Parsing Discourse Structures for Semantic Storytelling: Evaluating an efficient RST Parser

Pia Linscheid[1][0000-0001-7231-0137], Peter Bourgonje[2][0000-0003-3541-0678], and Georg Rehm[2][0000-0002-7800-1893]

[1] Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
[2] DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany

Corresponding author: pia.linscheid@hu-berlin.de

Abstract. We explore if an efficient RST parser achieves as good results with a register-balanced data set as with register-specific data of the training and test phase. In this paper, we present the evaluation of an efficient parser that identifies semantic connections between textual elements within English texts based on Rhetorical Structure Theory (RST). The parser was tested on data from the Georgetown University Multilayer Corpus (GUM). Its output was compared to the manual GUM RST annotations serving as gold standard by using the evaluation tool RST-Tace. This investigation is motivated by the underlying question if the parser can be considered for the processing of discourse structure for a broader application scenario we call "semantic storytelling", which is an approach for processing content to extract and analyse information as well as generate storylines to support knowledge workers such as journalists or scholars. Our results are based on a relatively small amount of data, yet they show a clearly recognizable difference between manual and parsed annotation of discourse structure. This is due to a number of factors, among them the assignment of default labels for ambiguous relations. The results demonstrate a need for improved disambiguation methods when it comes to assigning rhetorical relations.

Keywords: Rhetorical Structure Theory, Discourse Parsing, RST Parser, RST-Tace, Georgetown University Multilayer Corpus, Semantic Storytelling.

1 Discourse Structure and Semantic Storytelling

With the increase of digitalization in the workspace, the demands on the human and the machine are growing. In many industries, the demand for automated processing of big data and AI-supported text generation is increasing, too. Using text and data mining methods, text segments that are relevant for a specific topic can be extracted from a document collection. The question of how best to combine such textual fragments is not trivial. One important aspect of AI-driven text generation is (semi-)automatic production of suitable text structures. The discourse structure of a text reflects the content network of sentences and paragraphs [Ste11]. In other words, these links between text fragments (as semantic units) create a coherent flow of information.

Knowledge about how to order events and other types of semantic information are central aspects of a semantic storytelling system. The development of such a system is one of the goals of the innovation project QURATOR, which aims to develop a sustainable AI technology platform to support human experts in the process of curating digital content [RBHK$^+$20; RZMO$^+$20]. We aim to provide one component of such a system, aimed at supporting knowledge workers like journalists or scholars by extracting facts from a document collection and aligning these informational units to a storyline [RZMS19; RZBO$^+$20]. For this goal, insights into discourse structure and semantic links between informational units are crucial. A well-known approach to plot discourse structure as an alignment of informational units is the Rhetorical Structure Theory (RST) [MT88]. Based on RST (see Section 2), several corpora and discourse analysis tools have been developed [Neu15; Zel17]. The emerging QURATOR system would greatly benefit from the inclusion of such a tool into its semantic storytelling prototype [RMBS$^+$18]. The current paper examines the performance of an efficient RST discourse parser [HS15] running on register-balanced data from the Georgetown University Multilayer Corpus (GUM) [Zel17] followed by an evaluation of its accuracy with the evaluation tool RST-Tace [WKLS19]. Register-balanced means that the corpus is balanced according to different text types. Accordingly, no bias is to be expected due to certain text type-specific occurrences. This paper is a summary of [Lin20].

We will first outline the theoretical framework, previous work and relevant resources in Section 2. Section 3 describes the concept and procedure. The results are presented in Section 4, discussed in Section 5, and summarized in Section 6.

2 Framework, Research and Resources

In order to be able to analyze and generate texts and storylines automatically, a comprehensive understanding of text structure and its formalized or formalizable mapping is needed. With the help of computer-aided text linguistics, this requirement can be addressed. Text linguistics deals with the question of what discourse (or: a text) is and how it is structured.

Most texts consist of more than one sentence, yet a text can also consist of a single sentence [WEK12], illustrated by single word phrases like *Sorry!*. In this context, discourse and text are used synonymously. A further assumption is that discourse structure is characterized by the fact that the overall meaning of a text is greater than the sum of the meanings of the individual sentences due to their relatedness to each other [WEK12]. What constitutes a sentence descriptively is disputed in scientific discourse. One way to avoid the problem of defining sentences sufficiently and still be able to deal with texts is to look at semantic discourse units and thus become more independent of syntactic surface structures. Among the many proposals that deal with the subdivision of a text into units, we chose the one suggested in RST [MT88]. In RST, these textual building blocks are called Elementary Discourse Units (EDU). Thus, RST allows us to approach a text from an informational, semantic level and visibly grasp how EDUs are linked together – and to see how information “flows” through the text. This is what we want to be able to take up and use for the Semantic Storytelling system.

