The Good, the Bad, and the Disagreement: Complex ground truth in rhetorical structure analysis

Debopam Das
Dept. of Linguistics
University of Potsdam
Potsdam, Germany
debdas@uni-potsdam.de

Maite Taboada
Dept. of Linguistics
Simon Fraser University
Burnaby, BC, Canada
mtaboada@sfu.ca

Manfred Stede
Dept. of Linguistics
University of Potsdam
Potsdam, Germany
stede@uni-potsdam.de

Abstract

We present a proposal to analyze disagreement in Rhetorical Structure Theory annotation which takes into account what we consider "legitimate" disagreements. In rhetorical analysis, as in many other pragmatic annotation tasks, a certain amount of disagreement is to be expected, and it is important to distinguish true mistakes from legitimate disagreements due to different possible interpretations of the structure and intention of a text. Using different sets of annotations in German and English, we present an analysis of such possible disagreements, and propose an underspecified representation that captures the disagreements.

1 Introduction

The past ten years have seen continuous interest in RST-oriented discourse parsing, which aims at automatically deriving a complete and well-formed tree representation over coherence relations assigned to adjacent spans of text. For various downstream applications (e.g., summarization, essay scoring), such a complete structure is more useful than the purely localized assignment of individual relations, as it is done in PDTB-style analysis (Prasad et al., 2008).

At the same time, it is well known that RST parsing is difficult, and furthermore, it is more difficult to achieve good human agreement on RST trees, as compared to PDTB annotation. This latter problem has not been in the spotlight of attention, though, while the computational linguistics community developed a series of parsing approaches over the years (Hernault et al., 2010; Ji and Eisenstein, 2013; Feng and Hirst, 2014; Braud et al., 2016). Part of the reason for the focus on data-oriented automatic parsing is the availability of the RST Discourse Treebank (Carlson et al., 2003), a corpus large enough to supply training/test data in supervised machine learning (ML).

The central thesis of our paper is that the fundamental questions of RST annotation and agreement deserve to be re-opened. With powerful ML and parsing technology in place, it is timely to give more attention to the nature of the underlying data, and to its descriptive and theoretical adequacy. Our claim is that the "single ground truth asssumption" is essentially invalid for an annotation task such as rhetorical structure, which inevitably includes a fair amount of subjective decisions on the part of the annotator. As we will emphasize later, we regard this *not* as a fault of Rhetorical Structure Theory (Mann and Thompson, 1988; Taboada and Mann, 2006), but as a reality to accept, shared with labelling of other pragmatic phenomena, such as speech acts or presuppositions.

Specifically, we will argue that a certain amount of ambiguity is to be regarded as part of the "gold standard" or "ground truth". At the same time, it is clear that RST annotation is not a matter of "anything goes". So, the central challenge in our view is to differentiate between good and bad disagreement: Two annotators may legitimately disagree on some part of the analysis, when both alternatives are in line with the annotation guidelines, and they arise from, for instance, different background knowledge. This needs to be kept separate from disagreement with a not-so-well-educated annotator who misread the guidelines and thus sometimes makes analysis decisions that should not be regarded as legitimate.

Our overall project has two parts: Teasing apart the two types of disagreement, and adequately representing the space of legitimate alternative analyses. In this paper we focus on the first task and

To appear in *Proceedings of the 6th Workshop on Rhetorical Structure Theory and Related Formalisms.* Santiago de Compostela, Spain. September 2017. Current version: July 20, 2017.

provide a brief sketch of the second.

In the next section, we discuss relevant related work, and then present two agreement studies we undertook on German and English texts (Section 3). We draw conclusions from both in Section 4 and then sketch our framework for technically representing alternative analyses in Section 5. A brief summary (Section 6) concludes the paper.

