Discourse Parsing for German with new RST Corpora

Sara Shahmohammadi and **Manfred Stede**
Applied Computational Linguistics Group
Department of Linguistics
University of Potsdam, Germany
shahmohammadi@uni-potsdam.de stede@uni-potsdam.de

Abstract

For RST-style discourse parsing in German, so far there has been only one corpus available and used, the single-genre Potsdam Commentary Corpus (PCC). Very recently, two new RST corpora of other genres have been made available. In our work, we build a homogeneously-annotated German RST corpus by changing the PCC annotations so that they become compatible with the new corpora. We then run parsing experiments on different constellations of train/test splits over the three genres involved and report the results. A modified and streamlined version of the DPLP (Ji and Eisenstein, 2014) parser is prepared and made available, so that overall, the "resource situation" for German discourse parsing is notably improved.

1 Introduction

Rhetorical Structure Theory (Mann and Thompson, 1988) is a theory of discourse structure that models text coherence by a tree structure composed of discourse relations. Various corpora in several languages have been annotated within this framework since it was introduced. The Potsdam Commentary Corpus (PCC) (Stede and Neumann, 2014), was the first RST corpus for German, and just recently, two new German corpora have been annotated within this framework, viz. the APA-RST corpus (Hewett, 2023) of newspaper text, and a multimedia corpus of blogposts and podcast transcripts (Seemann et al., 2023). Although these two corpora followed the annotation guidelines of PCC for the most part, the authors modified the relation set, most importantly by adding the discourse relations *Same-unit* and *Attribution* (for compatibility with existing English corpora). This makes PCC incompatible with them at the levels of segmentation and relation set. In our work, we present a re-annotation of PCC texts, firstly in order to make it interoperable with the new corpora, and secondly because we observed that the annotations could also be improved in various other respects (which we will explain).

Taking the union of the three corpora, we perform discourse parsing using a modified version of the DPLP parser (Ji and Eisenstein, 2014). As there are three slightly different genres present in the corpora, we run experiments with different train/test splits in order to test generalizability. We find that the overall best model is obtained by training on PCC and the blogpost data. We make the re-annotated PCC data as well as the ready-to-use parser available to the community.[1]

After a brief introduction to RST and discussion of related work in Section 2, we discuss our PCC re-annotation and provide some corpus statistics in Section 3. Section 4 then gives details on the three corpora used in the parsing experiments, which we present in Section 5, and then conclude the paper in Section 6.

2 Background and Related Work

2.1 Rhetorical Structure Theory

RST (Mann and Thompson, 1988) models the structure of a text as a tree whose leaf nodes are given by the sequence of elementary discourse units (EDUs)[2] and whose internal nodes represent coherence relations holding between those leaf nodes and/or text spans (internal nodes of the tree) that are formed recursively. Coherence relations are built on the concept of nuclearity. If one discourse unit is more essential to the coherence relation than the other, it is deemed the nucleus (denoted by N); otherwise, it is deemed satellite (denoted by S). In Figure 1, for example, unit 4 and units 5-6 are the nucleus and the satellite, respectively. The majority

[1] Available at: https://github.com/mohamadi-sara20/pcc

[2] EDUs are the minimal parts of discourse. (Stede et al., 2017, p. 4). They are usually defined as clauses of the text. In Figure 1, for example, there are three EDUs.

Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024), pages 65–74
September 10-13, 2024

of relations are formed from elements with different weights (mononuclear relations), but some relations also connect multiple nuclei (multinuclear relations). The overall set of relations is not restricted to one closed list. Different corpora have proposed different relation sets; e.g., Mann and Thompson (1988) defined about 25 relation types in total, while the RST Discourse Treebank has 78 fine-grained relations, which are merged into 18 coarse-grained ones for automatic parsing purposes (Carlson et al., 2003, p. 32).

So-called "schemas" specify the constellations that may arise, e.g., whether multiple relation satellites can be attached to the same nucleus; if so, whether this is allowed only from one or from both directions in the text. In any case, relations always connect adjacent spans in such a way that no crossing dependencies arise.