The visualization of EDUs and their content-related connection within the RST framework can be described and illustrated in brief as follows. In RST, EDUs are set in relation to each other and are usually hierarchized. The more relevant EDU is considered the nucleus and the less relevant the satellite. The relational links are represented as a tree structure, with the EDUs representing the leaves. There are multiple types of defined semantic relations in a given RST implementation. RST sets can have up to two dozen different relations. Most relations between EDUs are asymmetrical, consisting of nucleus and satellite (see Fig. 2, *Background*). In case that EDUs are equally important, some relations are symmetrical consisting of two nuclei, so-called multi-nuclear relations (see Fig. 2, *Joint* and *Sequence*). A sketch of the relation between EDUs is shown in Fig. 1 [according to MT88].

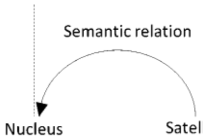

Fig. 1. Outline of the basics and technical terms of RST. The more relevant EDU (nucleus) is semantically related to the less important EDU (satellite).

This theory of analyzing rhetorical and content structure of texts is too extensive to detail here, but also too abstract not to illustrate when applied in the exploration. Therefore, a sample RST analysis is presented in Fig. 2 using a text excerpt from the GUM corpus [Zel17]. Furthermore, names and definitions of some RST relations are explained and linked to examples from the GUM corpus in Table 1.

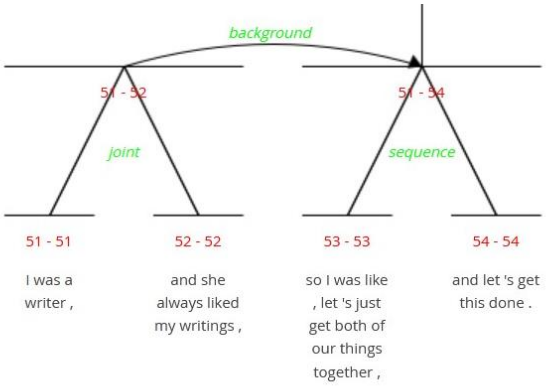
background
51 - 52
joint
53 - 54
sequence
51 - 51
I was a writer ,
52 - 52
and she always liked my writings ,
53 - 53
so I was like , let 's just get both of our things together ,
54 - 54
and let 's get this done .

Fig. 2. Sample analysis according to RST: excerpt of the GUM corpus (CC-BY attribution license); see the whole analysis under https://korpling.german.hu-berlin.de/annis3/?id=dc6232c5-6227-41e6-99e9-d6f57ccc8b70

Short meaning-bearing sentence segments or whole sentences are considered minimal units or EDUs (see Fig. 2). They are numbered and positioned in relation to each other. These relations can be different. If one EDU or complex constituent is considered more important than the other, there is a hierarchized relation, as in *Background*. If EDUs or complex constituents are considered equally important, they a labeled with multi-nuclear relations, as in *Joint* and *Sequence*. The difference between the relations *Joint* and *Sequence* is small: both give information (e.g., about facts). However, with *Sequence* these are in a temporal sequence and with *Joint* in non-specific relation (no enumeration, no sequence).

A set of RST relations is defined by determining certain conditions for the EDUs and their content connection. Table 1 shows some of the definitions.

Table 1. A selection of RST relations: Names and excerpts of the definitions [Ste16], as well as examples from ANNIS[3] [KZ16].

	Relations	**Definition**	**Example from GUM**
hierarchical relations	*Background*	The understanding of the satellite (S) facilitates the reader's understanding of the content of the nucleus (N); S contains orienting background information without which N would not be understandable or would be difficult to understand. […]	https://korpling.german.hu-berlin.de/annis3/?id=dc6232c5-6227-41e6-99e9-d6f57ccc8b70
	Cause	The circumstance described in N is caused by the circumstance described in S. […]	https://korpling.german.hu-berlin.de/annis3/?id=b1ed29fb-f996-46a1-a379-7a87dec12cd1
	Elaboration	S provides more precise information or details about the content of N – but not just about a single item mentioned in N. N precedes S in the text. […]	https://korpling.german.hu-berlin.de/annis3/?id=d4632d05-fedc-4646-b215-54bc3f0f390b
	Result	The circumstance described in S is caused by the circumstance described in N. […]	https://korpling.german.hu-berlin.de/annis3/?id=56a677b7-5c4d-41d1-a608-83eece518d71
multi-nuclei relations	*Sequence*	The nuclei describe facts of the world that take place in a certain temporal sequence. […]	https://korpling.german.hu-berlin.de/annis3/?id=13edbb54-8be4-4f87-b3db-4e4884868d68
	Joint	The nuclei give separate information, but do not stand in any clearly identifiable semantic or pragmatic relation to each other, nor do they have the character of an enumeration together. Nevertheless, there is a coherent connection because they serve the overarching textual function equally well.	https://korpling.german.hu-berlin.de/annis3/?id=6fbc1538-d964-440a-a7c9-1fb99f92b8ff

Examining these definitions, it becomes apparent that the definitions of RST relations have grey areas and overlap with one another [DTS17, Ste18]. Likewise, the hierarchy of information according to its relevance is governed by a certain subjectivity. Therefore, the same sequence of EDUs may have different legitimate analyses or annotations, as illustrated in Fig. 3. Despite precise definitions of the relations, ambiguity can only rarely be avoided, since an RST analysis should systematically depict the result of a (naturally subjective) interpretation of the text [Ste18]. For this reason, RST nonetheless enables a consistent, though not unambiguous, analysis and mapping of discourse structure.

