Elaborative Simplification for German-language Texts

Freya Hewett[1,2], **Hadi Asghari**[1,3], **Manfred Stede**[2]
[1]AI & Society Lab, Humboldt Institute for Internet and Society, Berlin, Germany
firstname.lastname@hiig.de
[2]Applied Computational Linguistics, University of Potsdam, Germany
lastname@uni-potsdam.de
[3]Faculty of Electrical Engineering and Computer Science, TU Berlin, Germany

Abstract

There are many strategies used to simplify texts. In this paper, we focus specifically on the act of inserting information or *elaborative simplification*. Adding information is done for various reasons, such as providing definitions for concepts, making relations between concepts more explicit, and providing background information that is a prerequisite for the main content. As all of these reasons have the main goal of ensuring coherence, we first conduct a corpus analysis of simplified German-language texts that have been annotated with Rhetorical Structure Theory (RST). We focus specifically on how additional information is incorporated into the RST annotation for a text. We then transfer these insights to automatic simplification using Large Language Models (LLMs), as elaborative simplification is a nuanced task which LLMs still seem to struggle with.

1 Introduction

There are many strategies used to simplify texts. Sentences can be shortened, split or paraphrased, complex words replaced with synonyms, and information can be reordered, dropped or inserted (Amancio and Specia, 2014; Alva-Manchego et al., 2019). In this paper, we focus specifically on the act of inserting information.

Inserting information is done for various reasons: providing definitions for concepts, making relations between concepts more explicit, and providing background information that is a prerequisite for the main content. These all should contribute to decreasing complexity and therefore ideally ensuring coherence; the semantic or pragmatic relationships that link units in a discourse to other units (Das and Taboada, 2018). Readers need to recognise these relationships in order to make sense of the text, so a more coherent text should increase comprehension and also allow readers to recognise the communicative function of the text (cf. Nussbaumer, 1993).

In our study, we focus on German-language texts and aim to transfer insights from a detailed corpus analysis to automatic simplification models, to improve their ability in inserting information and therefore their overall ability at simplification. We use a corpus of parallel newspaper articles that have been annotated with Rhetorical Structure Theory (RST). RST annotations provide information about how segments in a text are related to each other within semantic or pragmatic relations such as *cause*, *background*, or *contrast* (Mann and Thompson, 1988). Our corpus analysis examines how inserted information in simplified texts can affect the coherence, and also what purpose the additional information has.

In order to utilise these discourse structure annotations for the task at hand, we first add a new layer of annotations by labelling the transformations that are applied to the original sentences to create the simplified sentences. One of these labels is 'Insert complementary information' which we focus on in more detail. We examine the role that this inserted information plays in the overall RST annotation.

We then transfer these insights to automatic simplification using Large Language Models (LLMs), exploring the use of different prompts.

In summary, our contributions are: we extend the APA-RST corpus (Hewett, 2023) to include transformation labels. We show results of an extensive corpus analysis, showing how new information is inserted in text simplification, and how this affects the coherence. We then explore models for document-level text simplification for German using the insights from our corpus analysis, with results comparable with the state of the art.

In Section 2 we present an overview on work that has looked at the insertion of new information in simplified text. In Section 3 we present our annotations of alignment labels and fine-grained inserted information categories, before presenting our RST analysis. Section 4 gives details on our

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 29–39
September 18–20, 2024.

models and experiments with them, and we reflect on our results and possible avenues for future work in Section 5. We publish our annotations and models at https://github.com/fhewett/GermanElabSimplification.

2 Related work

Srikanth and Li (2021) introduce the term 'elaborative simplification' to describe content addition in text simplification. The elaborative content added consists of 'definitions, explanations or clarifications to improve readability' with effective elaborations providing background information 'in a contextual manner'. They focus on this contextual aspect, annotating 1,300 instances of elaborative simplification in the Newsela corpus (Xu et al., 2015), categorising them according to the level of contextual specificity. They experiment with GPT-2, fine-tuning it on the simplest texts in Newsela and their annotated elaborations. Their best-performing model has the four sentences preceding an elaboration in a simplified text as input, and generates an elaboration as output, with the level of context specificity as determined by the gold annotation. Wu et al. (2023) use these annotated instances and add Questions Under Discussion (QUD), to show which questions elaborations answer. They find that the most common purpose of the elaboration is to explain a concept, followed by elaborations explicitly describing the cause of consequence of an event. They use GPT-3 for zero-shot elaboration generation, experimenting with including an automatically generated or human annotated QUD in the prompt or not. The results show that manually written QUDs produce the best elaborations. These studies build on ideas proposed by Alva-Manchego et al. (2020), who list explanation generation as an area of future work (albeit in the context of sentence-level simplification), stating that it involves elaborating 'on the concept in a natural way that keeps the text grammatical, is meaning preserving, and is simple'. Additionally, the well-established evaluation metric for automatic simplification, SARI, rewards 'addition operations' (Xu et al., 2016).