2.2 The Potsdam Commentary Corpus

PCC is a freely available, multi-layer annotated corpus, whose latest revision of the RST layer was introduced by Stede and Neumann (2014). It consists of 176 commentary texts from a local German newspaper, i.e., it is a relatively small and deliberately homogeneous corpus. In our work, we inspected the RST layer and found some room for improvement, which will be described in Section 3.

2.3 RST Parsing for German

Early results for German RST parsing, using for the first time a support-vector machine and linguistic features for this purpose, had been presented by Reitter (2003). Recent results using neural systems were published by Braud et al. (2017), Liu et al. (2020), and Liu et al. (2021), who proposed multilingual parsers where the German part was trained and tested on PCC.

Braud et al. (2017), Liu et al. (2020) and Liu et al. (2021) respectively report performances of 0.80, 0.84, 0.84 on span detection; 0.54, 0.62 , 0.64 on nuclearity detection; and 0.35, 0.45, 0.47 on relation detection. In addition, Braud et al. (2023) report a performance of 0.32 on relation classification for German.

As a note of caution, we report that we tried to execute and reproduce the results of Braud et al. (2017) and Liu et al. (2021), but were unfortunately unable to do so and therefore turned to an alternative system.

Comparability is exacerbated by the fact that multilingual parsers are trained on large amounts of multilingual data, while we are dealing here with a single-language corpus, which is (still) comparatively small.

3 PCC-RST "reloaded"

3.1 Motivations for re-annotation

We found some improvable points in the RST layer of the PCC and thus made a number of changes to the annotations regarding segmentation, attachment point selection, and relations. For brevity, in the rest of the paper we call our revised RST layer *PCC**.

Segmentation. Occasionally, PCC annotators had used phrasal segments (*[And the town will hopefully not be brought down-][despite the bankruptcy of the State Development Corporation (LEG)][and occasional complaints within their own ranks.]*[3]). We decided to eliminate these, because their segmentation was not consistent. Phrasal segments were only kept in cases where a colon was present (*[Firstly:][The parking fees in the shopping area must be removed.]*[4]), as it was possible to remain consistent this way.

Further, since we aimed to add the *Attribution* and *Same-unit* relations to the data (see below), we had to modify the segmentation for these cases as well. For verbs of *Attribution*, we consulted a list of communication verbs provided by Tofiloski et al. (2009).[5]

Attachment Points. Non-adjacent attachments, which were present in several PCC trees, were avoided. Instead, we follow the suggestion of Egg and Redeker (2010): If all children have the same function, they are first joined as a list and then connected to the parent. For instance, the tree in the upper part of Figure 1 is turned into the tree in the lower part, because units 5 and 6 are both connected to their parent via an *Interpretation* relation. However, if children do not serve the same function, to avoid such connections, the adjacent child is prioritised and connected to the parent first, and then other children can be added. For an example, see Figure 2.

[3] From maz-8727.
[4] From maz-18914.
[5] https://github.com/sfu-discourse-lab/SL-Seg

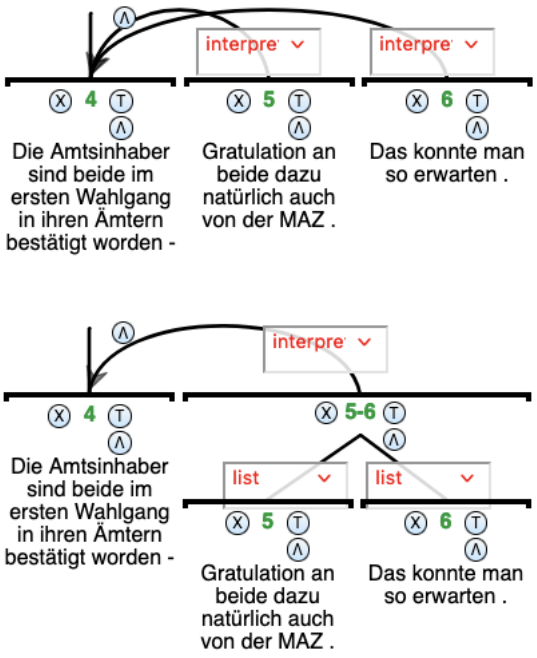

Figure 1: Non-adjacent connection, resolved by joining 5 and 6, because they are both connected to the nucleus via an *Interpretation* relation.