2 Related work

In Computational Linguistics, a discussion on ambiguity in RST started shortly after Mann and Thompson (1988) was published, mostly in the Natural Language Generation community. The well-known proposal by Moore and Pollack (1992) argued that certain text passages can systematically have two different analyses, one drawing on the intentional, the other on the subject-matter (informational) subset of coherence relations. In a pair of two sentences, for example, when the first states a subjective claim, the second might be interpreted as EVIDENCE for the first, or as merely providing ELABORATION. Moore and Pollack also gave examples where the alternative analyses coincide with conflicting nuclearity assignments.

These questions were never really resolved; instead, with the availability of the RST Discourse Treebank (RST-DT), attention shifted to automatic parsing with ML techniques, starting with Marcu (2000), who also suggested a way of measuring agreement between competing analyses, splitting the overall task into four subtasks (units, spans, nuclearity, relations); we will also use this approach below in our experiments. As to the results achieved, Carlson et al. (2003) reported these kappa results for an experiment with pre-segmented text (i.e., where there is no point in computing unit agreement): spans .93, nuclearity .88, and relations .79. Note that these results were obtained after annotators had already worked for several months on many texts.

More recently, van der Vliet et al. (2011) annotated a Dutch corpus, and computed agreement following Marcu's method, also using pre-segmented text. They report an average kappa agreement of .88 on spans, .82 on nuclearity, and .57 for relations. These figures should not be directly compared to those of Carlson et al., because there are differences in the relation set, the guidelines, and the amount of annotator training.

The problem of ambiguity was again studied by Schilder (2002), who worked in the framework of Segmented Discourse Representation Theory or SDRT (Asher and Lascarides, 2003) and approached the problem from a semantic viewpoint. He proposed that certain aspects of the analysis could be left unannotated. For instance, nuclearity may be assigned, but the specific relation between nucleus and satellite may be left blank, if a decision cannot be reached.

Around the same time, Reitter and Stede (2003) proposed the Underspecified Rhetorical Markup Language (URML), an XML language for encoding competing analyses in a single representation. We will describe this in more detail in Section 5.

More recently, Iruskieta et al. (2015) proposed a qualitative method for analysis comparison, teasing apart constituency, relation, and attachment. The most important aspect of their comparison method is that nuclearity and relation label are separated, unlike in Marcu's quantitative agreement metric.

3 Empirical studies

Both of our studies are attached to existing RST-annotated corpora, so that our results can be related to the earlier work. Also, we used nearly-identical annotation guidelines, which we describe first, before we turn to the actual experiments.

3.1 Annotation guidelines

In contrast to the RST-DT project of Carlson et al. (2003), our annotation guidelines follow the original RST paper (Mann and Thompson, 1988) relatively closely. This means that our relation set is much smaller than that of the RST-DT (31 relations instead of 78). We do not use the many nucleus-satellite variants, and we deliberately left out suggestions like TOPIC-COMMENT or ATTRIBUTION, which we do not regard as coherence relations in the same way as those of "classic" RST.[1] We group the relations in a slightly different way from Mann & Thompson into subject-matter and presentational ones, and we have an extra category for textual relations (LIST, SUMMARY).

For technical reasons, at the moment we avoid the SAME-UNIT relation of the RST-DT by not

[1] We are of course not claiming that phenomena of Topic/Comment and Attribution do not exist. Instead, notions of information structure in our view belong to a separate level of analysis—not to that of coherence relations.

separating center-embedded segments. This decision may be revised later, and it is not critical for the purposes of this paper.

For the German experiment, we used the annotation guidelines developed for the Potsdam Commentary Corpus (Stede, 2016) and which are publicly available. Then, for annotating the English texts, we produced an English version of those guidelines and made minimal changes to the descriptions of relations (clarifications on how to distinguish between certain contrastive and argumentative relations). Further, we used language-specific segmentation guidelines that we borrowed from the implementation of SLSeg (syntactic and lexically based discourse segmenter) (Tofiloski et al., 2009).[2] In addition to many individual examples for the relations, the guidelines finish with a sample analysis of a complete text with 14 elementary discourse units (EDUs).

