Disagreements in analyses of rhetorical text structure: A new dataset and first analyses

Freya Hewett and Manfred Stede
Applied Computational Linguistics
University of Potsdam
Germany
lastname at uni-potsdam.de

Abstract

Discourse structure annotation is known to involve a high level of subjectivity, which often results in low inter-annotator agreement. In this paper, we focus on 'legitimate disagreements', by which we refer to multiple valid annotations for a text or text segment. We provide a new dataset of English and German texts, where each text comes with two parallel analyses (both done by well-trained annotators) in the framework of Rhetorical Structure Theory. Using the *RST-Tace* tool, we build a list of all conflicting annotation decisions and present some statistics for the corpus. Thereafter, we undertake a qualitative analysis of the disagreements and propose a typology of underlying reasons. From this we derive the need to differentiate two kinds of ambiguities in RST annotation: those that result from inherent linguistic ambiguity, and those that arise from specifications in the theory and/or the annotation schemes.

1 Introduction

Natural language contains many ambiguities with varied possible interpretations, especially in the domains of pragmatics and discourse. The differences and similarities of annotations from individual coders, the inter-annotator agreement (IAA), is often used to demonstrate that annotation guidelines are effective, the annotators have worked in a precise way, and that overall, the annotations are of a high quality. In recent years, however, the instances of disagreement have gained interest as a resource for more informative models of the underlying task, often under the heading of 'perspectivism' (Uma et al., 2021).

In this study, we focus on the annotation of discourse structure using Rhetorical Structure Theory (RST; Mann and Thompson, 1988). RST annotations provide information about how segments in a text are related to each other with semantic or pragmatic relations such as `cause`, `background`, or `contrast`; we give a brief overview in Sct. 2.1.

With its focus on pragmatic aspects of language use, RST annotation is generally considered to be highly subjective, and as discussed by Marchal et al. (2022), disagreement in alternative annotations can reflect either incorrect annotations or – more interestingly – instances of item ambiguity or of inherent task subjectivity. So far, empirical studies on annotator disagreement in RST (and also for similar frameworks) have been scarce, as we show in Sct. 2.2; one reason is probably the fact that comparing entire tree structures as alternative analyses is a relatively complicated undertaking. To make it more effective, in this paper, we utilise the RST-Tace software (Wan et al., 2019) to compute the individual points of disagreement between two annotators, which we then analyse further.

We use a dataset of English and German corpora that have recently been made available and partly were extended by us with a secondary annotation (see Sct. 3), and we add to this the double-annotated part of the English RST Discourse Treebank (Carlson et al., 2003), which to our knowledge has so far not been analysed for the reasons of the disagreements. For these corpora, we manually inspect a motivated subset of the points of disagreement and build a typology of categories for legitimate alternative analyses.

Our results have multiple implications. Firstly, they provide insights into the variability of discourse structure, as it is comprehended by different annotators. Secondly, our results can lead to improvements on the RST annotation process, with guidelines being made more precise and annotators being made aware of areas of particular difficulty. Thirdly, our disagreement data and typology can be used to improve evaluation methods of discourse parsers and provide inspiration for evaluation of other similarly subjective tasks.

In Sct. 2 we give a brief overview of RST and outline previous work that has looked at annotation disagreement, and in Sct. 3 we introduce the composition of our dataset. Sct. 4 explains RST-Tace (henceforth: Tace), which provides us with the starting point for our analyses that we present in Sct. 5. In Sct. 6 we discuss these results, before Sct. 7 concludes and outlines possible avenues for future work.

2 Background and Related Work

2.1 A brief overview of RST

Idea. According to Mann and Thompson (1988), an analysis in Rhetorical Structure Theory is conducted by first breaking the text into its Elementary Discourse Units (either simple sentences, or certain types of clauses), which we henceforth call 'EDUs', and then recursively combine adjacent EDUs to form larger units (henceforth: 'spans'). We will use the term 'unit' to refer to a portion of text that is either an EDU or a span. Each combination of adjacent units is labelled with a coherence relation; Mann and Thompson proposed a set of ca. 25 relations. Most of them join one unit that is "more important for the author's purposes" – the 'nucleus' – with a unit that is less important – the 'satellite'. The result is a projective tree where units are marked for their nuclearity status. An example in the original notation proposed by Mann and Thompson (but with actual text removed for brevity) can be seen in Figure 1. Nucleus units have an incoming arrow and a vertical line connecting it to the next upper level.

Corpora. For English, the RST Discourse Treebank (RST-DT; Carlson et al., 2003) was introduced in 2003; it is based on annotation guidelines by Carlson and Marcu (2001), where the size of the relation set has been increased to 78. A part of the corpus comes with two annotations and will be part of our dataset (see Sct. 3). A second important English corpus is GUM (Zeldes, 2017), which is being continuously extended with new data and also with new annotation layers. The annotation guidelines of RST-DT and GUM differ in terms of EDU characterisation and relation set, so that the corpora are not immediately comparable. A smaller English corpus that was recently released contains speeches from the UN Security Council (Zaczynska and Stede, 2024). A part of that has two distinct RST analyses, and these will also be used in our study.

For German, a collection of RST data was recently made available by Shahmohammadi and Stede (2024). A part of that material is double-annotated and will be used in our analyses. This data, as well as the UNSC data, were annotated according to the guidelines by Stede et al. (2017).