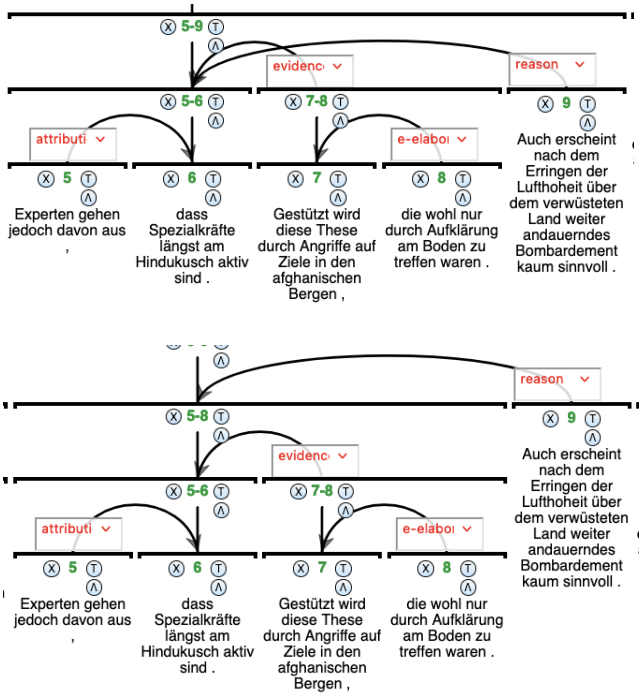

Figure 2: Non-adjacent connection, resolved by a hierarchical structure.

Schemas. We avoided the schema where a node is the parent of its left and right adjacent node at the same time. See Figure 3 for one such example. This step was taken because the annotation guidelines did not clearly specify the conditions for applying this schema, and we believe it is in fact not possible to avoid considerable ambiguity in such a formalization.

Relations. We made some changes to the relation list, by adding some new relations, eliminating some infrequent relations, and merging some relations. To see the definitions of the relations, consult Stede et al. (2017).

- *Attribution* and *Same-unit* were added to improve compatibility with existing large English RST corpora. The former relation is used for ascribing speech/thought content to a speaker ("John explained that the earth is flat"), while the latter handles parenthetical segments ("John explained – against his own belief – that the earth is flat"), which are in fact quite frequent in PCC.

- *Enablement*, which occurred only twice, was merged with the *Means* relation, following the practice of the two new German corpora mentioned above.

- *Unless* occurred only once in PCC and was removed from the inventory (the instance was re-annotated as *Condition*).

- *Disjunction* was merged with *Conjunction*, as it is not documented in the annotation guidelines.

- *Preparation*: Preparation usage was extended. We decided to use this relation whenever the satellite "consists of an introductory formula" (Stede et al., 2017, p. 19), announcing a nucleus, regardless of the information the satellite holds.

3.2 Inter-annotator Agreement

The data was annotated by the first author of this paper. Roughly ten percent of the corpus (18 texts) was double annotated. The second set of annotations were done by a student assistant, well-trained in RST.

The standard agreement measuring scores are span detection (S), nuclearity detection (N) and relation detection (R) scores, which are also widely used in evaluating automatic parsing results. These are reported in Table 1, which we obtained after converting our trees to parenthetical format using *discoursegraphs* (Neumann, 2015).