The guidelines merely guide the annotators in their task. They could in principle be written in such a way as to "strongly encourage" agreement when cases of ambiguity arise (e.g., by specifying preference hierarchies), but they make only minimal use of that move. The interesting issue from a theoretical viewpoint is that the same general guidelines can give rise to what we consider as legitimate disagreements.

3.2 Study I: German

For the German study, we selected ten texts from the publicly available Potsdam Commentary Corpus[3], which has been annotated at various levels of linguistic description, including RST. They are editorials or "pro and con" commentaries from local newspapers, with a typical length of 8 to 10 sentences (with an average length of 16 words, sentences often consist of more than one EDU). We picked texts of general-interest topics and which do not make too many references to local events or people, which might confuse annotators.

The idea of the annotation experiment was to assess the influence of the amount of training that annotators receive. Thus we worked with four annotators, all with university education. Two of them received fairly extensive training (henceforth: GE1 and GE2 for German Expert 1 and 2): They first read the guidelines and studied the analysis of the sample text, then discussed their questions with us. Thereafter, they were asked to individually annotate three texts (from the same genre, but not used in the experiments), and the results were jointly discussed and adjudicated. The other two annotators (henceforth: GL1 and GL2) were only lightly trained; they read the guidelines, could ask questions, and then worked on one text together, which was subsequently discussed with us. The overall procedure stretched over several days, and each annotator spent between 12 and 15 hours on the experiment. They all received €50 as reimbursement.

One variable that for present purposes we are not interested in is the segmentation of texts into EDUs. We therefore decided to present the pre-segmented text (as found in the corpus) to the annotators. For one thing, this reduces the effort of the annotators, and—more importantly—it makes it easier to focus on the evaluation on the aspects we are targeting: decisions on spans, nuclearity, and relations.

In our evaluation, we first looked at the pairwise agreement of the two annotators within the groups GE1-GE2 and GL1-GL2, respectively. When applying the measures of Marcu (2000), one consequence of our using pre-segmented text needs to be discussed: Since EDUs are a priori identical for all annotators, an artificial agreement arises for the span decisions pertaining to EDUs. We decided to disregard all the spans consisting of just one EDU from the calculation. Had we included them, the overall agreement values would be higher, but the surplus would not reflect decisions made by the annotators themselves.

	Span	Nuclearity	Relation
GE1 - GE2	65.6	43.7	24.0
GL1 - GL2	51.6	25.4	9.7

Table 1: Percent agreement of annotators in the two groups (German study, 10 texts)

In this study, we calculated percent agreement among the annotators. The results for the group-internal agreement are given in Table 1. All figures are substantially better for the expert annotators, with the clearest margins for nuclearity and relations. We have to be careful in drawing conclusions, since each group consisted of just two an-

[2] Annotation guidelines: http://www.sfu.ca/~mtaboada/docs/research/RST_Annotation_Guidelines.pdf

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

notators, but the result indicates that the difference in training time and content—in particular, we surmise, the difference in the number of jointly-discussed sample analyses—leads to a marked difference in annotator agreement.

In order to measure the agreement between expert and non-expert annotators, we computed the precision and recall values for GE1 and GL1, following the method documented in Marcu (2000). GE1 was considered as the "gold" annotation. The precision and recall values, provided in Table 2, show relatively higher agreement for spans and nuclearity, but low agreement for relations. Precision and recall are the same, because there are equal numbers of false positives and false negatives.

	Precision	Recall
Span	0.65	0.65
Nuclearity	0.56	0.56
Relation	0.30	0.30

Table 2: Precision and recall for expert versus student annotation (GE1-GL1)

We also conducted various more detailed analyses, but for reasons of time, only a randomly chosen subset of five texts and their RST trees could be handled in this phase. In Table 3, we report the percent agreement results for all pairs of annotators.

