APA-RST: A Text Simplification Corpus with RST Annotations

Freya Hewett[1,2]
[1]AI & Society Lab
Humboldt Institute for Internet and Society

[2]Applied Computational Linguistics
University of Potsdam
firstname.lastname@hiig.de

Abstract

We present a corpus of parallel German-language simplified newspaper articles. The articles have been aligned at sentence level and annotated according to the Rhetorical Structure Theory (RST) framework. These RST annotated texts could shed light on structural aspects of text complexity and how simplifications work on a text level.

1 Introduction

The goal of text simplification is to reduce the complexity of a text whilst retaining the main information, in order to make a text easier to understand. In this paper, we present a corpus of German-language parallel simplified newspaper articles, at three different complexity levels. Each text has been annotated according to the Rhetorical Structure Theory (RST, Mann and Thompson, 1988) framework. RST posits that segments (or Elementary Discourse Units, EDUs) in a text are related to each other and that one component of the pair serves as a 'nucleus' and the other as a 'satellite', or in some cases both components are considered to have 'nucleus' status. An example of a relation that connects EDUs is 'evidence' where the nucleus is a claim and the satellite provides evidence for the claim. RST therefore provides information about the structure of texts; an area which has thus far not been the focus of much research on text simplification. Whilst much of previous work has focused on sentence-level simplification, text-level simplification is a promising area of research, as it represents a 'more real use-case scenario for a simplification model' (Alva-Manchego et al., 2020). For the German language, *Leichte Sprache* is a term often used in the context of simplified texts: *Leichte Sprache* texts are written according to clearly defined rules, however text-level aspects, including coherence, are often neglected in the guidelines, even though they are highly relevant when producing accessible texts (Bock, 2019; Maaß, 2020).

Aside from contributing to the general collection of RST annotated texts, this corpus could also be used to help carve out what text-level simplification actually constitutes, how simplified texts are structured and how this differs to their standard counterparts, and could also be used to answer questions surrounding the complexity of different types of text structures. Section 2 provides an overview on previous work that has considered questions at the intersection of discourse structure and text simplification, Section 3 provides details on the annotation process of the corpus, Section 4 outlines some statistical analysis on the corpus and Section 5 summarises the paper and provides inspiration for potential use-cases for the corpus. The corpus can be downloaded at https://github.com/fhewett/apa-rst.

2 Related work

Datasets which combine discourse structure and text simplification are relatively rare. LeiKo is a German-language corpus of newspaper articles simplified at various levels (including *Leichte Sprache*, Jablotschkin and Zinsmeister, 2022). A subset of 40 articles from the corpus has (manual) Penn Discourse Treebank (PDTB) annotations. Ko et al. (2023) expand their corpus of English-language texts annotated according to the Questions under Discussion (QUD) framework to include six Newsela articles and their counterparts at middle and elementary school level. In the context of text-level simplification, the task of sentence deletion has also been approached using various discourse structures. Zhang et al. (2022) also look at Newsela texts, and automatically annotate them with a 'news genre-specific functional discourse structure' and with sentence alignments. They train a model to predict when a sentence should be deleted and find that the discourse structure improves the accuracy. Zhong et al. (2020) also focus on the task of sentence deletion and analyse various discourse-based features (from RST and PDTB) for this purpose and find that the position of a sentence in an RST tree as well as some specific relations play a key role. The link between discourse structure and other aspects of text-level simplification have also been considered; Siddharthan (2003) proposes a rule-based system – using cue words, for example – to help preserve coherence when restructuring texts during the simplification process. Niklaus et al. (2021, 2016) split complex sentences by using automatically parsed syntax trees. They use a pre-determined list of cue words (such as 'although') to determine the rhetorical relation within sentences to ensure that the split sentences are still coherent. Davoodi and Kosseim (2016) implement pairwise classification of texts of varying complexity using discourse features, using a subset of 30 articles from the PDTB which have been annotated with a complexity level (Pitler and Nenkova, 2009) and an automatically parsed subset of the Simple English Wikipedia corpus (Coster and Kauchak, 2011).