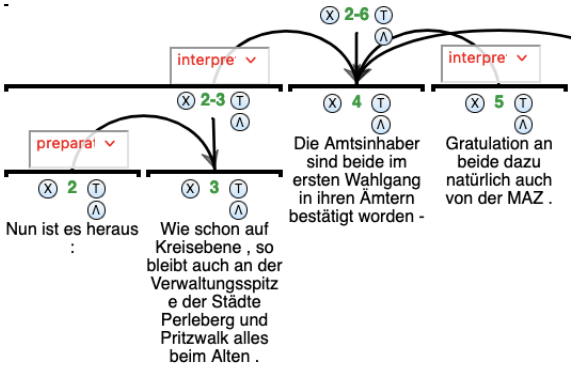

Figure 3: Parent with left and right children

In addition, we also report the inter-annotator agreement using the RST-Tace (Wan et al., 2019; Iruskieta et al., 2015) criteria in Table 2. RST-Tace is a tool that measures the agreement of RST annotations of different coders.

	S	N	R
PCC*	0.85	0.65	0.44

Table 1: Inter-annotator agreement. Results computed with the script released by Joty et al. (2015).

Agreement Ratio

	NR	RR	CR	AR	Average
mean	0.58	0.38	0.51	0.46	0.48
std	0.19	0.19	0.19	0.20	0.18

Agreement Kappa

	NK	RK	CK	AK	Average
mean	0.39	0.32	0.52	0.42	0.41
std	0.27	0.20	0.19	0.21	0.20

Table 2: Inter-annotator agreement computed by RST-Tace. NR, RR, CR, AR in the upper table denote Nuclearity Ratio, Relation Ratio, Constituent Ratio, Attachment-Point Ratio. NK, RK, CK, AK in the lower table denote Nuclearity Kappa, Relation Kappa, Constituent Kappa, and Attachment-Point Kappa.

3.3 Corpus Statistics

Taking a brief look at the changes in some relation groups, namely causal[6], additive[7], contrastive[8], context[9], and commentary relations[10] can give us an overview of how PCC and PCC* differ in terms of relations.

The proportion of additive relations overall changed drastically ($\chi^2 = 46.26$, p-value < 0.0001). A significant change is also present in causal relations ($\chi^2 = 8.59$, p-value $= 0.0034$), contrastive relations ($\chi^2 = 6.41$, p-value $= 0.0113$), relations of context ($\chi^2 = 6.55$, p-value $=0.0105$), as well as commentary relations ($\chi^2 = 14.18$, p-value $=0.0002$). On the other hand, relation groups like elaborative relations[11], conditionals[12] or summary, did not change significantly in proportion.

Figure 4 portrays the kernel density estimation of the relations whose proportions changed significantly. We have used Kernel Density Estimation from SciPy (Virtanen et al., 2020) to obtain them.

A more detailed comparison of our annotations with the original PCC annotations would be possible if the original PCC annotations were minimally changed – at least minor modifications at segmentation level – so that they can become comparable to ours, which can be the done in the future.

4 Data and Preprocessing

For our RST parsing experiments, we can now utilize the following German corpora: Blogposts from a multimedia corpus (Seemann et al., 2023), RST annotations of the original texts from the APA corpus (Hewett, 2023), and PCC. In addition to our new version PCC*, we also include the original PCC annotations in order to see if the parsing performance improves as a result of the re-annotation. The original PCC has 2,676 relations and 3,018 EDUs, while PCC* has 2,935 relations and 3,111 EDUs.

Blogposts. The blogposts come from several publishers (both commercial companies and scientific writers), and have been written for the weblog of various podcasts (Seemann et al., 2023). Each blogpost corresponds to one episode and usually either summarizes the content of the episode or more briefly announces the topic of discussion. In total, there are 78 RST trees, with 1,309 relations and 1,387 EDUs.

