.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi and B. Webber. 2008. The Penn Discourse Treebank 2.0. In *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC)*. Marrakech, Morocco.

I. Rehbein, M. Scholman and V. Demberg. 2016. Annotating Discourse Relations in Spoken Language: A Comparison of the PDTB and CCR Frameworks. In *Proc. of the 10th International Conference on Language Resources and Evaluation (LREC)*. Portoroz, Slovenia.

T. Sanders, W. Spooren and L. Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes*, 15:135.

M. Stede. 2011. *Discourse Processing*. Morgan & Claypool.

M. Stede and A. Neumann. 2014. Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*. Reykjavik.

M. Taboada and D. Das. 2013. Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue & Discourse*, 4(2):249–281.

B. Webber. 2006. Accounting for discourse relations: Constituency and dependency. *Intelligent linguistic architectures*, 339–360.