	Span	Nuclearity	Relation
GE1 - GE2	63.6	43.8	27.0
GL1 - GL2	60.6	35.2	15.4
GE1 - GL1	56.6	38.8	13.2
GE1 - GL2	48.8	31.2	19.6
GE2 - GL1	63.4	44.2	23.8
GE2 - GL2	44.2	35.2	15.4

Table 3: Percent agreement of all annotator pairs (German study, 5 texts)

First of all, notice that the results for GL1-GL2 are considerably closer to those of GE1-GE2 than in the comparison of the full 10 texts; this indicates that the texts selected are "easy" ones. But the main insight to be gained from Table 3 is that the poor results of GL1-GL2 are mainly due to the performance of GL2, who consistently reaches low agreement with all three other annotators (the single exception being the Relation agreement with GE1), while GL1 does a fairly good job; in particular s/he agrees with GE2 essentially as much as GE1 does.

One other factor we investigated is the "difficulty" of individual RST relations. On the basis of the five texts, we computed how many pairs of annotators achieve at least one perfect agreement for a particular relation type. The results are given in Table 4. The second column gives the number of pairs of annotators that agree on the relation label (and also on spans and nuclearity) in at least one text.

Relation	Ann.pairs	Percent
Preparation	6	100
Condition	6	100
Evaluation-S	5	83
List	4	66
Circumstance	4	66
Elaboration	3	50
Conjunction	3	50
Background	2	33
E-Elaboration	2	33
Contrast	2	33
Cause	2	33
Reason	1	16
Joint	1	16
Antithesis	1	16
Restatement	1	16
Result	1	16

Table 4: Pairwise annotator agreement (%) on relations (German study, 5 texts)

Again, the figures have to be taken with some caution; while the number of annotator pairs entering the calculation is not so low, we studied only five texts here. The ranking, however, confirms the intuition that those relations that tend to occur low in the tree (relating EDUs), and are often clearly marked by connectives, receive the most agreement in annotation.[4]

3.3 Study II: English

In the interest of comparability with the German study, we selected the text material from an RST-annotated corpus, in this case the RST Discourse

[4] Running this calculation on the different levels of the hierarchy has not been done but is an interesting step for future work.

Treebank (Carlson et al., 2003), but we did not use the associated annotation guidelines, as explained earlier. To match the genre of "commentary", we looked especially for argumentative text (which in general we expect to be more prone to competing analyses, since more interpretation and subjectivity is involved than in plain news text). In total we found 19 such documents in the RST-DT, which are letters to the editor, editorials, op-ed pieces, or reviews. For our present experiment, we selected four of the documents. One document contains multiple letters; we split it up and thus have a set of seven individual texts to work with. With an average length of 205 words per text, they are somewhat shorter than the German texts.

Also in line with the German study, we performed a pre-segmentation (following the rules mentioned in Section 3.2) of all the texts, so that annotators started from a basis that allows for a solid comparison of span, nuclearity and relation decisions. In terms of annotator teams, however, we could not exactly replicate the setting of the previous study. Instead, two authors of this paper (who have many years of experience with various RST annotation projects) served as "expert" annotators (henceforth: EE1 and EE2). On the "non-expert" side, we recruited a student of Linguistics (EL1) who carefully studied the guidelines, practiced, and discussed her questions with us. All annotations were done with RSTTool (O'Donnell, 2000).

Quantitative analysis. To determine the extent to which expertise leads to higher agreement, we again computed the percent agreement on spans, nuclearity and relations between the two experts (EE1 and EE2), and between one expert and the lightly-trained annotator (EE1 and EL1). These figures are given in Table 5. As in the German study (Table 3) we see a difference between E-E and E-L agreement, which is much less pronounced for spans than for nuclearity and relations. The main difference between the two studies, however, is that overall the English results are considerably better than the German ones. To a large extent we can attribute this to the difference in having experienced expert annotators (English) as opposed to well-trained students (German). This does not explain the better results for the EE1-EL1 pair in comparison to all the GE-GL pairs, though. There must be an additional factor, and we suspect it is the fact that the English texts are shorter and thus somewhat easier to annotate in the sense that there is less room for different interpretations.