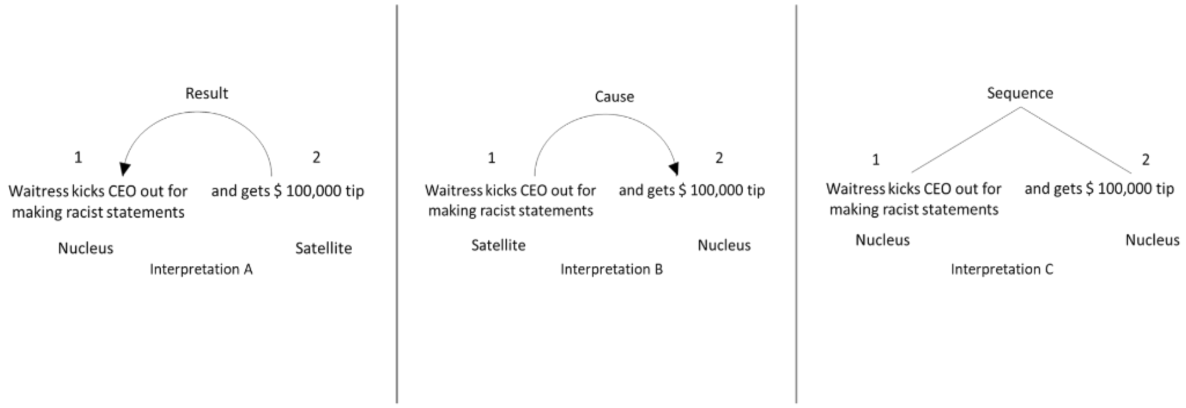

Fig. 3. Three different but valid annotations of the relation between two EDUs illustrating the subjective and definitorial vagueness of RST exemplified by a news ticker entry, originally in German: ‘Kellnerin schmeißt CEO wegen rassistischer Äußerung raus und erhält fast 100.000 Dollar Trinkgeld' from https://www.stern.de/news/ (published on 2020-08-20).

Another theoretical approach used in this work is the black box approach, which disregards the internal structure and coding of a program and refers exclusively to its functionality in terms of the output [Nid12].

Regarding the research, one crucial aspect of RST-based resources like annotated corpora is that their construction is highly time- and resource-consuming. The same applies to the manual evaluation of RST analyses [WKLS19]. Therefore, the development of sustainable tools for automated RST analyses is of great importance both for research and for possible implementations in projects like QURATOR. The existing corpora with which the developed tools are trained and tested are often limited in their scope and application due to these circumstances. Most discourse processing tools for English are based on one or two prevalent corpora: namely the Penn Discourse Treebank (PDTB) [PWLA19] and the RST Discourse Treebank (RST DT) [CMOM02]. PDTB comprises 53600 tokens from Wall Street Journal articles. RST DT, the main English language RST corpus [Ste18], also includes articles from the Wall Street Journal. There is a sizeable overlap between the articles of the two corpora [Ste18]. Although they are relatively large and well established, they are register-specific. Register is understood as text type in this context. The register of a text is assumed to have major impact on the structure of a text [Ste18].

Therefore, it can be assumed that a database containing one single register might not allow reliable conclusions to be drawn about text structures of other or all registers. Which in turn raises the question whether this corpus is sufficient for the development of tools alone, if the tools are to be used generally (i.e., also register-independently) for the RST analysis of texts. Regarding a possible industry-focused implementation in a project such as QURATOR, the question arises whether tools achieving good results in a register-specific manner would also be suitable for non-register-specific text analysis.

However, for the last ten years many RST-based text processing tools have been developed. A good overview is given on the *RST Workbench*.[1] These include several RST parsers, such as HILDA [HPDI10] and CODRA [JCN15]. Shift-reduce discourse parsers were also developed like DPLP [JE14]. They can learn to improve RST parsing by means of surface structures [JE14]. The fastest RST parser tested with comparable accuracy and robustness to others was the shift-reduce RST parser from Heilman and Sagae [HS15]. As performance is relevant for our purposes, this parser was chosen to be evaluated in this project. It was trained and tested using RST DT and PBDT. Within the scope of the register-specific corpus it achieved state-of-the-art accuracy at high speed [HS15], see Table 2.