APA. This corpus contains 25 news articles from the Austrian news agency, along with their

[6] cause, result, justify, reason, reason-N, evidence, solutionhood, solutionhood-N, and motivation combined
[7] joint, conjunction, list, and disjunction combined
[8] antithesis, contrast, and concession combined
[9] background, circumstance, and preparation combined
[10] evaluation-n, evaluation-s, interpretation combined
[11] elaboration and e-elaboration combined
[12] condition, unless

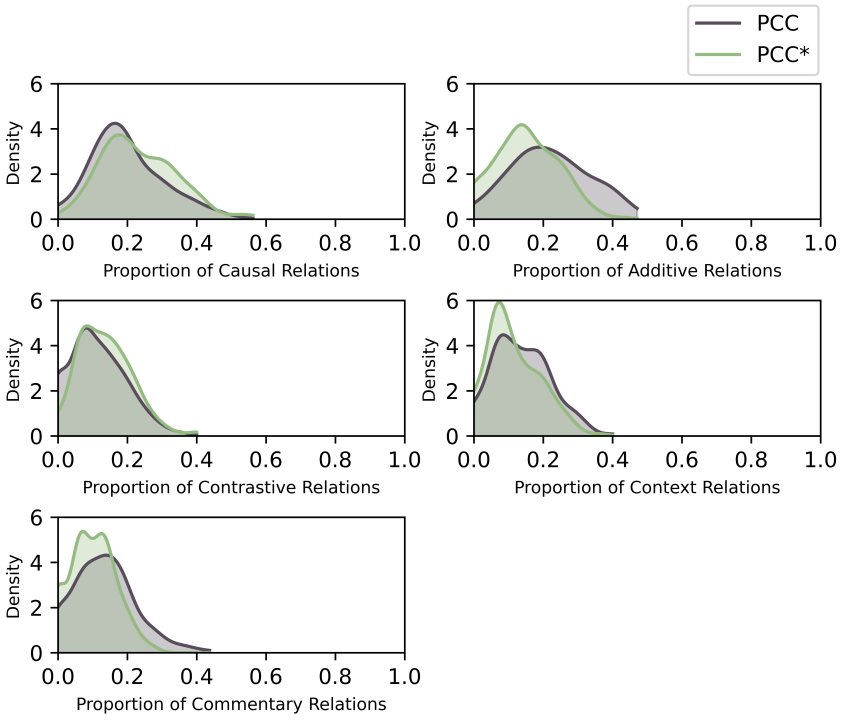

Figure 4: Kernel Density Estimation of proportions of causal, additive, contrastive, context and commentary relations

manually-produced simplifications to the language-learning levels of B1 and A2; hence in total there are 75 texts (Hewett, 2023). We only use the 25 original articles, because they are more comparable to the other corpora. RST trees have been annotated per paragraph, yielding a total of 61 trees with 852 relations and 938 EDUs.

Total data size. The original PCC corpus, as well as our version PCC*, contains 176 texts and thus the same number of trees. In total, there are now 78+61+176 = 315 different German texts (or for APA, paragraphs) with RST trees. They contain 5,096 discourse relations and 5,436 EDUs. In terms of relations, this represents a roughly 70 % increase in data size when compared to the original PCC RST corpus.

Preprocessing. In line with other parsing approaches, we use a Lisp-inspired parenthetical format of the RST trees as input to the parser. To obtain this format from the .rs3 XML standard used by the manual-annotation tools, we make use of the *discoursegraphs* library (Neumann, 2015). POS tagging and dependency parsing was done with stanza (Qi et al., 2020). All first segments of the original PCC trees, which are the headings of the text and not connected to the RST tree, have been removed.

5 Parsing Experiments and Results

5.1 Parser

We use the DPLP parser (Ji and Eisenstein, 2014), publicly available on github, because in comparison to others, it is well-documented, well-structured, lightweight, and rather easily adaptable to new data.

This shift-reduce parser is based on a set of linguistic and positional features, viz.: sentence ID, segment ID, word ID (in the sentence), word, POS tag, dependency label, dependency head for each EDU and also the two EDUs on top of the stack, and the EDU on top of the stack and at the front of the queue. In addition to these features, Brown clusters are also used as a means of contextualizing words.[13]

As a downside, the code was rather old, requiring discontinued versions of some libraries. To solve this issue, runtime dependencies are containerized in a Docker image and shared on Docker Hub[14]. The code was adapted by extending some

[13] https://github.com/mheilman/tan-clustering

[14] https://hub.docker.com/repository/docker/mohamadisara20/dplp-env

of the original modules and writing a number of new scripts.