3 Corpus creation

The data used in the corpus is from the Austria Press Agency (APA), who publish four to six articles every weekday, (manually) simplified to two language levels: B1 and A2[1], according to the Common European Framework of Reference for Languages (CEFR). More details on the APA data can be found in Ebling et al. (2022); the version used to create the APA-RST subset contains articles up to April 2022. APA-RST covers a total of five randomly selected days from a time-frame between 2018 and 2022, with five articles each day. The corpus therefore consists of a total of 75 articles, with 25 at each level (original, B1 and A2), covering different topics such as politics, culture and sport.

3.1 RST annotations

Each article has been annotated according to the RST guidelines from Stede et al. (2017). In addition to the relations[2] present in those guidelines, we also include two additional relations from RST-DT (Carlson and Marcu, 2001): *sameunit* and *attribution*. We also remove *means* from the relation set (due to its similarity to *enablement*) as well as *unless* (due to its similarity to *condition*). The titles of the newspaper articles were excluded from the annotation, as well as glossary entries for complicated words, which were occasionally included in the A2 texts. Longer texts were (manually) split into separate parts for a total of 111 parts; information on these parts can be found in Table 1. These texts were segmented into EDUs and given to the annotators in pre-segmented form. Five annotators used rstWeb to annotate the texts (Zeldes, 2016). The annotators were undergraduate students of computational linguistics, who were trained for the annotation task and had regular feedback sessions during the annotation process.

Level	**OR**	**OR parts**	**B1**	**A2**
Total sent.	558	558	184	204
Total tok.	9567	9567	2009	1871
Sent./text	22.3	9.1	7.4	8.2
Tok./text	382.7	156.8	81.0	74.2
Tok./sent.	17.1	17.1	10.9	9.2
Char./tok.	6.2	6.2	5.8	5.6
EDUs/text	-	15	9	9

Table 1: General information on APA-RST. Sent. stands for sentence(s), tok. for token(s), char. for characters and OR for original level. The values which are not totals represent averages.

Approximately one third of the corpus (36 texts) has three sets of annotations. The inter-annotator agreement (IAA) was calculated using RST-Tace (Wan et al., 2019), which is based on a proposal by Iruskieta et al. (2015), and considers four different aspects: nuclearity, relations, constituents and attachment points. RST-Tace is designed for comparing two sets of annotations, so to adapt it for our three sets we simply calculated the IAA for all possible combinations, i.e. between set 1 and 2, set 1 and 3, and set 2 and 3. Overall, the average Kappa score is 0.27, and the aspect with the most disagreement between annotations was the relations. Out of all non-matching relations, *elaboration* and *e-elaboration* are the main source of disagreement, i.e. one annotator chose *elaboration* whilst a second annotator chose *e-elaboration* for the same set of EDUs. Although a certain level of subjectivity is to be expected in RST annotations, due to the relatively low agreement, all annotations were manually checked by two doctoral students. Additionally, the texts with multiple annotations

[1] The A2 level corresponds (approximately) to *Leichte Sprache*.

[2] For readers unfamiliar with RST, short descriptions of the relations mentioned in the following Sections can be found in the Appendix A.4.