	Span	Nuclearity	Relation
EE1 - EE2	95.1	67.0	49.8
EE1 - EL1	94.8	57.1	35.2

Table 5: Percent agreement of two annotator pairs (English study, 7 texts)

	Span	Nuclearity	Relation
EE1 - EE2	75.6	42.3	40.3
	90.2	50.5	48.2
EE1 - EL1	74.4	24.1	23.0
	89.7	35.7	33.2

Table 6: Chance-corrected agreement of two annotator pairs (English study, 7 texts); for each group, line 1 provides fixed marginal kappa, line 2 free marginal kappa

In addition, we computed kappa values for the same pairs of annotators in order to see the influence of chance agreement. These results are shown in Table 6. In the calculations, the span agreement includes the (implicit agreement on) non-existing spans (i.e., spans that neither annotator marked), while these were left out for computing the nuclearity and relation agreement. In the related work, this point is usually not mentioned; we believe it is important to make explicit how the "virtual" spans are being handled.

Finally, as in the German study (see Table 2), we determined the agreement in terms of precision and recall between EE1 and EL1. For this purpose, we made use of RSTEval, a tool that provides precision and recall statistics between a "gold" human annotation and a parser-produced annotation.[5] EE1 was considered as the "gold" annotation here and thus we have the same scenario as in evaluations of automatic parsers against human annotations. Table 7 provides these results, showing once again high agreement in spans and nuclearity, but quite low agreement in relations. Precision and recall are the same, because there are equal numbers of false positives and false negatives.

[5] http://www.nilc.icmc.usp.br/rsteval/

	Precision	Recall
Span	0.88	0.88
Nuclearity	0.58	0.58
Relation	0.41	0.41

Table 7: Precision and recall for expert versus student annotation (EE1-EL1)

Qualitative analysis. We are also interested in a qualitative comparison: Which phenomena in the texts triggered discrepancies in the two analyses, and of what kinds are the resulting structural differences? We carried out a study of the disagreements in the English data, and found that disagreements involving spans, nuclearity and relations emerge from a number of sources. This is evident in the pairwise comparison between the expert annotations, and to a larger degree, between the expert and non-expert annotations.

Differences in spans primarily result from differences in the point of attachment of EDUs or larger segments. Figures 1 and 2 below exemplify two structures produced by the expert annotators who attach the spans at either different points or different levels. Both annotations employ CONTRAST and BACKGROUND relations, but the spans constituting these relations are different in length and hierarchy.

The situation is more complicated in cases for nuclearity where there are two main sources of disagreement. In the first case, the annotators assign equal or unequal importance to the respective spans, resulting in the formation of a mononuclear and a multinuclear relation. In the second case, both the annotators choose a mononuclear relation, but each assigns a different nucleus-satellite order (NS vs. SN order) to the respective spans.

More importantly, the differences in nuclearity assignment have a follow-up effect on choosing relevant relation labels. First, assigning a mononuclear vs. multinuclear structure further constrains the choice of relation labels, as the mononuclear and multinuclear relations in an RST taxonomy contain two mutually exclusive sets of relations. For instance, in one of our analyses, assigning equal vs. unequal importance to spans results into a mononuclear ANTITHESIS and a multinuclear CONTRAST relation (Note: both relations are of contrastive type). Second, assigning an opposite nucleus-satellite order also contributes to selecting different relations, most of which are mirror relations (differing primarily according to the nucleus-satellite order), such as CAUSE vs. RESULT.

Finally, the differences in relation are also caused by choosing an altogether different or similar relation label for the otherwise same discourse structure involving the same spans and identical nuclearity assignment. We have one such example in our corpus, with the two labels being SUMMARY and RESTATEMENT.

4 Conclusions from the experiments

The most popular method to measure agreement Marcu (2000) computes precision and recall with four factors: Elementary Discourse Units (EDUs), units linked with relations (Spans), nucleus or satellite status (Nuclearity), and relation label (Relation). One problem with this method is that it measures twice the same type of decision: Whether the units are linked (Span), and the status of each unit as nucleus or satellite. This problem is extensively discussed by Iruskieta et al. (2015).