2.2 Earlier research: disagreement in discourse structure

Annotation projects in all areas of NLP feature some level of disagreement, with possible sources of disagreement at the level of the annotator, the data, or the context (Basile et al., 2021). In the case of RST, disagreements can arise at the annotator level due to ambiguous EDUs being interpreted differently or genuine errors being made (Mann and Thompson, 1988). At the context level, the same annotator can acknowledge that multiple annotations are reasonable – but in traditional annotation practice has to select one of them. At the data level, text spans (whether they are ambiguous or not) can belong to multiple categories simultaneously.[1]

This final aspect of multiple concurrent relations is included in the proposal by Zeldes et al. (2024) for eRST, which aims to provide solutions for some of the limitations of RST. It allows for so-called 'secondary relations' to be annotated on a unit, which breaks the tree property of the overall structure. Zeldes et al. (2024) mention that allowing for multiple relations could also help in providing more information on RST parser 'errors', which in fact constitute legitimate predictions. Liu et al. (2023) explore the types of errors that RST parsers make, finding that implicit discourse relations and long-distance relations are difficult to identify. They use the double annotated English-language RST-DT corpus subset and find that some of the 'errors' found when comparing a parsers' output to a gold annotation, do actually correspond to plausible relations in alternative trees produced by other annotators.

In a recent study, Zikánová (2024), using the Prague Dependency Treebank in addition to a small set of five Czech texts with RST annotations, outlines seven factors which lead to different interpretations of coherence. These include the interpretation of relations due to polysemous or under-

[1] A discussion on the systematicity of many such ambiguities, due to RST's supplying both 'intentional' and 'informational' relations, originated shortly after RST was originally published; see, e.g., (Moore and Pollack, 1992). Correspondingly, ambiguities arising from the multi-faceted notion of nuclearity were dissected by (Stede, 2008).

specified nature of discourse connectives, or the interpretation of scope due to abstract coreferential expressions.

In the context of discourse parsing, Huber et al. (2021) propose using nuclearity distributions rather than a binary nucleus-satellite distinction, for the benefit of nuclearity-sensitive downstream applications. They create 'silver-standard' trees using summarisation and sentiment analysis data, which feature nuclearity distributions and compare these to the doubly annotated section of the RST-DT. They find that these distributions capture disagreement more than the binary assignment.

3 The corpus

Overall, the corpus used in this study consists of 156 texts in English and German, coming from four sources. All texts have two annotations that were produced by well-trained annotators, and the pair always features identical EDU segmentation. This makes a systematic disagreement analysis much easier, and it reflects an annotation procedure convention to separate the segmentation process from the tree building step. (But see our remark in the Limitations section at the end.)

The English texts are from the RST-DT (Carlson et al., 2003) and the UNSC-RST corpus (Zaczynska and Stede, 2024). The texts in the RST-DT are articles from the Wall Street Journal from the late 1980s. We use a subset of the corpus which consists of texts having two annotations that are based on identical segmentation. The UNSC-RST corpus contains transcripts of speeches from the UN Security Council in the years 2014/15, and we work with its doubly-annotated subset.

The German-language data consist of the doubly-annotated subsets of the APA-RST corpus, which are newspaper articles and their manual simplifications into 'easy language' (Hewett, 2023), and of the Potsdam Commentary Corpus (PCC), which collects commentaries from local newspapers (Shahmohammadi and Stede, 2024).

Five different trained annotators created the analyses of the APA-RST texts, and there was a follow-up step that corrected obvious errors or violations of the schema. The same procedure was applied in UNSC-RST, with a team of four annotators. Two well-trained annotators were involved in building the PCC subset, and also at the time in producing the RST-DT.

Since the two German corpora are based on the same annotation guidelines, we fuse them into a single set that we call APA+PCC. UNSC-RST had the same guidelines but is in English; the RST-DT features a much more fine-grained relation set and hence different guidelines. We thus have three subcorpora for which disagreements can be analysed, but cross-corpus comparisons have to keep in mind the differences. For instance, the PCC/UNSC-RST guidelines were conceived for opinionated text, with the goal of supporting argumentation analysis. Hence they distinguish between the relations `Evidence`, `Reason` and `Cause` with different constellations of objective/subjective material. The RST-DT uses many relations that are absent in the PCC/UNSC-RST, such as six fine-grained versions of `Elaboration`, or the relations `Topic-Shift` and `Example`. (A proposal for mapping between the relations sets was made as part of a shared task on RST parsing (Braud et al., 2023).)

Statistics on our corpus size can be found in Table 1. We make available the parallel APA+PCC and UNSC data as XML files in the customary `rs3` format, and as a `csv` that builds on the output of Tace (see below).[2] The RST-DT data is licensed from the LDC[3]; therefore, only the list of IDs of the texts that we used is part of the repository.

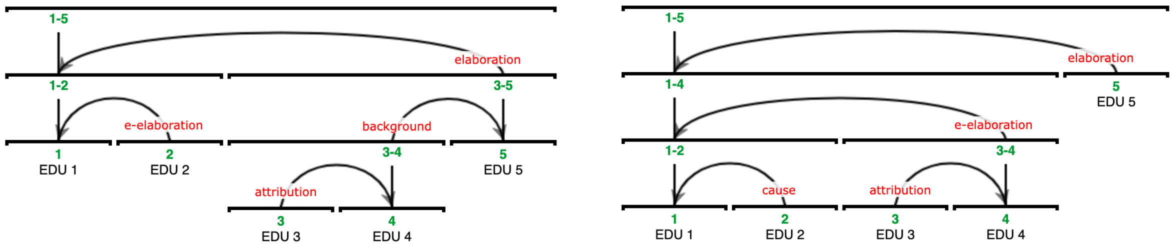
4 Mapping out the disagreements: RST-Tace

We use Tace (Wan et al., 2019) on our corpus to compare the pairs of plausible annotations. Tace takes two RST annotated texts as input, which have identical segmentation, and produces a table comparing the two annotations. Tace calculates IAA using four different aspects: nuclearity (N), relations (R), constituents (C) and attachment points (A), based on a proposal by Iruskieta et al. (2015). A constituent is the satellite span, the attachment point is the span which the constituent is linked to. Pairs of annotated units are matched according to the overlap between central subconstituents (CS); the nuclear units of the satellite of the relation above, or the satellite if the relation is between two EDUs. In Figure 1a, for the `e-elaboration` relation spanning the EDUs 1 and 2, the constituent is 2, the attachment point is 1, and the CS is 2.