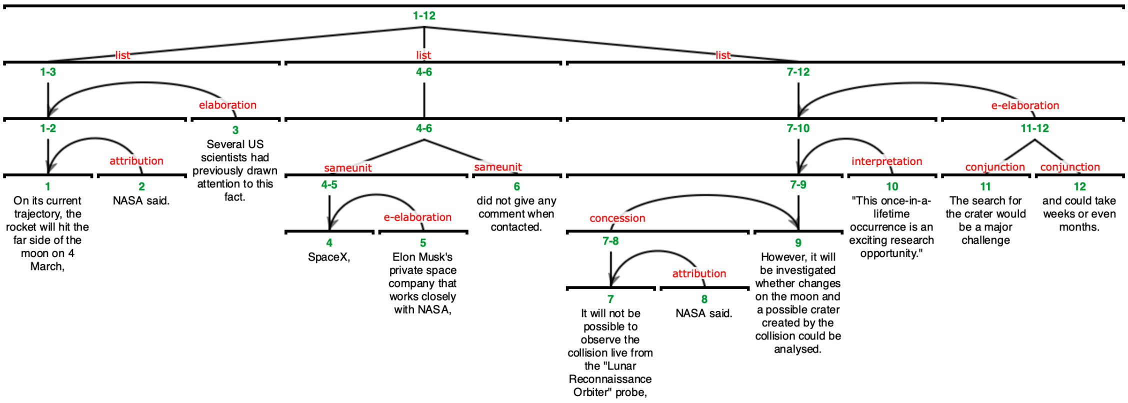
1-12
list
list
list
1-3
4-6
7-12
elaboration
e-elaboration
1-2
3
Several US scientists had previously drawn attention to this fact.
4-6
7-10
11-12
attribution
sameunit
sameunit
interpretation
conjunction
conjunction
1
On its current trajectory, the rocket will hit the far side of the moon on 4 March,
2
NASA said.
4-5
6
did not give any comment when contacted.
7-9
10
"This once-in-a-lifetime occurrence is an exciting research opportunity."
11
The search for the crater would be a major challenge
12
and could take weeks or even months.
e-elaboration
concession
4
SpaceX,
5
Elon Musk's private space company that works closely with NASA,
7-8
9
However, it will be investigated whether changes on the moon and a possible crater created by the collision could be analysed.
attribution
7
It will not be possible to observe the collision live from the "Lunar Reconnaissance Orbiter" probe,
8
NASA said.

Figure 1: The tree of the second part of an original level text (4-freitag-28-1-22-or-pt2).

have been manually harmonised and checked by one annotator from the original group and two doctoral students.

3.2 Alignment annotations

The original articles and the B1 articles have also been annotated with alignments to the simplified levels. Two annotators (an undergraduate student and a doctoral student) looked at the sentences in the B1 and A2 texts and labelled the sentence(s) in the original texts which provided the content for the simplified sentences. The inter-annotator agreement (Kappa) for aligning the sentences in the original texts to their B1 counterparts was 0.77, and 0.9 for B1 to A2. Information on the types of alignments can be found in Table 3. An example of an aligned text can be seen in Table 2; the B1 sentence [1] consists of content from two sentences in the original, this is an n:1 alignment.

4 Corpus statistics

As we can see in Table 1, the simplified texts are approximately a third of the length of the original articles, showing that simplification of newspaper articles mostly results in a shorter version. The A2 texts are slightly longer than the B1 texts, owing to more descriptions and explanations of complicated concepts.

Relation distribution. Figure 2 shows the distribution of the relations at the different complexity levels. The texts at levels B1/A2 contain more *elaborations* and *e-elaborations* than the standard texts. With regards to the multi-nuclear relations – relations which consist of two nuclei segments – the simplified texts contain more *sequences* and slightly more *conjunctions*, whereas the original texts contain almost the same amount of *lists*. For the original texts, the *list* relations are found at a higher level in the tree, as they encompass an average of 8.2 EDUs, as compared to approximately 4 in the simplified texts, as can be seen in Figure 3. The example tree in Figure 1 shows a *list* relation at the highest level in the tree. When it comes to causal relations, the simplified texts contain more *causes*, slightly more *results* but less *reasons*. According to the annotation guidelines, *reason* should be used to link two subjective claims, which suggests that the original texts have more subjectivity. *Attribution* relations occur more frequently in the standard texts. Attributing information or a quote to an external source increases the number of perspectives in a text, whereas the simplified texts have less attributions and therefore less perspectives. *Sameunits* do not occur at all in the simplified versions. In the original texts they are used for nested constructions; these are not present in the simplifications.