Another problem with this type of evaluation is that it is just quantitative, that is, it does not distinguish between different types of disagreements and their "quality". We believe that on the one hand there are true mistakes in discourse annotation, maybe due to lack of experience in annotation, carelessness, or any other human factor. We also believe, however, that other differences in annotation may be considered legitimate disagreement, i.e., annotations that are both valid from a theoretical point of view. This is particularly the case in argumentative texts, where the analysis hinges on how the annotator perceives the writer's intentions. Those may not be equally clear to annotators in argumentative texts, as they are more subjective than descriptive text types.

In particular, what we find with inter-annotator agreement studies, is that (i) spans are relatively easy to identify; (ii) nuclearity increases complexity and leads to disagreements; and, most importantly, (iii) relation assignment seems particularly difficult. We propose that some of the more fine-grained distinctions among relations may not be relevant in all cases and all uses of RST trees. Thus, an underspecified representation of spans and nuclearity, plus reliably annotated relations, may be sufficient in many cases. We propose such a representation in the next section.

5 The complex gold: Capturing ambiguity

As we mentioned earlier, the second part of our overall project is to represent the expert annotations in a common data structure; in this paper, we describe the direction we are taking. Below, we briefly describe the framework we are using for this, and illustrate the conversion with an example from the English expert annotations.

5.1 URML

The Underspecified Rhetorical Markup Language (URML) was introduced by Reitter and Stede (2003) primarily to facilitate automatic RST parsing: The authors envisaged a pipeline analysis where subsequent modules can refine underspecified intermediate results of earlier modules. To some extent, this was implemented in the early SVM-based parser by Reitter (2003).

Our proposal here is that URML can serve to represent the complex ground truth derived from multiple expert annotations. In brief, URML is an XML format that regards every node of an RST tree as a data point to be described with various attributes and with elements pointing to the daughters (satellite and nucleus, or two nuclei). URML was designed to represent only binary trees, but that is in line with most existing implementations (including the RST parsers mentioned earlier), which usually work with binarized versions of the data.

An URML file for a text consists of three major blocks: an enumeration of all the RST relations in the set, a sequence of the EDUs of the text, and a sequence of node descriptions. This node-centric representation allows for *subtree sharing*: Competing analyses can be encoded to share common subtrees by referring to the same node ID. Other ways of underspecification are: (i) The name of the relation for a node can be specified or left out; as an intermediate variant, it is also possible to only state whether it is some mono- or multinuclear relation. (ii) The nucleus/satellite status of daughter nodes can be left open. (iii) The mechanism of *local ambiguity packing* allows for representations of alternative subtrees, whose root node IDs are specified to belong to the same *group*. Each relation node can also have a numerical score attribute, so that probabilities or preferences among the alternatives in a group can be encoded.

A limitation of URML lies in the fact that dependencies between different decisions cannot be represented. For example, the choice between relation R1 and R2 at node X might entail a preference for subtree S1 over S2 at one of X's daughter nodes. If such constellations need to be covered, the only way is to use alternative analyses, i.e., two (or more) complete URML graphs.[6]

5.2 Coding alternative expert trees in URML

We coded the RST trees resulting from our empirical study on the seven English texts (Section 3.3) in URML and found that all the phenomena of "legitimate disagreement" can be captured in this framework. In contrast to the original uses of URML envisaged by Reitter and Stede (2003), who focused on underspecification accompanying an incremental parsing technique, our goal here is to effectively represent genuine ambiguity. We thus make use of structure sharing and ambiguity packing, but not of unspecified relation names or types.