Based on the type of mis/match between the two annotators, we create five bins of "annotation deci-

[2] The repository can be found at `https://github.com/discourse-lab/RSTmulti/`.

[3] `https://www.ldc.upenn.edu`



(a) Annotator 1

(b) Annotator 2

Figure 1: Two parallel example annotations.

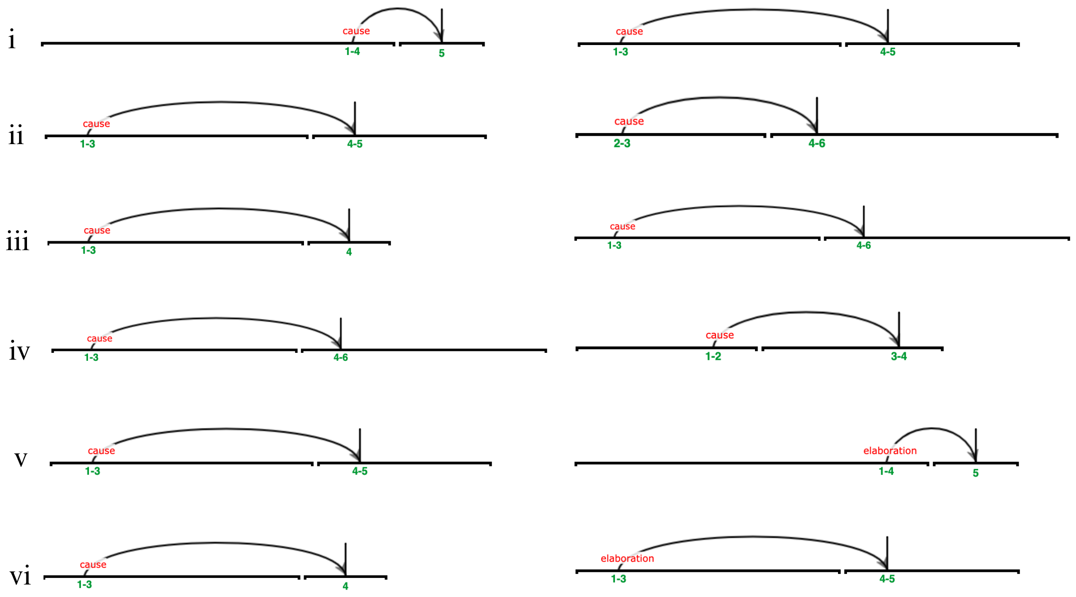


Figure 2: Two parallel extracts from example annotations to illustrate different versions of 'scope mismatch'.

sions" that can be extracted from Tace's output[4], in the form of a spreadsheet where each row contains inter alia the EDU numbers participating in the annotation decision, the actual text spans, and the relations assigned by the annotators. We illustrate the bins with examples from Figures 1a, 1b and 2:

1: Perfect match – Annotators analysed two units in the same way. Example: The `attribution` relation in Fig. 1 constitutes a perfect match.

2: Relation mismatch – Annotators identified the same pair of units but chose a different relation. We can distinguish (i) two mononuclear relations with the same N/S distribution, (ii) one mono- and one multinuclear relation, and (iii) the same units but the N/S distribution is reversed. Example: The different relations between EDUs 1 and 2 (`cause` versus `e-elaboration`) in Fig. 1 belong to category 2(i).

3: Scope mismatch – Annotators disagree on the scope of a relation. This comprises six different constellations: (i) identical overall span; identical relation; different split points; (ii) different overall spans; identical relation; identical split point; different argument spans; (iii) different overall spans; identical relation; one identical argument span; (iv) different overall spans but one common end point; identical relation; different split point, different argument spans; (v) identical overall span, different relations, different split points; (vi) different over-

[4] Details on how we convert the output from Tace to these annotation decisions can be found in Appendix A.2.

all spans, different relations, identical split point, one identical argument span. Example: The `elaboration` relation that encompasses the EDUs 1 to 5 in Fig. 1 belongs to the category 3(i). All cases of scope mismatch can be seen in ascending order from top to bottom in Fig. 2.

4: Left/right priority mismatch – Annotators identified one identical unit, but one attaches it to the left context and one to the right context. Example: The span 3-4 in Fig. 1.

5: No match – Decisions of the first annotator that are not matched at all by the second annotator.

5 Analysis

Table 1 provides some corpus statistics and the distributions of the five bins and the average unit lengths for our three corpora.[5]

In this Section, we cover the three biggest (ignoring "no match") mismatch groups: We will make observations on the perfect matches and then give the results of a qualitative analysis of all relation and scope matches in the corpora APA+PCC and RST-DT. (Analysis of the UNSC corpus and of the remaining bins for the other corpora is left for future work.) For this qualitative analysis, we approach the task from the perspective of a third trained annotator who, however, does not add a third annotation but instead makes a qualitative judgement on the existing two annotations, for each individual mismatch. Section 5.1 discusses the statuses of mismatch and judgement, while in Section 5.2, we present a categorization of underlying *reasons* for the disagreements.