Relations and nuclearity of aligned sentences. Figure 4 shows the relations together with the nuclearity assignments for the original sentences which align with the B1 version, i.e. the original sentences which contain the content chosen for the simplification. Only those that occurred at least 5 times were included. Any bars above the line means that the corresponding relation occurs over-proportionally in the selected sentences. *Elaboration N* and *e-elaboration N* feature heavily. As the simplified texts are more concise, they mostly consist of more salient information. *E-elaboration S* are also selected more frequently, indicating that elaborations on specific entities or examples are useful for a simplified text. The high frequency of *sequence MN* also suggests that the simplified texts may have more of a linear tone; the standard deviation for the

B1 sentences	**Original sentences**
[1] In 2015, a rocket from the company SpaceX was sent into space.	[1] A part of **a SpaceX rocket** could collide with the moon in early March, according to calculations by scientists at the US space agency NASA. [3] The rocket was launched from the Cape Canaveral Cosmodrome **in 2015** and had brought the "Deep Space Climate Observatory", an Earth observation satellite, into space.
[2] The rocket's fuel ran out before it could return to earth, which is why it is still in orbit.	[4] Afterwards, however, **the rocket's fuel ran out before it could return to Earth, so it's been in orbit ever since**.
[3] According to the US space agency NASA, a part of the rocket could collide with the moon in early March.	[1] **A part of a SpaceX rocket could collide with the moon in early March, according to** calculations by scientists at **the US space agency NASA.**
[4] NASA announced this on Thursday.	[2] The trajectory of the "Falcon 9" rocket is currently being monitored, a **NASA** spokeswoman **told** the Deutsche Presse-Agentur **on Thursday.**
[5] It will not be possible to observe the collision live.	[8] **It will not be possible to observe the collision live** from the "Lunar Reconnaissance Orbiter" probe, NASA said.
[6] However, it will be investigated if there are any changes on the moon afterwards.	[9] **However, it will be investigated whether changes on the moon** and a possible crater created by the collision could be analysed.
[7] The search for the crater could however take weeks or months.	[11] **The search for the crater** would be a major challenge and **could take weeks or even months.**

Table 2: Example of an aligned text. The B1 sentences (on the left) were aligned with the original sentence (on the right) that contains the content; the relevant content is highlighted in bold. The full original text can be found in Appendix A.1.

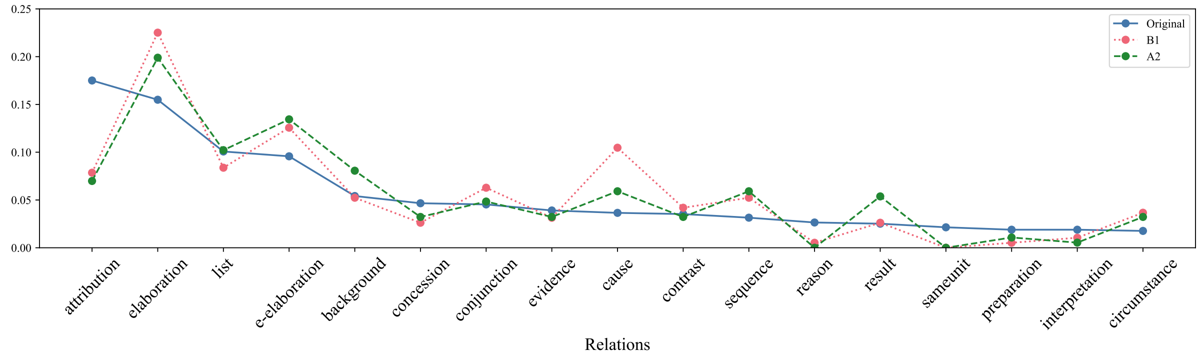

Figure 2: Relation distribution at the three different complexity levels. The counts of relations have been normalised. For readability purposes, only the top 17 relations are shown (out of a total of 30).