We demonstrate the functionality with an excerpt from one text of our study (from wsj_1117), looking at the expert annotators EE1 and EE2. For reasons of space and readability, we replaced the text segments with segment identifiers and show the two expert annotations in Figures 1 and 2. This is in fact one of the worst cases of disagreement that resulted from our study. At first sight the trees look quite different, but notice that: 1) both versions picked up a CONTRAST whose spans meet between S2 and S3; 2) both versions picked up a BACKGROUND whose spans meet between S5 and S6; and 3) the analyses for S3 – S5 are identical.

The disagreement thus amounts to the extension of the spans of the CONTRAST and BACKGROUND, the relation between S1 and S2, and the subtrees for S6 – S8. Here is an excerpt from the URML encoding of the node descriptions:

```
<parRelation id="N1a" group="N1"
     type="Contrast"
     annotator="V1"
     score="0.5">
   <nucleus id="N2a">
   <nucleus id="N5">
</parRelation>
<hypRelation id="N1b" group="N1"
 type="Background"
     annotator="V2"
     score="0.5">
   <nucleus id="N6b">
```

[6] For our present purposes, we did not encounter the need for this step.

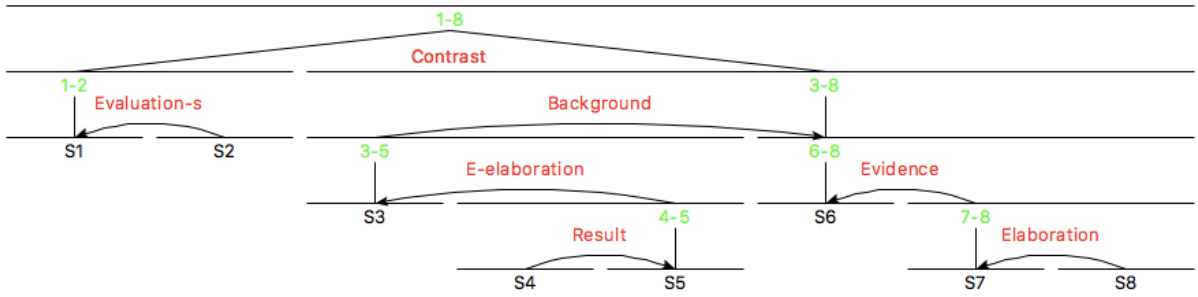

Figure 1: Annotation by EE1 for part of a corpus text (English study)

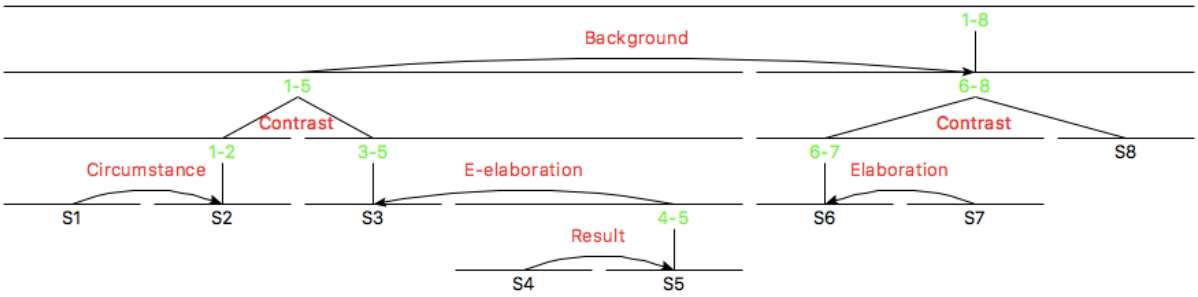

Figure 2: Annotation by EE2 for part of a corpus text (English study)

```
   <satellite id="N4">
</hypRelation>

<parRelation id="N4" type="Contrast"
     annotator="V2">
   <nucleus id="N2b">
   <nucleus id="N3">
</parRelation>
```

The declarations state that nodes N1a and N1b are alternative analyses provided by annotators EE1 and EE2. They are alternatives because they belong to the same group N1, and cover the same sequence of EDUs (S1–S8). In contrast, N4 does not belong to a group, i.e., it occurs only in EE-2's analysis. The first nucleus of both CONTRAST relations is an alternative of group N2 (not shown here), which represents the analyses for segments S1–S2.