5.2 Evaluation procedure

To evaluate parsing performance for the within-corpus experiments, we did a 5-fold cross validation, averaging over five runs for each of the five folds (first and second blocks in Table 3). For cross-corpus experiments, we train on the complete source corpus and test on the complete target corpus, averaging over five runs (third, fourth and fifth block in Table 3).

To divide the data into five partitions, we randomly shuffled the data and created five batches from PCC, PCC*, and APA data. For blogposts, however, we created a stratified sample, i.e., we partitioned the data such that texts from their different publishers are represented proportionally. This decision seemed advisable because we observed great linguistic variability among the texts from these different sources.

5.3 Results

All our results are collected in Table 3.

Within-corpus evaluation. The first block of the table shows the results of training and testing on each corpus separately (5-fold CV).

As evident, using PCC* annotations, the performance has improved on nuclearity and relation detection. This indicates that we have managed to improve the annotations and reduce inconsistency to a certain degree. However, part of the improvement is due to the addition of the *Attribution* relation, which is rather easy to learn due to its syntactical and lexical features.

On blogposts, performance is notably higher than on PCC. We assume that this is due to the lower complexity of these texts: They are shorter, and overall have either the straightforward purpose of introducing, or (less frequently) summarizing a podcast episode. This leads to more formulaic structures than in editorials, which exhibit relatively high stylistic and argumentative variation.

For APA texts, results are in the range of PCC, which at first sight hints at similarities between the rhetorical structures of newspaper texts, irrespective of their degree of subjectivity.

Finally, we ran a test on the complete corpus of 315 texts and found that compared to when only including PCC* data, the performance improves only minimally. The cross-corpus experiments can explain the potential reasons to some extent.

	S	N	R
PCC	0.77	0.52	0.28
PCC*	0.77	0.55	0.35
Blogs	0.81	0.61	0.40
APA	0.81	0.56	0.32
APA+Blogs+PCC*	0.78	0.56	0.36
Blogs+PCC* → PCC*	0.77	0.54	0.34
Blogs+PCC* → Blogs	0.82	0.64	0.43
PCC* → APA	0.77	0.48	0.24
APA → PCC*	0.75	0.47	0.24
APA → Blogs	0.78	0.53	0.31
Blogs → APA	0.76	0.45	0.21
PCC* → Blogs	0.80	0.59	0.39
Blogs → PCC*	0.76	0.50	0.28

Table 3: Parser performance results for the various train/test settings (see Section 5.3). The arrow notation is "training corpus" → "test corpus".

Cross-corpus evaluation. Since the parsing performances differ somewhat between the corpora, we decided to explore how well a model learnt on one corpus would predict the structure on another.

Firstly (second block of the table), we found that adding blogs to the PCC* training data does not increase the performance on PCC*. However, we achieve the overall best results by testing the combined PCC* and blogs model on blogs, among the individual corpora as well as the pairings. This may also be an effect of the corpus size: PCC parsing does not benefit as much from the addition of (small) out-of-domain data as the blog parsing does from adding a larger amount of out-of-domain data.

The third block of Table 3 shows results for the PCC*/APA pair. The two results are very close to each other and at the same time lower than those of the individual corpora, so it seems that neither is able to generalize well to the other. This may contradict the impression of their similarity that we formulated for the first experiment above. It can also partly be due to the fact that PCC* annotations cover the complete text, while this is not true for APA.

Finally, the fourth and fifth blocks of the table give the results for the pairings with the "top performing" individual corpus, i.e., blogposts. Results are higher than for PCC*/APA throughout, with the odd exception of a rather low relation recognition

in the Blogs→APA setting (for which we have no explanation hypothesis). The much better results for PCC*→Blogs in comparison to Blogs→PCC* can be an effect of training corpus size, given that PCC* has twice as many texts as the Blogs corpus.