Level	**1:1**	n**:1**	**1:**n	**1:0**	**0:1**
OR:B1	85	5	33	430	30
B1:A2	123	1	33	26	10

Table 3: The types of alignments that were annotated, where n is more than one.

depth of EDUs[3] is in fact lower in the simplified texts (1.18 at A2 level and 1.33 at B1 level, compared to 1.4 in the original texts), indicating that the RST trees for the simplified texts are slightly more shallow. The mean depth of the trees of the original texts is 4.7 EDUs, compared to 4.1 and 3.9 for B1 and A2, respectively.

5 Conclusion and outlook

We have introduced a new German-language corpus of 75 parallel texts at three different complexity levels. The texts have been annotated according to the RST framework and have also been aligned at sentence level. We have shown how the relation distribution differs across the complexity levels, as well as how the relations differ in terms of what level they are used in the tree. We have also looked at the sentence alignments together with the RST annotations and shown the specific relations and nuclearity assignments of the content that is selected for a simplification. We pro-

[3] For example, the depth of the EDUs 5 and 6 in Figure 1 are 5 and 4, respectively.

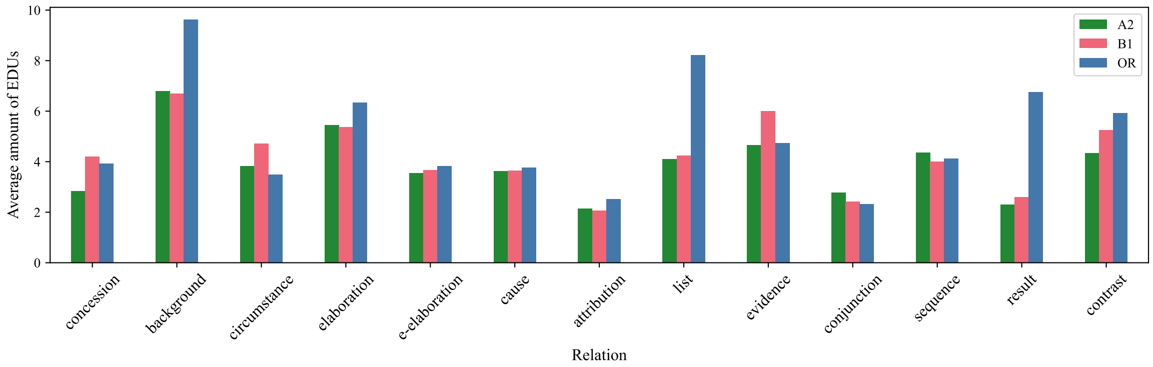

Figure 3: Average (mean) amount of EDUs that the relations encompass. Only relations that occur more than once in the A2 texts are included.

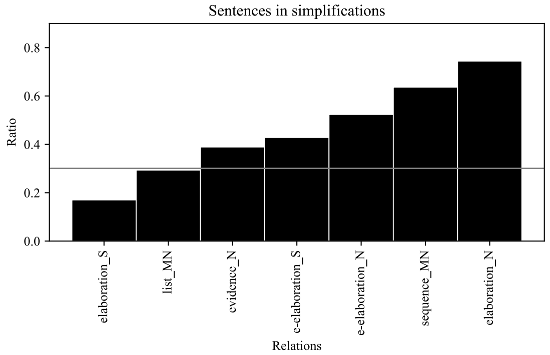

Figure 4: Relations and nuclearity of aligned sentences, out of the total amount of relations. Any relation ratio above 30% (the line on the graph) is above average. Relations and nuclearity assignments occurring less than a total of 5 times are excluded. N stands for nucleus, S for satellite, MN for nucleus in a multi-nuclear relation.

vide the corpus for download, enabling research on German-language RST in general, but also on specific questions which consider the interaction of text complexity and discourse structure.

Limitations

The corpus presented in this paper is relatively small and so the conclusions made should be considered in this context. We have also only focused on the specific text type of the newspaper article; other text types have different structures and are also simplified in different ways.