5.1 Status of mismatches

For the *status* of a mismatch, we distinguish four types of judgement that the third annotator can make on a mismatch:

- **Dis**agree: One of the annotations does not seem agreeable, but the other does.[6]

- **Both** are correct and important: A "good" annotation would actually use both relations to do full justice to the text unit (this is the situation that is captured by eRST, as mentioned above).

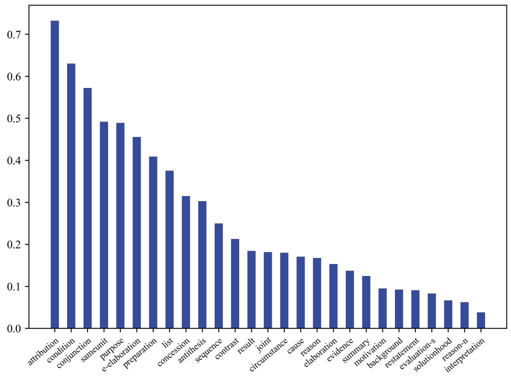


Figure 3: Corpus APA+PCC combined with UNSC-RST: The proportion of relations that occur in a 'perfect match': i.e. the constituent, attachment point, nuclearity and relation are the same.

- **V**ague: One could see things either way, depending on some factors that are to be analysed further (see below).

- **E**ither/**O**r: One can see things either way, but the two ways are actually mutually exclusive.

5.1.1 Perfect match

Fig. 4 shows the confusion matrix for APA+PCC, bins 1 and 2 combined.[7] The diagonal corresponds to perfect matches, which make up between 26% and 49% of all decisions – see Table 1. The avg. number of involved EDUs shows that perfect matches have a clear tendency to occur at the leaf nodes of the trees. Figure 3 shows the relations that occur in a perfect match in the UNSC and the APA+PCC subcorpora combined.[8] `Attribution`, `condition`, and `conjunction` occur frequently in perfect matches, which are relations that often have a clear signal.

For our present purposes, we decided to not analyse the perfect matches; i.e., no status labels were assigned.

5.1.2 Relation mismatch

According to Table 1, this is the largest group of mismatches, and similar to the perfect matches it occurs predominantly at the leaf nodes. When two annotators link the same units but use different

[5] We note that all matches consist of two annotated spans, except for 'no matches', which are counted individually. Therefore the counts for no matches are inflated.

[6] In principle, the situation of disagreeing with both annotations could also arise, but we did not encounter this.

[7] The confusion matrices for the UNSC (Fig. 5) and for the RST-DT (Fig. 6) can be found in Appendix A.1.

[8] We do not include RST-DT in this plot, as it uses a different relation set.

Subcorpus	**APA+PCC**				**UNSC**				**RST-DT**			
Size	46 texts		640 EDUs		84 texts		1346 EDUs		26 texts		768 EDUs	
Agreement	**N**	**R**	**C**	**A**	**N**	**R**	**C**	**A**	**N**	**R**	**C**	**A**
	.50	.33	.46	.42	.60	.38	.55	.51	.56	.37	.53	.49
Tace output bin	n		**Span length**		n		**Span length**		n		**Span length**	
Perfect match	183 (26%)		3.1		410 (29%)		4.4		397 (49%)		5.9	
Relation mismatch	135 (20%)		3.7		288 (20%)		4.2		165 (20%)		4.9	
Scope mismatch	152 (22%)		7.1		301 (21%)		5.9		125 (15%)		11.3	
Left/right mismatch	25 (4%)		3.2		49 (3%)		3.2		8 (1%)		4.4	
No match	197 (28%)		7.4		369 (26%)		7.6		115 (14%)		13.6	

Table 1: Statistics on the corpora and the six bins from Tace output. The average span length is the average number of EDUs contained in the overall relation span. Agreeement values are calculated by Tace and represent F1 values.

relations, this provides the clearest indications for problems with the relation set or with individual definitions provided in the annotation guidelines.

For the 135 instances in APA+PCC, we limit the scope of our analysis to the relation text span that we extracted, i.e., we do not study them in their surrounding context. We find 25 cases of **Dis**, many of which are mismatches between `elaboration` and `entity-elaboration`, where only one appears to actually apply. In 15 cases, no judgement seemed possible because of the missing context; the vast majority are from group 2(ii), involving a mononuclear relation and a `list`, where it is not clear whether other `list` members would warrant the analysis. Of the 28 **Both** cases, many involve a `conjunction` relation, where the other annotator opted for a more informative relation (which points to a guideline problem; see Sct. 6). Roughly half of the **Both** cases do not exhibit a clear linguistic signal and thus would not be annotated in the eRST approach. We find 72 **Vague** cases, and their two biggest subgroups are (i) those where annotators use one of the contrastive relations `contrast`, `antithesis`, `concession`; and (ii) those involving one or two causal relations. When both annotators chose a causal relation, the mismatch is due to different decisions on subjectivity (e.g., `cause` vs. `reason`), while cases with one annotator using a causal relation it is not clear whether a causal connection should be inferred or not (these cases all have no explicit connective).

Within the 165 instances of relation mismatches in the RST-DT, approximately 90 were **Vague**, with a large subset of these (around 50) involving relations that seem to be very similar, such as `analogy` and `comparison`. The second largest subset involved a causal relation in one annotation. Overall, around half of the **Vague** category have some kind of `elaboration` relation in at least one annotation. Around 50 of the relation mismatches represented cases where one annotation does not seem agreeable (**Dis**). The RST-DT has a larger relation set with more fine-grained relations, which has several implications, particularly for this **Dis** category. 12 **Dis** cases involved the same relation, where one relation had the additional suffix '-e' to signify an embedded unit, 19 cases involved a mismatch between `elaboration-object-attribute` and `elaboration-additional`, which mostly differ due to the elaboration being restrictive or non-restrictive. We note that the majority of the **Dis** cases were of this nature and therefore represented negligible 'errors'.