5.4 Error Analysis

Table 5 shows a sample confusion matrix from PCC*. The rows signify the true labels and the columns signify the predicted labels. We can see that relations such as *Concession*, *Conjunction*, *List*, *E-elaboration*, as well as *Attribution* and *Reason* have been recognized better.

The confusion matrix is, however, rather sparse. Since it can be beneficial to see the performance on relation groups as well, we trained another model by merging all additive, causal, commentary, context, and elaboration relations.[15] to see on a more general level what relations are better recognized as well as what relation groups are confused with each other. Some relations, such as *Attribution* or *Same-unit*, were kept as they were, since we believe they do not have enough in common with each other or with other groups.

Table 4 represents the confusion matrix of a model with merged relations. As the table shows, additive, conditional, and context relations are in general detected more reliably, while contrastive relations are often confused with additives and causals. Causal relations are also confused with commentary or elaborative relations. Less frequent relations such as *Sequence* were also often confused with additives.

It should be noted that although the merged model can give us a more general overview, it must be looked at with care, since the numbers in most cases are still not high enough to draw solid conclusions.

6 Conclusion

So far, the only resource for RST parsing in German has been the Potsdam Commentary Corpus. Prompted by the recent release of two additional RST corpora, we created a unified resource by changing the PCC annotations, on the one hand for compatibility with the new corpora, on the other hand for improving certain shortcomings in the existing annotations. Using the new homogenous set of corpora, we performed various RST parsing experiments with different train/test splits, and report the results here as baselines for further studies. We showed that parsing performance improves for nuclearity and relations when moving from the original PCC to our PCC* trees, which may indicate higher annotation consistency.

Furthermore, we are making a revised version of the DPLP parser available (ready to use for German), as well as the re-annotated PCC texts.

In future work, the enlarged data set can be used to test other parsing architectures. In addition, the old and new versions of the PCC RST layer can be used to study the phenomena of ”legitimate disagreement” in discourse annotation – a topic that has recently become popular also under the label ”perspectivist approaches to NLP”. This can include approaches to systematically including both variants in training parsing models.

Limitations

As hinted at in the conclusion, RST annotation is known to be subjective, and thus we do not regard our new PCC annotations as ”the single ground truth”; instead it represents a set of possible text interpretations. The corpus that can now be used for parsing has more genre variety than the PCC had, but is still relatively homogeneous (opinion articles, news, well-edited blogs); additional genre diversity could be achieved, for example, by adding more user-generated text, e.g. from social media.

Ethics Statement

All annotations were performed by the first author of this paper and reviewed by the second author of this paper. One regularly-paid student assistant annotated part of the data.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 317633480 – SFB 1287. We would like to thank Elisa Lübbers for helping with annotations and the reviewers for their thoughtful comments.

[15] The groups have been specified in Section 3.3.

	additive	attribution	causal	commentary	condition	context	contrastive	elaboration	means	purpose	sameunit	sequence	summary-restatement
additive	71	2	8	6	1	3	3	9	0	1	3	0	0
attribution	3	12	0	0	3	1	0	0	0	1	0	0	0
causal	4	0	26	11	0	2	5	11	0	0	0	0	0
commentary	4	0	6	8	0	3	3	8	0	0	0	0	0
condition	0	0	0	0	12	1	0	0	0	0	0	0	0
context	2	0	1	0	2	15	1	3	0	0	0	0	0
contrastive	12	0	15	0	2	2	28	6	0	0	0	0	0
elaboration	8	0	8	1	0	2	1	26	0	0	1	0	0
means	0	0	0	0	0	0	1	0	0	0	0	0	0
purpose	0	0	1	0	1	0	0	1	0	4	0	0	0
sameunit	2	0	2	0	0	1	2	3	0	0	7	0	0
sequence	5	0	0	0	0	0	0	0	0	0	0	0	0
summary-restatement	2	0	0	1	0	0	0	1	0	0	0	0	0