Acknowledgements

We would like to thank Dietmar Benndorf, Louis Cho, Elisa Lübbers, Hugo Meinhof, Costanza Rasi, Christopher Richter and Sara Shahmohammadi for their work on annotating the data. Furthermore, we would like to thank the reviewers for their very helpful comments. This research was funded by a grant from the German Ministry of Education and Research (BMBF).

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-Driven Sentence Simplification: Survey and Benchmark. *Computational Linguistics*, 46(1):135–187.

Bettina M. Bock. 2019. *'Leichte Sprache' – Kein Regelwerk. Sprachwissenschaftliche Ergebnisse und Praxisempfehlungen aus dem LeiSA-Projekt.* Frank & Timme.

L. Carlson and D. Marcu. 2001. *Discourse tagging reference manual (TR-2001-545).* USC Information Sciences Institute.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Elnaz Davoodi and Leila Kosseim. 2016. On the Contribution of Discourse Structure on Text Complexity Assessment. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 166–174, Los Angeles. Association for Computational Linguistics.

Sarah Ebling, Alessia Battisti, Marek Kostrzewa, Dominik Pfütze, Annette Rios, Andreas Säuberli, and Nicolas Spring. 2022. Automatic Text Simplification for German. *Frontiers in Communication*, 7.

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

Sarah Jablotschkin and Heike Zinsmeister. 2022. LeiKo. Ein Vergleichskorpus für Leichte Sprache und Einfache Sprache. In Mark Kupietz and Thomas Schmidt, editors, *Neue Entwicklungen in der Korpuslandschaft der Germanistik. Beiträge zur IDS-Methodenmesse 2022*. Narr, Tübingen.

Wei-Jen Ko, Yating Wu, Cutter Dalton, Dananjay Srinivas, Greg Durrett, and Junyi Jessy Li. 2023. Discourse Analysis via Questions and Answers: Parsing Dependency Structures of Questions Under Discussion. *arXiv:2210.05905v2*.

Christiane Maaß. 2020. *Easy Language – Plain Language – Easy Language Plus : Balancing Comprehensibility and Acceptability*. Frank & Timme.

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.

Christina Niklaus, Bernhard Bermeitinger, Siegfried Handschuh, and André Freitas. 2016. A Sentence Simplification System for Improving Relation Extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 170–174, Osaka, Japan. The COLING 2016 Organizing Committee.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2021. Context-Preserving Text Simplification. *arXiv:2105.11178 [cs]*.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.

Advaith Siddharthan. 2003. Preserving Discourse Structure when Simplifying Text. In *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003) at EACL 2003*, Budapest, Hungary. Association for Computational Linguistics.

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

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

Amir Zeldes. 2016. rstWeb - A Browser-based Annotation Interface for Rhetorical Structure Theory and Discourse Relations. In *Proceedings of NAACL-HLT 2016 System Demonstrations*, pages 1–5, San Diego, CA.

Bohan Zhang, Prafulla Kumar Choubey, and Ruihong Huang. 2022. Predicting sentence deletions for text simplification using a functional discourse structure. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 255–261, Dublin, Ireland. Association for Computational Linguistics.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse Level Factors for Sentence Deletion in Text Simplification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(5), pages 9709–9716.

A Appendix

A.1 Translated example text (4-freitag-28-1-22-or)

A part of a SpaceX rocket could collide with the moon in early March, according to calculations by scientists at the US space agency NASA. The trajectory of the "Falcon 9" rocket is currently being monitored, a NASA spokeswoman told the Deutsche Presse-Agentur on Thursday. The rocket was launched from the Cape Canaveral Cosmodrome in 2015 and had brought the "Deep Space Climate Observatory", an Earth observation satellite, into space. Afterwards, however, the rocket's fuel ran out before it could return to Earth, so it's been in orbit ever since. On its current trajectory, the rocket will hit the far side of the moon on 4 March, NASA said. Several US scientists had previously drawn attention to this fact. SpaceX, Elon Musk's private space company that works closely with NASA, did not give any comment when contacted. It will not be possible to observe the collision live from the "Lunar Reconnaissance Orbiter" probe, NASA said. However, it will be investigated whether changes on the moon and a possible crater created by the collision could be analysed. "This once-in-a-lifetime occurrence is an exciting research opportunity." The search for the crater would be a major challenge and could take weeks or even months.