Table 2. Test set of discourse saving performance as F1-scores (%); the Human Agreement from Penn Treebank is serving as gold standard, table with data from [HS15]

Results of RST Parsing	Span	Nuclearity	Relation
Heilman & Sagae (2015)	83.5	68.1	55.1
Li et al. (2014a)	84.0	70.8	58.6
Joty et al. (2013)	82.5	68.4	55.7
Joty & Mochitti (2014)	–	–	57.3
Feng & Hirst (2014)	85.7	71.0	58.2
Li et al (2014b)	82.9	73.0	60.6
Ji & Einstein (2014)	81.6	71.0	61.8
Human agreement	88.7	77.7	65.8

To test the performance of the parser for cross-register text processing, we used the GUM Corpus, a recent, register-balanced and relatively large RST-annotated corpus comprising 130,000 tokens from 148 English texts across eight registers [Zel17].

Several additional tools have been developed for automating the evaluation of RST annotations. The comparison of multiple RST annotations of the same text involves three challenges: first, the actual parsing of RST trees to determine their structure, second, adequate evaluation and comparison methods have to be found and defined, and finally, this methodology has to be sufficiently well applied [WKLS19]. Facing this challenge, RST-Tace, an up-to-date tool for the evaluation of RST annotations [WKLS19][2], was developed to overcome weaknesses of previous automated evaluation tools like RSTeval [MP09].

[1] Available online at https://github.com/arne-cl/rst-workbench

[2] Available online at https://github.com/tkutschbach/RST-Tace

3 Procedure

This study aims to evaluate an efficient RST parser with regard to the question if it achieves state-of-the-art results when processing texts across registers. The parser is tested on data of the register-balanced GUM corpus and then compared to the manual RST annotations from GUM serving as gold standard. For the analysis the evaluation tool RST-Tace is used. Fig. 4 illustrates the multi-step procedure of the work.

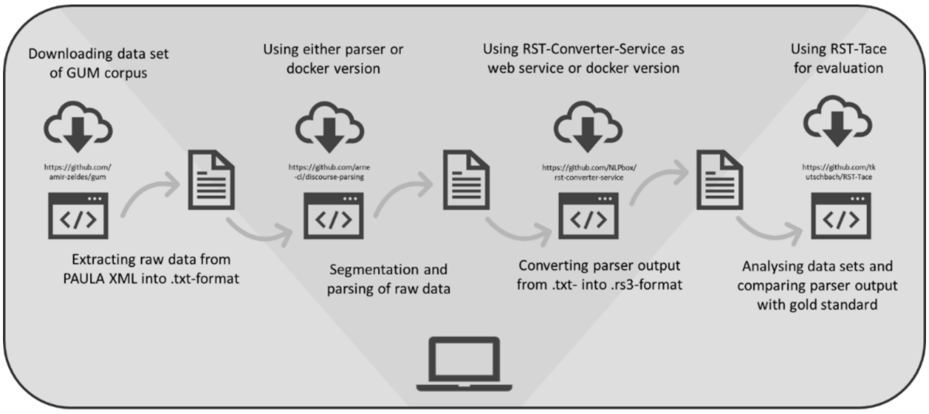
Downloading data set of GUM corpus
https://github.com/amir-zeldes/gum
Extracting raw data from PAULA XML into .txt-format
Using either parser or docker version
https://github.com/arne-cl/discourse-parsing
Segmentation and parsing of raw data
Using RST-Converter-Service as web service or docker version
https://github.com/NLPbox/rst-converter-service
Converting parser output from .txt- into .rs3-format
Using RST-Tace for evaluation
https://github.com/tkutschbach/RST-Tace
Analysing data sets and comparing parser output with gold standard

Fig. 4. Schematic representation of the sequence of work steps.

Several formats, conversion steps and tools were needed to test the parser on the GUM texts and to compare its output with the GUM annotations in RST-Tace. The evaluation with RST-Tace is based on a recent method that uses the agreement of central constituents [IDT15]. Inter-annotator agreement (IAA) and matching ratio (MR) are evaluated with respect to nuclearity, relation, constituent and attachment point. IAA is calculated by means of Kappa (κ) and MR by means of F1.

The nuclearity variable records whether EDUs were classified differently or equally as nucleus or satellite in the annotations – it records the direction of the relation (from satellite to nucleus). The relation variable reflects whether there is agreement or differences in the assignment of the relation between EDUs. The constituent variable records the unit(s) in which the satellite (or a nucleus in the case of multi-nuclear connections) is located. The variable called attachment point is superior to the constituent and compares the annotations with each other with respect to the unit(s) in which a constituent is linked to [WKLS19].

In the workflow data loss occurred due to processing errors for many files.[3] The data loss could be attributed to errors of the parser, but it could also be attributed to hidden issues within the large number of steps in the workflow. Due to the black box approach, however, these conditions cannot be verified in this paper. The results presented below are, thus, based on a rather small amount of data.

[3] From 130 files at the beginning, only 15 parser files could be analyzed at the end, and 12 of them could be compared to the gold standard.