5.1.3 Scope mismatch

In APA+PCC, of the various subcategories listed for (3) at the end of Sct. 4, (i), (ii) and (iv) each occur at most eight times in the data, so that we ignore them here. (iii) has 50 instances and is actually quite close to a 'perfect match', the only difference being that one of the arguments of the relation is of different length in the two annotations. Since this can only be evaluated in context, we studied the 50 instances in their full tree context. In 8 cases (16%), the judgement was **Dis**, as the underlying 'logic' in one of the two analyses seemed implausible. We found a single instance of **EO**, where the different scopes of a `background` relation actually lead to different implications in the surrounding context. The vast majority is **Vague**, usually involving an EDU or very short span being attached to the tree one level lower/higher in the two analyses. One example is a sequence 'If A, then B. Then C.'[9] which can be analysed by first linking B and C into a `list` that forms the satellite of the `condition`, or by stacking two separate `conditions`.

[9] This sounds somewhat uncommon in English, but in German, it is a way of deriving two conclusions from the same antecedent.

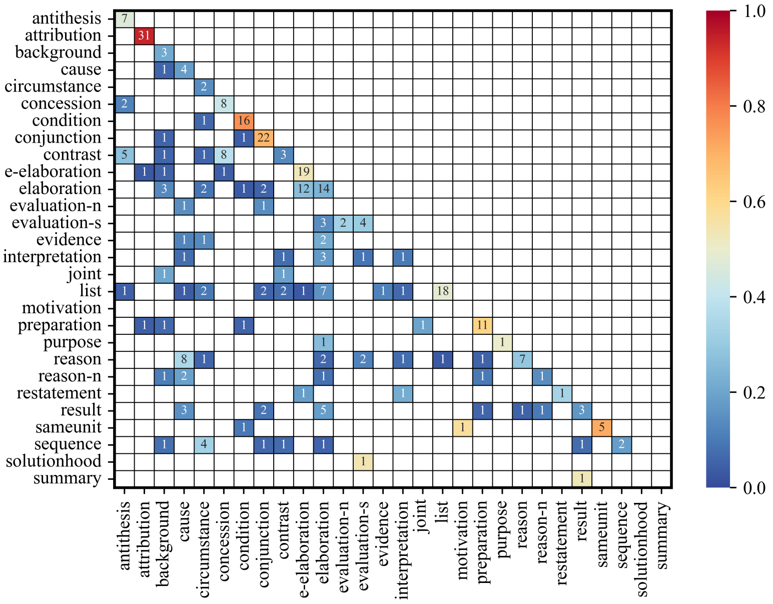


Figure 4: Relations in the categories ‘Perfect match’ or ‘Relation mismatch’ in the double annotated subsets of the German-language subcorpora (APA+PCC).

For longer spans, one recurring pattern stems from annotators applying the “strong nuclearity principle”.[10] In one example, annotator A sees span 8-13 as evaluating the preceding span 1-7; for annotator B, EDU 13 evaluates span 1-12, but therein, span 1-7 is the central nucleus. Both analyses are plausible, the preference depends on the “weight” one gives to the strong nuclearity principle in the decision process.

Another prominent group of disagreements results from ambiguous contrastive/concessive adverbials such as *aber* and *dabei* (which in English are best rendered by the conjunction ‘but’) or *stattdessen* (‘instead’). When they appear sentence-initial, their scope is not restricted by syntax, and their function can be a “strong” contrast between propositions or merely a “weak” signal of topic change, which can lead to different assignments of the boundary of the preceding span (and sometimes of the following span).

Regarding (v) (16 instances) and (vi) (68 instances), they are by their definition rather different, sharing only the overall span (v) or only one argument span (vi). Thus they are the closest constellations to “no match”, and for now we leave their investigation to future work.

The same patterns can be found in our RST-DT subcorpus within the subcategory 3(iii), which consists of 49 cases (of a total of 125 scope mismatches). We note that of these 49, the relation `elaboration-additional` is present in 19 of these cases (almost 40%), compared to its presence in the whole corpus at 17%. The over-proportional presence of this relation makes it clear that it is difficult to pinpoint boundaries between what is being elaborated upon and what constitutes an elaboration, particularly at a higher level in an RST tree. `Attribution` also occurs frequently within 3(iii), and whilst some cases were judged to be **Dis**, i.e. the scope of the attribution did not seem plausible, other cases were ambiguous, with it being difficult to tell how much of the information can be attributed to a source. Examples of this include citing a report or statement without direct quotes. Overall, as the RST-DT has segmentation rules that result in more EDUs per text, and generally more embedded segments, other scope mismatches involved relations such as `sameunit`, and both annotations are equally correct. We also note that the RST-DT texts are mostly longer than those in the German subcorpora and often consist of multiple paragraphs; this formal aspect leads to some annotations which follow these text boundaries, and others which do not, resulting in scope mismatches or left/right mismatches. The RST-DT texts also represent different types of text that can be found in a newspaper; some feature multiple

[10] This principle states that when a relation holds between two spans, it also holds between the central nuclei of the spans (Marcu, 2000).

different topics which each have a lead sentence. An annotator can choose to include the lead sentence directly in the block of text related to the lead, or can separate the lead with a relation such as summary. The nature of this relation, as well as, e.g., comment or circumstance, combined with the mention of specific entities, can make it difficult to pin down exactly what is being commented on or summarised. We also have three cases which we classified as **EO**: These were all due to decisions higher up in the tree, where more specific relations were used, which then limit the scope of elaborations in a specific way. One example of this involved the relation Topic-Drift at the highest level in the tree, which meant that an elaboration was limited to the left-hand side of this relation.