In the same way, the other disagreements between EE1 and EE2 can be captured in the same URML representation, which thus plays the role of a "complex gold" annotation.

6 Summary

With two empirical studies, we demonstrated that annotator agreement depends on the amount of training and expertise the annotators have acquired. While this is hardly surprising, our next step is to differentiate between non-expert disagreement (some of which can arise from failure to adhere to the given guidelines, annotation flaws, or other human factors) and what we call "legitimate disagreement", i.e., that between expert annotators. Our proposal here is that competing expert analyses should be regarded as part of the "ground truth" in an annotated corpus. Besides differentiating between annotator expertise by means of quantitative measures, we undertook a first qualitative analysis of the types of disagreements encountered among experts. In future work, this needs to be elaborated.

The second point we made is that we can use the URML representation framework (which had originally been designed for a somewhat different purpose) to capture the disagreement in annotations in a single representation for a text. Our initial result is that the analyses used in the English study could all be mapped to URML and adequately represent the alternatives in the annotations. Here, the next step for us is to provide tools for automatic mapping (and merging) from the rs3 format of RSTTool to URML, and to devise ways of computing annotator agreement between a "new" annotator, or an RST parser for that matter, and the URML graph representing the "complex gold".

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

Chloé Braud, Barbara Plank, and Anders Søgaard. 2016. Multi-view and multi-task training of RST discourse parsers. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*. Osaka, Japan, pages 1903–1913.

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, pages 85–112.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, MA, pages 511–521.

Hugo Hernault, Hemut Prendinger, David duVerle, and Mitsuru Ishizuka. 2010. Hilda: A discourse parser using support vector machine classification. *Dialogue and Discourse* 1(3):1–33.

Mikel Iruskieta, Iria da Cunha, and Maite Taboada. 2015. Principles of a qualitative method for rhetorical analysis evaluation: A contrastive analysis english-spanish-basque. *Language Resources and Evaluation* 49(2):263–309.

Yangfeng Ji and Jacob Eisenstein. 2013. Discriminative improvements to distributional sentence similarity. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 891–896.

William Mann and Sandra Thompson. 1988. Rhetorical Structure Theory: Towards a functional theory of text organization. *Text* 8:243–281.

Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT Press, Cambridge/MA.

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

Michael O'Donnell. 2000. RSTTool 2.4 – A markup tool for Rhetorical Structure Theory. In *Proc. of the International Natural Language Generation Conference*. Mizpe Ramon/Israel, pages 253–256.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC)*. Marrakech, Morocco.

David Reitter. 2003. Simple signals for complex rhetorics: On rhetorical analysis with rich-feature support vector models. *LDV Forum* 18(1/2):38–52.

David Reitter and Manfred Stede. 2003. Step by step: underspecified markup in incremental rhetorical analysis. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC)*. Budapest.

Frank Schilder. 2002. Robust discourse parsing via discourse markers, topicality and position. *Natural Language Engineering* 8(2-3):235–255.

Manfred Stede. 2016. Rhetorische Struktur. In Manfred Stede, editor, *Handbuch Textannotation: Potsdamer Kommentarkorpus 2.0*, Universitätsverlag, Potsdam.

Maite Taboada and William Mann. 2006. Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies* 8(4):423–459.

Milan Tofiloski, Julian Brooke, and Maite Taboada. 2009. A syntactic and lexical-based discourse segmenter. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (Short Papers)*. Suntec, Singapore, pages 77–80.

Nynke van der Vliet, Ildikó Berzlánovich, Gosse Bouma, Markus Egg, and Gisela Redeker. 2011. Building a discourse-annotated Dutch text corpus. In *Proceedings of the Workshop Beyond Semantics: Corpus-based annotations of Pragmatics and Discourse Phenomena*. Göttingen, Germany, pages 157–171.