4 Results

The small amount of data reflects a relatively uniform picture. With respect to the IAA,[4] the results of the annotation comparison give different mean values for nuclearity (0.206; SD: 0.093), relation (0.142; SD: 0.058), constituent (0.079; SD: 0.033) and attachment point (0.015; SD: 0.045). The MR also shows varying averages with respect to nuclearity (0.201; SD: 0.060), relation (0.091; SD: 0.037), constituent (0.112; SD: 0.016) and attachment point (0.087; SD: 0.037). For the IAA, the differences in nuclearity and relation are greater than for constituent and attachment point. With regard to MR, the only major difference is in terms of nuclearity. The distribution of the values as well as the mean values and standard deviations are shown in Fig. 5 and Fig. 6.

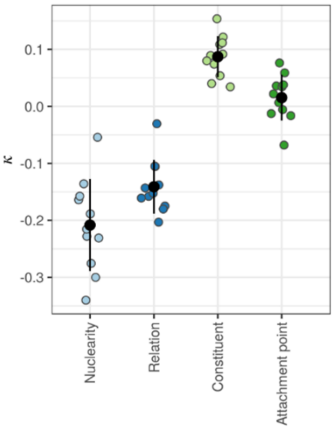

Fig. 6. Distribution by variables in the IAA, *κ* (mean value as black dot, SD as interval marked in black)

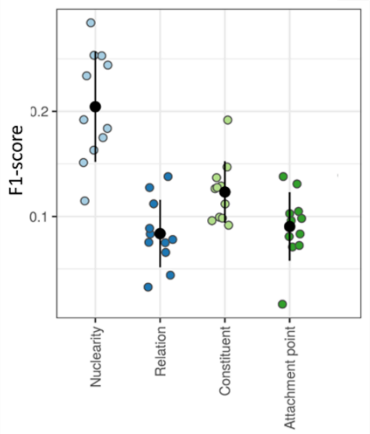

Fig. 5. Distribution by variables for MR, F1-score (mean value as black dot, SD as interval marked in black)

An indicator for the parser's performance are the F1-scores from Table 2 by converting the relative values to absolute ones and by comparing these values with those of the MR. Here, the category span should be compared with that of attachment point. In Table 2, the F1-scores are generally much higher than in Fig. 6. Yet, differences between the syntactic and the semantic categories are apparent both in Table 1 (Agreement in F1-score) and in Fig. 5 (IAA measured in *κ*). In Table 2, the F1-scores of the syntactic category span (with an F1-score of 0.835) are 15-28% higher than the semantic-hierarchical categories nuclearity (0.681) and relation (0.551). In Fig. 6, all F1-scores are low, but here the category nucleus (0.206) has slightly better values than the others (relation: 0.091; const.: 0.112; att. point: 0.087).

[4] The measurement of the IAA in this case does not refer to the comparison of the annotations of several (human) annotators but to the annotations of one annotator and an automated one. Nevertheless, the term IAA is used here in favor of the value specifications of RST-Tace.

Fig. 7 shows the ten most frequent relations using a confusion matrix. The absolute frequency of agreement of the relation annotation is shown and its relative frequency is highlighted in color. According to this, in case of a perfect match, the highest value would have to be recorded on the diagonal and the values of the remaining cells would have to be zero. This means, the more the higher values (and thus the colored highlighting) are moving away from the diagonal, the less the relations are matching. Again, it is clearly visible that there is a considerable deviation of the relations between the gold standard and the parser output. In 14% of the cases (n=89) the parser assigned the relation *Elaboration*, in which an elaboration was annotated in the gold standard as well. The confusion matrix in Fig. 7 also illustrates that the parser has annotated the label *Elaboration* at considerably more nodes where other relations occur in the gold standard. This can be seen in the line *Elaboration*, which shows that in the gold standard the relations *Concession*, *Evidence*, *Joint*, *Preparation*, *Same-unit* and *Sequence* were annotated at these points instead of *Elaboration*.

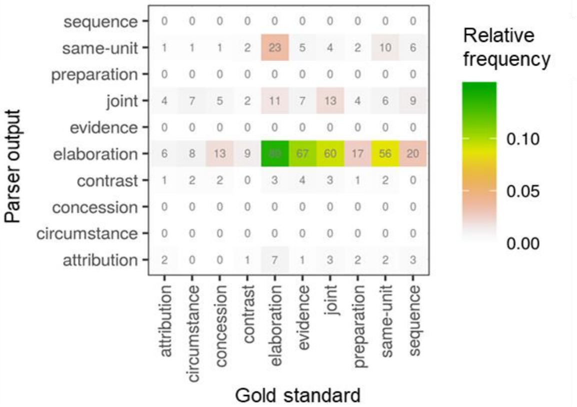

Fig. 7. Confusion matrix showing the distribution of the most frequent relations in the parser output and gold standard

In Fig. 7, the column *Elaboration* shows that the parser has annotated *Same-unit* or *Joint* in a percentage of one digit, in which the gold standard contains the relation *Elaboration*. The colors appear weak in the diagonal, stronger in the horizontal and weak in the vertical, creating rake-shaped symmetries between the distributions of *Elaboration*, *Joint,* and *Same-unit*. This suggests that there is a certain amount of overlap between *Elaboration*, *Joint,* and *Same-unit.* The results also suggest that *Elaboration* is used as a kind of default relation by the parser. This matches with the results of the individual annotation analyses, as shown in Fig. 8.