Table 4: Confusion matrix (merged relations)

	antithesis	attribution	background	cause	circumstance	concession	condition	conjunction	contrast	e-elaboration	elaboration	evaluation-S	evidence	interpretation	list	means	preparation	purpose	reason	reason-N	restatement	result	sameunit	sequence	solutionhood	solutionhood-N	summary
antithesis	5	0	1	0	0	0	1	0	1	2	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
attribution	1	12	0	0	0	0	2	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
background	0	0	3	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
cause	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
circumstance	0	0	0	0	4	0	3	0	1	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
concession	1	0	0	0	1	19	3	2	0	0	1	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
condition	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
conjunction	1	0	0	0	0	0	0	51	0	0	1	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0
contrast	0	0	0	0	0	4	0	5	0	1	1	1	0	1	6	0	0	1	5	1	0	0	0	0	0	0	0
e-elaboration	0	0	0	0	0	0	0	1	0	23	1	0	0	0	3	0	1	0	0	0	0	0	1	0	0	0	0
elaboration	1	1	1	0	0	1	0	2	1	1	5	0	1	1	1	0	0	0	2	0	0	0	0	0	0	0	0
evaluation-S	0	0	1	0	0	2	0	1	1	0	1	0	0	1	2	0	1	0	2	0	0	0	0	0	0	0	0
evidence	0	0	0	1	0	0	0	0	2	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0
interpretation	0	0	1	0	0	1	0	1	6	1	1	1	0	2	2	0	1	0	3	0	0	0	0	0	0	0	0
list	0	3	2	0	0	2	1	8	3	4	3	1	1	1	25	0	1	1	3	0	0	0	4	0	0	0	0
means	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
preparation	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0
purpose	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0
reason	0	0	0	0	0	1	0	1	1	5	0	1	0	4	7	0	0	0	13	0	0	0	0	0	0	0	0
reason-N	0	0	0	0	0	0	0	0	0	0	1	1	0	2	3	0	0	0	3	0	0	0	0	0	0	0	0
restatement	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
result	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0
sameunit	0	0	0	0	0	2	1	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	7	0	0	0	0
sequence	0	1	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
solutionhood	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
solutionhood-N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
summary	0	0	0	0	0	0	0	1	0	1	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 5: Confusion matrix (unmerged relations)

References

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST discourse parsing. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 292–304, Valencia, Spain. Association for Computational Linguistics.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol T Rutherford, and Amir Zeldes. 2023. The disrpt 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023), pages 1–21. ACL: Association for Computational Linguistics.

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.

Markus Egg and Gisela Redeker. 2010. How complex is discourse structure? In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, 19-21 May, 2010, pages 1619–1623. European Language Resources Association (ELRA).

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.

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, 49:263–309.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers), pages 13–24.

Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. Computational Linguistics, 41(3):385–435.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2020. Multilingual neural RST discourse parsing. In Proceedings of the 28th International Conference on Computational Linguistics, pages 6730–6738, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy F Chen. 2021. Dmrst: A joint framework for document-level multilingual rst discourse segmentation and parsing. arXiv preprint arXiv:2110.04518.

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.

Arne Neumann. 2015. discoursegraphs: A graph-based merging tool and converter for multilayer annotated corpora. In Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015), pages 309–312.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.

David Reitter. 2003. Simple signals for complex rhetorics: On rhetorical analysis with rich-feature support vector models. Journal for Language Technology and Computational Linguistics, 7(1):38–52.

Hannah J Seemann, Sara Shahmohammadi, Tatjana Scheffler, and Manfred Stede. 2023. Building a parallel discourse-annotated multimedia corpus. Proceedings of the 10th International Conference on CMC and Social Media Corpora for the Humanities, 8(3):17.

Manfred Stede and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 925–929, Reykjavik, Iceland. European Language Resources Association (ELRA).

Manfred Stede, Maite Taboada, and Debopam Das. 2017. Annotation guidelines for rhetorical structure. Manuscript. University of Potsdam and Simon Fraser University.

Milan Tofiloski, Julian Brooke, and Maite Taboada. 2009. A syntactic and lexical-based discourse segmenter. In Proceedings of the ACL-IJCNLP 2009 conference short papers, pages 77–80.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17:261–272.

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.