5.2 Reasons for disagreement

Following the categorization of mismatches in the Tace-induced five "formal" bins (step 1) and our judgements on the statuses for a large subset of the mismatches in APA+PCC and RST-DT (step 2) in the previous subsection, we now propose categories of the underlying reasons of the disagreements; they resulted from our observations while conducting the status judgements that we just discussed above.

Formal structural alternatives. When a sequence of EDUs plays the same rhetorical role toward a common nucleus, this can be represented either by stacking the same relation, or by first linking the EDUs into a List, which is then attached to the nucleus. Annotation guidelines should provide guidance for these situations. Likewise, they should specify whether multinuclear relations with more than two nuclei should be binarized or not. (The GUM guidelines[11] do this; others do not.)

Relation definition overlap. As RST definitions operate with different notions, they are by no means mutually exclusive. Elaboration, for example, applies to many EDU pairs where another relation (causal or other) is also appropriate, as our mismatch data shows. Guidelines can suggest to prefer relations that are more informative over very general ones. Another domain where annotators struggle to distinguish similar relations is Antithesis/Concession/Contrast, as our confusion matrices show.

Epistemic status of propositions. Evidence, reason and cause differ in whether the satellite is presented as a factual or as a subjective statement. In many of our corpus instances this is a case of vagueness, where two analyses are equally plausible.

Presupposed knowledge, subjective bias. We found many cases where the decision on non/indentity of referents (e.g., two names of local geolocations) entails topic continuity or switch and hence different coherence relations. Besides such factual knowledge, other mismatches result from subjective interpretation. One example from a corpus text about raising children is the coherence relation depending on whether the expression *all families* includes single parents with their children, or not.

Assignment of 'importance'. When annotators apply the aforementioned strong nuclearity principle, they assign degrees of importance to spans and recursively to EDUs. This can be done by using relations with a 'good' nucleus/satellite assignment (e.g., choosing between Background and Elaboration, or between Cause and Result) or preferring a multinuclear relation like Joint. Perception of relative importance can be highly subjective, however, and the interdependencies between relation/nuclearity decisions on low and high levels of the tree lead to ensuing annotator disagreements.

Text structure. Attachment decisions on higher levels can be influenced by the tension between accounting either for common text structure patterns (in editorials: opening—core—conclusion) or for topic shift, which can run across the borders of the structure blocks. Similarly, in the RST-DT we found examples where the format of the article, esp. paragraph breaks, seems to affect annotation decisions.

Scope of adverbial connectives etc. This is not as much an underlying reason but rather a surface phenomenon that facilitates disagreements. We mentioned examples of ambiguous connectives in Sct. 5; other cases concern demonstratives (*Due to this, ..*) and also ambiguous boundaries of indirect speech: *A said that B. C.* Sometimes it is not clear whether C is in the scope of *said*.

6 Discussion

Our findings on disagreements confirm and extend those of Zikánová (2024), and provide a much larger dataset for further study. We also find that the ambiguity of coreferential expressions or attributive verbs lead to scope mismatches in parallel

[11] https://wiki.gucorpling.org/gum/rst

annotations, while on the annotator level the perception of importance can lead to relation mismatches. These sources of ambiguities are not specific to RST annotation but a fact of language use, and they connect to earlier findings that implicitness – the lack of an overt signal clearly associated with a specific relation – leads to more disagreement (Liu et al., 2023; Pastor and Oostdijk, 2024). This is of particular relevance to automatic discourse parsing and led to the emphasis on signal annotation in eRST (Zeldes et al., 2024).

Ambiguity that is inherent in language, however, needs to be kept distinct from aspects of the theory and the annotation guidelines that create some undesirable choice points for annotators. Our observations on the interaction between perception of importance and nuclearity assignments on all levels of the tree reinforces the concerns stated by Morey et al. (2018), who pointed out that the strong nuclearity principle – and the degree to which annotators rely on it – leads to an inherently unclear notion of the *argument* of a coherence relation in an analysis. 'Perception of importance' is inherently subjective, like the ambiguities discussed above, but it should not propagate to an array of other annotation decisions and cause additional variability in the structures of longer texts. A large number of disagreements that we classified as due to **Vague**ness result from this.

The second important source for them is the routine applicability of multiple relation definitions to a given text span. Our 'status' categories distinguish **Vague** from **Both**, where the former may to some extent be curable by clearer relation definitions, while the latter corresponds to the situation where an annotator should have the option to in the first place assign two relations rather than one. The eRST approach offers this, though only in the presence of overt signals; it can be worthwhile to investigate annotators' behaviour if it would also be allowed in implicit contexts. In addition, other forms of underspecification (of the scopes of certain relations) could be a way of reflecting actual vagueness from the viewpoint of an annotator.

Offering annotators the means to make their uncertainties transparent requires a revised model of discourse structure, and still we will usually work with multiple annotators, so that their potentially-underspecified representations need to be compared in systematic ways to one another. In addition, the consequences for machine learning in discourse parsers and for their evaluation need to be considered – all aspects of perspectivism need to be attended to.

7 Conclusions

This is the first study of RST annotation disagreement that uses a sizeable English/German dataset with two alternative trees, which (except for the RST-DT) we also make publicly available. We have proposed a method for systematically studying the disagreements in three steps of analysis: (i) A formal analysis that extends the output of Tace and builds a list of individual points of disagreement between the annotators. (ii) An evaluation of the status of these disagreements. (iii) A typology of reasons for these disagreements. Using parts of our corpus – 480 instances of disagreements in total – we undertook a first qualitative analysis in this way, and then discussed some implications for potential improvements of annotation guidelines and for incorporating uncertainty into the annotation process.