Fig. 7 and Fig. 8 illustrate that the gold standard has a greater variation of relations, especially in cases where the parser annotates with *Elaboration*. Both suggest that *Elaboration* is a default relation of the parser. The figures reflect the fact that in the gold standard *Elaboration*, *Joint*, and *Same-unit* also occur frequently, but *Elaboration* is not as frequent as in the parser output.

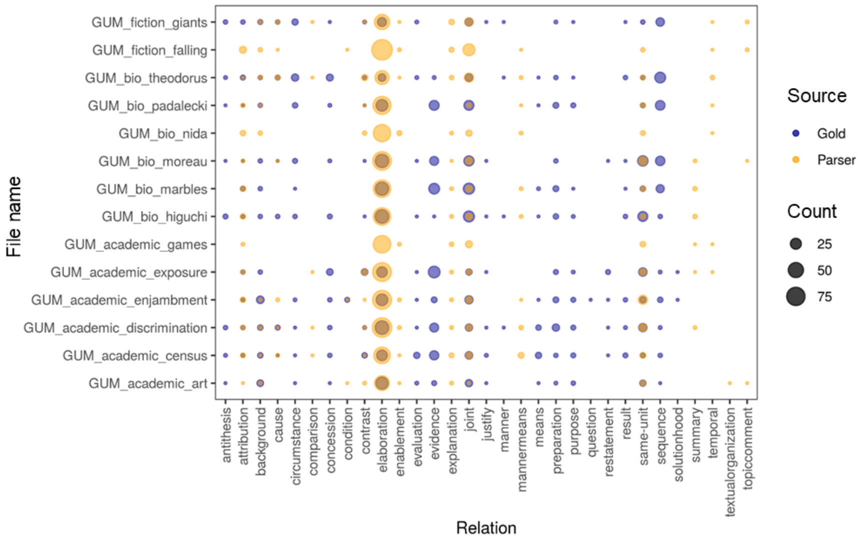

Fig. 8. Absolute frequency of relations in the texts, distribution in parser output (orange) and gold standard (blue); without gold-annotations for the files *falling, *nida, and *games. Some low frequent and thus negligible relations do not match in.

5 Analysis and Discussion

If the compared data can be assessed as valid despite its small size, then a rather poor match is apparent: the negative κ values could be explained by systematic differences between parser output and gold standard. The overall low F1-scores also indicate a low agreement between the two sets of annotations. For the IAA, the differences in nuclearity and relation are greater than for constituent and attachment point. With regard to MR, the major difference is in terms of nuclearity. The F1-scores of the test phase of the parser (Table 2) are far better. Thus, the state-of-the-art results on register-specific data could not be reproduced when applying the parser to cross-register data.

The confusion matrix and absolute frequency of the relations suggest that *Elaboration*, *Joint*, and *Same-unit* are the most common labels in the gold standard and parser output. Furthermore, there seems to be some overlap between *Elaboration*, *Joint*, and *Same-unit*. However, *Elaboration* occurs disproportionately often in the parser output. A considerable number of the *Elaboration* annotations in the parser output deviate from the gold standard, suggesting that *Elaboration* serves as a default relation for the parser.

The results are to be expected with regard to certain differences concerning the relations. Regarding the theoretical framework, we expected that the differences in the hierarchical classification of the text units and the labeling of their relationship could be greater than in the determination of the sentence constituents and their connection. As illustrated in Fig. 3, a certain degree of blurring and overlapping in the definition spectrum of the relations allows (legitimate) differences between EDU weighting and the labeling based on it. These decisions relate more to semantic aspects and the subjective interpretation of the illocution structure. This in turn could be a reason for the overlapping effect between *Elaboration*, *Joint*, and *Same-unit*. The determination of phrases and their combination is more strongly influenced by the formal surface structure and by syntactic restrictions and thus has less room for interpretation. This could account for the smaller deviations in constituent and attachment point.

Vagueness and ambiguity are hardly avoidable [Ste18]. Thus, the approach in RST to restrict itself to one tree structure interpretation per text becomes a recurring problem in case of structural and semantic ambiguities [Ste18]. This underlines the relevance for text-linguistic tasks of finding resolutions to deal with high-level ambiguity. This challenge can be addressed by under-specification or over-specification [Ste18] or by tolerating that a certain percentage of ambiguity should be considered as part of the gold standard or ground truth: legitimate disagreement included in a complex gold standard [DTS17].

Under-specification would require a reduction of the inventory of relations in favor of more vague relations, e.g., by merging existing relations or adding superordinate categories as relations. An analysis of all gold standard files showed that there seem to be register-specific relation patterns. This is consistent with the assumption that discourse structure can vary depending on the register [Ste18]. However, high frequencies of *Elaboration* and *Joint* across registers are also apparent in the gold standard. It can be assumed that a larger amount of data would show comparable results. In other words, it is probably representative that *Elaboration* and *Joint* occur in the parser output with such high frequency. According to the data, *Elaboration* is a potential relation for under-specification, as it is common across registers and has a certain intersection with other relations, which could motivate a possible grouping of relations.