A.2 Original example text (4-freitag-28-1-22-or)

Ein Teil einer SpaceX-Rakete könnte nach Berechnungen von Wissenschaftern der US-Raumfahrtbehörde NASA Anfang März mit dem Mond zusammenstoßen. Die Flugbahn der "Falcon 9"-Raketenstufe werde derzeit beobachtet, sagte eine NASA-Sprecherin am Donnerstag der Deutschen Presse-Agentur. Die Rakete war 2015 vom Weltraumbahnhof Cape Canaveral gestartet und hatte das "Deep Space Climate Observatory", einen Erdbeobachtungssatelliten, ins All gebracht.

Danach reichte jedoch der Treibstoff der Raketenstufe nicht aus, um zurück zur Erde zu kommen, weswegen sie seitdem im All unterwegs ist. Auf ihrer jetzigen Flugbahn werde die Raketenstufe am 4. März auf der Rückseite des Mondes einschlagen, hieß es von der NASA. Zuvor hatten mehrere US-Wissenschaftler darauf aufmerksam gemacht. Von SpaceX, der privaten Raumfahrtfirma von Elon Musk, die viel mit der NASA zusammenarbeitet, gab es auf Anfrage zunächst keine Reaktion. Der Aufprall werde von der Sonde "Lunar Reconnaissance Orbiter" nicht live beobachtet werden können, hieß es von der NASA. Es werde aber untersucht, ob danach Veränderungen auf dem Mond und ein möglicher durch den Aufprall entstandener Krater analysiert werden könnten. "Dieses einmalige Vorkommnis stellt eine aufregende Forschungsmöglichkeit dar." Die Suche nach dem Krater werde eine große Herausforderung und könne Wochen oder sogar Monate dauern.

A.3 Original example text, B1 (4-freitag-28-1-22-b1)

2015 ist eine Rakete der Firma SpaceX ins All gestartet. Der Treibstoff der Rakete reichte aber nicht mehr aus um zur Erde zurückzukehren, weshalb sie seither im All unterwegs ist. Laut Berechnung der US-Weltraumbehörde NASA könnte nun Anfang März ein Teil der Rakete in den Mond krachen. Das gab die NASA am Donnerstag bekannt. Der Aufprall wird nicht live beobachtet werden können. Allerdings wird untersucht werden, ob danach Veränderungen auf dem Mond erkennbar sind. Die Suche nach dem Krater könnte aber Wochen bis Monate dauern.

A.4 Descriptions of RST relations

Relation	Description
elaboration	‘S provides details or more information on the state of affairs described in N’
e-elaboration	‘S provides details or more information on a single entity mentioned in N’
sequence	‘the nuclei describe states of affairs that occur in a particular temporal order’
conjunction	‘the nuclei provide information that can be recognized as related, enumerating [...] and they are linked by coordinating conjunctions’
list	‘the nuclei provide information that can be recognized as related, enumerating’
cause	‘the state/event in N is being caused by the state/event in S’
result	‘the state/event in S is being caused by the state/event in N’
reason	S and N are ‘subjective statement[s]/thes[e]s/claim[s]’ and ‘understanding S makes it easier for [the reader] to accept N’
attribution	the attribution predicate is the S, the attributed material the N
sameunit	used for linking two discontinuous text fragments that are really a single EDU, but which are broken up by an embedded unit

Table 4: These descriptions are taken from the Annotation Guidelines from Stede et al. (2017); more detailed information can be found there. The descriptions for *sameunit* and *attribution* are adapted from the RST-DT guidelines (Carlson and Marcu, 2001).