Limitations

Our study started out with alternative RST analyses that are built on identical EDU segmentations. We believe this is a good decision when first embarking on the empirical analysis of RST structures, but ultimately, segmentation needs to be included into the overall picture.

The judgements made from the perspective of the 'third annotator' in Sct. 5 are the decisions of one of the authors of this paper; from a methodological perspective they can be strengthened by adding a second expert and determining agreement.

Our approach makes inspecting many types of agreement more efficient, but removing the context from the material that is being judged obviously creates some limitations. For scope mismatches, we consulted the full text, but for relation mismatches on identical spans we did not. This might lead to some inaccurate judgements.

Finally, using Tace limits the approach to handling concurrent annotations pairwise; if more than two are available, they cannot be immediately integrated into the present workflow.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 317633480 – SFB 1287.

References

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes, editors. 2023. *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*. The Association for Computational Linguistics, Toronto, Canada.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. Technical report, Univ. of Southern California/ISI. Unpublished manuscript.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Jan van Kuppevelt and Ronnie Smith, editors, *Current Directions in Discourse and Dialogue*. Kluwer, Dordrecht.

Freya Hewett. 2023. APA-RST: A text simplification corpus with RST annotations. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 173–179, Toronto, Canada. Association for Computational Linguistics.

Patrick Huber, Wen Xiao, and Giuseppe Carenini. 2021. W-RST: Towards a weighted RST-style discourse framework. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3908–3918, Online. Association for Computational Linguistics.

Mikel Iruskieta, Iria da Cunha, and Maite Taboada. 2015. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language Resources and Evaluation*, 49(2):263–309.

Yang Janet Liu, Tatsuya Aoyama, and Amir Zeldes. 2023. What's Hard in English RST Parsing? Predictive Models for Error Analysis. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–42, Prague, Czechia. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3659–3668, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Daniel Marcu. 2000. The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. *Computational Linguistics*, 26(3):395–448.

Johanna D. Moore and Martha E. Pollack. 1992. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544.

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2018. A dependency perspective on RST discourse parsing and evaluation. *Computational Linguistics*, 44(2):197–235.

Martial Pastor and Nelleke Oostdijk. 2024. Signals as Features: Predicting Error/Success in Rhetorical Structure Parsing. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 139–148, St. Julians, Malta. Association for Computational Linguistics.

Sara Shahmohammadi and Manfred Stede. 2024. Discourse parsing for German with new RST corpora. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 65–74, Vienna, Austria. Association for Computational Linguistics.

Manfred Stede. 2008. RST revisited: Disentangling nuclearity. In Cathrine Fabricius-Hansen and Wiebke Ramm, editors, *'Subordination' versus 'coordination' in sentence and text*. John Benjamins, Amsterdam.

Manfred Stede, Maite Taboada, and Debopam Das. 2017. Annotation Guidelines for Rhetorical Structure. Unpublished manuscript.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Shujun Wan, Tino Kutschbach, Anke Lüdeling, and Manfred Stede. 2019. RST-Tace A tool for automatic comparison and evaluation of RST trees. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 88–96, Minneapolis, MN. Association for Computational Linguistics.

Karolina Zaczynska and Manfred Stede. 2024. Rhetorical strategies in the UN security council: Rhetorical Structure Theory and conflicts. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 15–28, Kyoto, Japan. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2024. eRST: A Signaled Graph Theory of Discourse Relations and Organization. *Computational Linguistics*, pages 1–47.

Šárka Zikánová. 2024. Text Structure and Its Ambiguities: Corpus Annotation as a Helpful Guide. In *Proceedings of the 24th Conference Information Technologies – Applications and Theory (ITAT 2024)*, pages 2–12, Drienica, Slovakia.

A Appendix

A.1 Confusion matrices

Figures 5 and 6 show the confusion matrices for perfect matches and relation mismatches in the UNSC and the RST-DT, respectively.

A.2 Tace categories

Table 2 shows how we produced our annotation labels using the output from Tace.[12] In a first step, we used all the matches from Tace. Tace distinguishes between three different categories when comparing two RST trees: ‘no matching’, ‘partially identical CS’ and ‘completely identical CS’. For each category, it is further specified which of the four aspects match (nuclearity, relations, constituents, and attachment points). More information on what constitutes a match can be found in Wan et al. (2019). We used the categories outlined in Table 2. We then went through the ‘no matches’ category, according to Tace, and applied simple rules to find further members of our categories. We did this as we are interested in all cases of e.g. relation mismatch, regardless of whether the central subconstituent is the same (which is the method Tace uses to classify matches). We applied the rules in the following order: relation mismatch, relation mismatch with nuclearity switched, left/right mismatch, scope mismatch. An annotated unit can only occur once in our categorisation.

[12] More information can be found in our script: `https://github.com/discourse-lab/RSTmulti/`. Tace is available here: `https://github.com/tkutschbach/RST-Tace`.