For over-specification and legitimate disagreement, a multi-level structure could be implemented, in which several versions of annotations could be systematically recorded and examined [Ste18]. With regard to the evaluation of an RST parser, a manual analysis would be necessary to be able to make the following distinctions between annotation agreement and disagreement: (i) concordant and correct, (ii) not concordant and correct and (iii) not concordant and not correct.

In view of the effort needed for such an implementation and the focus on practicable, application-oriented solutions, pursuing under-specified discourse analysis is more appropriate in the context of semantic storytelling. Various proposals were made regarding the under-specification of discourse structure [Ste11] which we will now examine for their applicability to semantic storytelling.

6 Summary and Outlook

In this paper, an efficient RST parser was evaluated using data from the GUM corpus with the evaluation tool RST-Tace. The research question was whether the parser still achieves state-of-the-art results when processing texts across registers. If yes, the parser could be considered for processing discourse structure for semantic storytelling. Many processing and conversion steps were needed. In the workflow, there was a huge loss of data as RST-Tace attested segmentation errors in a disproportionately large share of the files. The results are, therefore, based on a small amount of data only. Due to the adopted black box approach, the erroneous conditions are not verified in this paper.

According to the results, the performance of the parser for cross-register texts does not come close to the performance shown for the trained and tested text base from a single register. Yet, the data situation might be too weak to adequately evaluate the performance of the parser. Despite the small data basis, two aspects are apparent in the results beyond the small agreement: first, the results suggest that there is a sort of default relation. Second, they suggest a certain overlap of several relations in the gold standard and parser output. The latter speaks for systematic ambiguities, which cannot be fully resolved in the context of RST. This indicates that high-level ambiguity resolution is needed for text-linguistic tasks: either by countering it through under-specification, over-specification and/or broader gold standards.

Due to the application-oriented approach of the QURATOR project and the sheer amount of resources required for over-specification approaches, it is advisable to test under-specification approaches with regard to their applicability for semantic storytelling in future work. In addition, the sources of error in the processing workflow should be revealed using white box testing in order to repeat the test with a larger data set.

Acknowledgements

The storytelling system presented in this paper has received funding from the German Federal Ministry of Education and Research (BMBF) through the project QURATOR (grant no. 03WKDA1A). We would like to kindly thank Anke Lüdeling for her advisory and commendatory support.

References

[CMOM02] Carlson, Lynn; Marcu, Daniel; Okurowski, Mary Ellen (2002): RST discourse treebank. [Philadelphia, Pa.]: Linguistic Data Consortium.

[DTS17] Debopam Das, Maite Taboada, and Manfred Stede (2017): The Good, the Bad, and the Disagreement: Complex ground truth in rhetorical structure analysis. In Workshop on Recent Advances in RST and Related Formalisms, Sant. de Compostela, Spain.

[HPDI10] Herneault, Hugo; Prendinger, Helmut; du Verle, David; Ishizuka, Mitsuru (2010): HILDA: A Discourse Parser Using Support Vector Machine Classification. In: D&D 1 (3), S. 1–33. DOI: 10.5087/dad.2010.003

[HS15] Heilman, Michael; Sagae, Kenji (2015): Fast Rhetorical Structure Theory Discourse Parsing. Available online at http://arxiv.org/pdf/1505.02425v1.

[IDT15] Iruskieta, Mikel; da Cunha, Iria; Taboada, Maite (2015): A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. In: Lang Resources & Evaluation 49 (2), pp. 263–309. DOI: 10.1007/s10579-014-9271-6.

[JCN15] Joty, Shafiq; Carenini, Giuseppe; Ng, Raymond T. (2015): CODRA: A Novel Discriminative Framework for Rhetorical Analysis 41:3.

[JE14] Ji, Yangfeng; Eisenstein, Jacob: Representation Learning for Text-level Discourse Parsing. Available at https://www.aclweb.org/anthology/P14-1002.pdf.

[KZ16] Krause, Thomas; Zeldes, Amir (2016): ANNIS3: A new architecture for generic corpus query and visualization. In: *Digital Scholarship Humanities* 31 (1), pp. 118–139. DOI: 10.1093/llc/fqu057.

[Lin20] Linscheid, Pia (2020): Parsing von Diskursrelationen nach der Rhetorical Structure Theory: Evaluation eines RST-Parsers mittels eines RST-Korpus und Evaluationstools (Unpublished Master Thesis). Humboldt-Universität zu Berlin, Berlin.

[MT88] Mann, William; Thompson, Sandra (1988): Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. In: Text 8 (3), pp. 243–281. Available at https://www.aclweb.org/anthology/W15-1843.pdf.

[MP09] Maziero, E.; Pardo, Thiago AS. 2009. Metodologia de avaliação automática de estruturas retóricas. In Proceedings of the III RST Meeting (7th Brazilian Symposium in Information and Human Language Technology), Brasil.

[Nid12] Nidhra, Srinivas (2012): Black Box and White Box Testing Techniques – A Literature Review. In: IJESA 2 (2), S. 29–50. DOI: 10.5121/ijesa.2012.2204.

[Neu15] Neumann, Arne (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), pp. 309–312.

[PWLA19] Prasad, Rashim; Webber, Bonnie; Lee, Alan; Josji, Aravind (2019): Penn discourse treebank version 3.0. [Philadelphia, Pa.]: Linguistic Data Consortium.

[RBHK+20] Rehm, Georg; Peter Bourgonje; Stefanie Hegele; Florian Kintzel; Julián Moreno Schneider; Malte Ostendorff et al. (2020): QURATOR: Innovative Technologies for Content and Data Curation. In: Adrian Paschke, Clemens Neudecker, Georg Rehm, Jamal Al Qundus and Lydia Pintscher (eds.). *Proceedings of QURATOR 2020 – The conference for intelligent content solutions*. Berlin, Germany.

[RMBS+18] Rehm, Georg; Moreno-Schneider, Julián; Bourgonje, Peter; Srivastava, Ankit; Fricke, Rolf; Thomsen, Jan et al. (2018): Different Types of Automated and Semi-automated Semantic Storytelling: Curation Technologies for Different Sectors. In: Georg Rehm and Thierry Declerck (eds.): *Language Technologies for the Challenges of the Digital Age*. 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings, Vol. 10713. Cham: Springer International Publishing; Springer (Lecture Notes in Computer Science, 10713), pp. 232–247.

[RZBO+20] Rehm, Georg; Zaczynska, Karolina; Bourgonje, Peter; Ostendorff; Malte; Moreno-Schneider, Julián; Berger, Maria; Rauenbusch, Jens; Schmidt, André; Wild, Mikka; Böttger, Joachim; Quantz, Joachim; Thomsen; Jan; Fricke, Rolf (2020): Semantic Storytelling: From Experiments and Prototypes to a Technical Solution. In Tommaso Caselli, Eduard Hovy, Martha Palmer, and Piek Vossen (eds.), *Computational Analysis of Storylines: Making Sense of Events*. Cambridge University Press, In print.

[RZMO+20] Rehm, Georg; Zaczynska; Karolina; Moreno Schneider, Julián; Ostendorff, Malte; Bourgonje, Peter; Berger, Maria; Rauenbusch, Jens; Schmidt, André; Wild; Mikka. Towards Discourse Parsing-inspired Semantic Storytelling. In Adrian Paschke, Clemens Neudecker, Georg Rehm, Jamal Al Qundus, and Lydia Pintscher (eds.), *Proceedings of QURATOR 2020 – The conference for intelligent content solutions*, Berlin, Germany, 2020. CEUR Workshop Proceedings, Volume 2535. 20/21 January 2020.

[RZMS19] Rehm, Georg; Zaczynska, Karolina; Moreno Schneider, Julian (2019): Semantic Storytelling: Towards Identifying Storylines in Large Amounts of Text Content. In: Alípio, Jorge; Campos, Ricardo; Jatowt, Adam; Bhatia, Sumit (eds.). *Proceedings of Text2Story – Second Workshop on Narrative Extraction from Texts*. Co-located with 41st European Conference on Information Retrieval. Köln. Available at https://dblp.org/db/conf/ecir/text2story2019.html.

[Ste11] Stede, Manfred (2011): Discourse Processing. In: Synthesis Lectures on Human Language Technologies 4 (3), pp. 1–165. DOI: 10.2200/S00354ED1V01Y201111HLT015

[Ste16] Stede, Manfred (2016): Handbuch Textannotation. Potsdamer Kommentarkorpus 2.0. Universität Potsdam, Potsdam.

[Ste18] Stede, Manfred (2018): Korpusgestützte Textanalyse. Grundzüge der Ebenen-orientierten Textlinguistik. Second edition. Tübingen: Narr Francke Attempto. Available online at https://elibrary.narr.digital/book/99.125005/9783823392040.

[WEK12] Weber, B.; Egg, M.; Kordoni, V. (2012): Discourse structure and language technology. In: Nat. Lang. Eng. 18 (4), pp. 437–490. DOI: 10.1017/S1351324911000337

[WKLS19] Wan, Shujun; Kutschbach, Tino; Lüdeling, Anke; Stede, Manfred (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. Minneapolis, MN: Association for Computational Linguistics, pp. 88–96. Available at https://www.aclweb.org/anthology/W19-2712.

[Zel17] Zeldes, Amir (2017): The GUM Corpus: Creating Multilayer Resources in the Classroom. In: *Language Resources and Evaluation* 51 (3), pp. 581–612. DOI: 10.1007/s10579-016-9343-x.