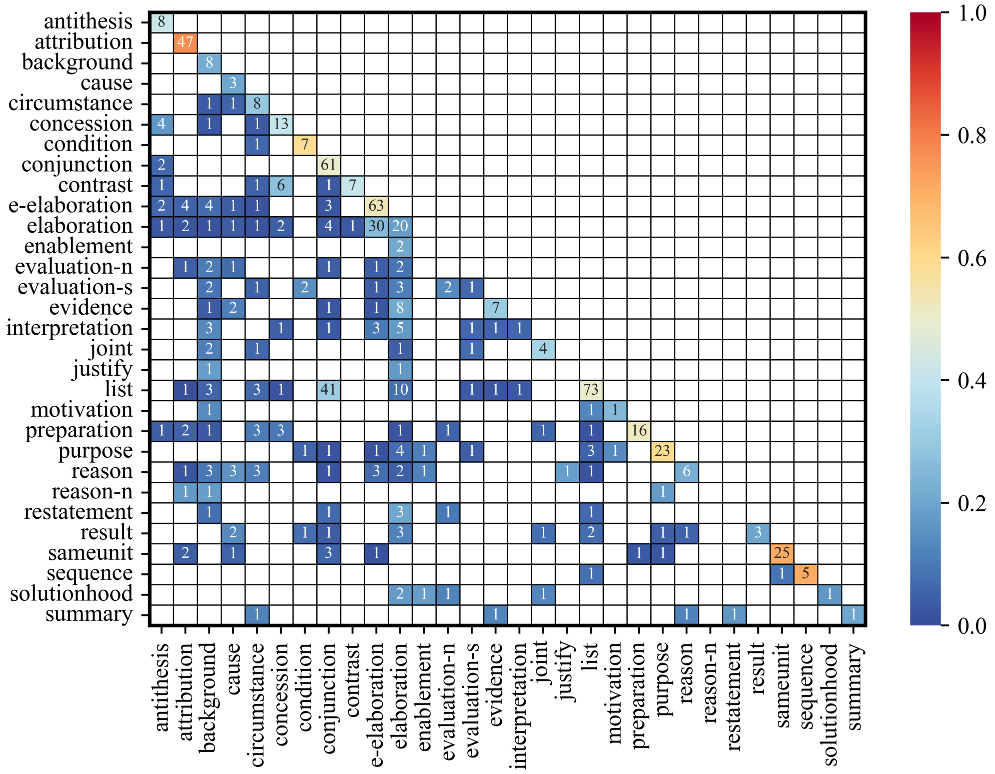


Figure 5: Relations in the categories ‘Perfect match’ or ‘Relation mismatch’ in the double annotated subset of the UNSC (Zaczynska and Stede, 2024).

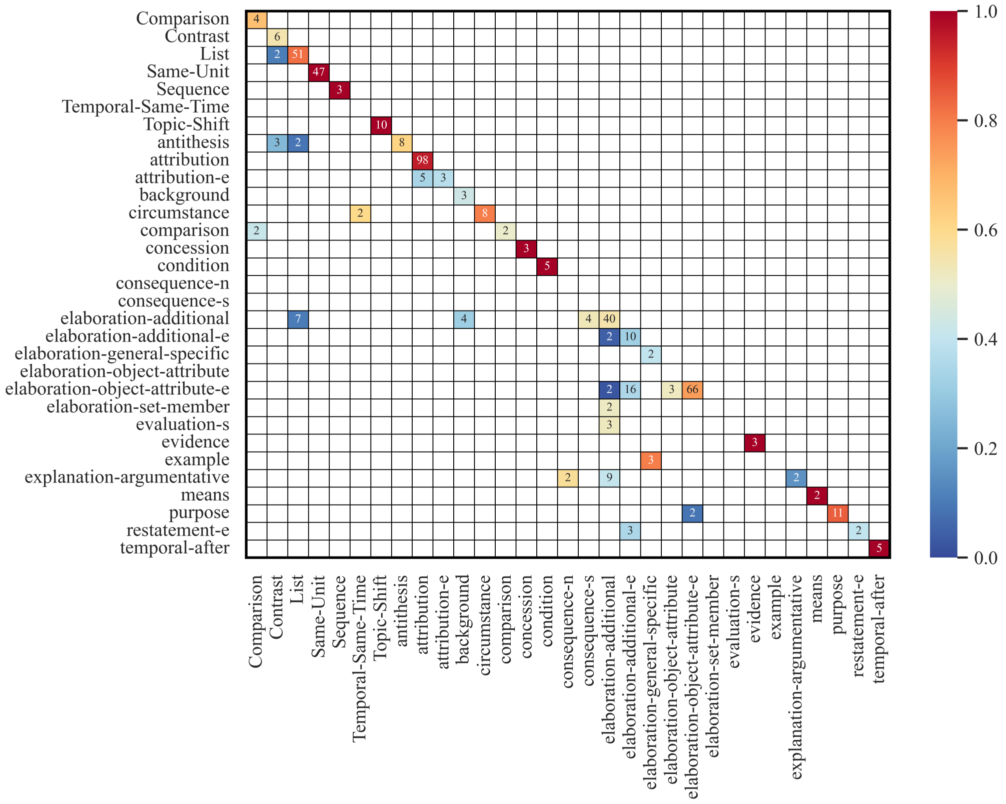


Figure 6: Relations in the categories ‘Perfect match’ or ‘Relation mismatch’ in the double annotated subset of RST-DT. Relation pairs which only occur once are not shown, for readability reasons.

Tace output	**Matching**	**Agreement**	**Disagreement**	**Other conditions**
Perfect match		*NRCA*		
Relation mismatch		*NCA*		
	C1=C2 and A1=A2 or *C1=A2 and A1=C2*		*N/N-N/S*, $\neq$ *R*	
	C1=C2 and A1=A2	*A*	*N/N-N/S*, $\neq$ *R*	
	C1=A2 and A1=C2		*N/S*, $\neq$ *R*	
Left/right mismatch	Completely identical CS	*C*	*N/S*, $\neq$ *R*	
	Partially identical CS		*N/N-N/S*, $\neq$ *R*	One span identical, the non-identical span on left in first annotation and on right in second annotation
Scope mismatch		*NR*		
		NRC		
		NRA		
				Not in any of the above categories, other conditions are outlined in Section 4
No match				Not in any of the above categories

Table 2: Information on how our categories were derived using Tace's (Wan et al., 2019) output.