Another related area of text simplification is conceptual complexity, defined as accounting for 'the background knowledge necessary to understand mentioned concepts as well as the implicit connections that the reader has to access between the mentioned concepts in order to fully understand a text' (Hulpuș et al., 2019).

Our work is also related to the field of factuality (evaluation) of language model outputs: Devaraj et al. (2022) create a taxonomy of factual errors in automatic simplification, including 'Information Insertion' which is described as inserting 'irrelevant or erroneous content'. They differentiate between these insertion errors and useful insertions, such as 'defin[ing] jargon or provid[ing] explanatory content'. In the field of automatic summarisation, Maynez et al. (2020) differentiate between intrinsic and extrinsic hallucinations, where the latter refers to 'adding information not directly inferable from the input document'. They find that 'over 90% of extrinsic hallucinations were erroneous' i.e. are 'neither faithful nor factual'. Maynez et al. (2020) also find factual hallucinations to be 'acceptable if they lead to better summaries that are factual with respect to the document and the associated background knowledge'. This last point is particularly relevant to the task of simplification.

In various guidelines on *Leichte Sprache* (LS) – a highly simplified rule-based version of German – inserting factual information is allowed and also even desirable, in order to increase the level of comprehension on the one hand, and to allow readers to potentially learn new information on the other hand (Maaß, 2015). The guidelines state that translators of LS are allowed to provide explanations, additional remarks, and (concrete) examples, in order to make abstract concepts or difficult words more comprehensible. Maaß (2015, p.130) does however state that translators, after adding these definitions, explanations and examples, should make sure that the text still has an argumentative flow. Bredel (2016) state that additional explanations in texts in LS can hinder the flow of the text and potentially also cause problems on the text level. These aspects are the specific focus of the current study, i.e. what happens to the structure and coherence of the text overall when elaborative simplification is used.

Other corpus studies which focus on the transformation operations between non-simplified and simplified text often define an operation for inserting information. This category encompasses subcategories such as inserting *eliciting information*, *complementary external information*, *spurious information*, *pre-requisite information*, *concrete examples of abstract concepts or phenomenona* (Amancio and Specia, 2014; Alva-Manchego et al., 2019; Sun et al., 2021; Laban et al., 2023). This

category has also been used in German-language corpus studies: Stodden et al. (2023) manually align parallel texts with a category for additional information and Jablotschkin et al. (2024) find that phrases such as 'for example' or 'that means' feature heavily in simplified texts and are used for *explaining difficult words*, *making abstract concepts more concrete* and *connecting the sentences of a text explicitly*.

3 Corpus analysis

The main corpus we work with is the APA-RST. The corpus consists of German-language newspaper articles, which are classified as being at B1 and A2 level, according to the Common European Framework of Reference for Languages (CEFR), which is a scale from A1 (beginner language learner) to C2 (native speaker). There are 75 parallel articles in the corpus, with 25 at each level (original[1], B1 and A2), covering various topics such as politics, culture and sport. The articles have been annotated with RST and manually aligned at sentence level; further information can be found in the original publication (Hewett, 2023). Due to the relatively small sample size, we extend our analysis to label 200 instances of the 'APA' subcorpus of DEplain (Stodden et al., 2023) which features a larger number of newspaper articles from the same publisher as APA-RST. This subcorpus has been aligned at the sentence level, between the versions B1 and A2.

3.1 Adding transformation labels

Two annotators added transformation labels to the sentence alignments in the APA-RST, i.e. a label to describe how the original content was transformed for the simplification. We determined our labels by first selecting a subset of the most relevant labels from previous work (cf. Section 2). We then annotated a few texts and refined the definitions and added or removed labels. Our final label set consisted of **Paraphrase** (the content is the same, but the wording and/or structure are different), **Simple split** (original sentence has been split into two or more sentences, the structure and vocabulary are similar), **Complex split** (a split combined with a paraphrase), **Join** (content from two or more original sentences is combined in one simplified sentence), **Drop extra information** (sentences are fairly similar, but some content has been dropped for the simplification), **Insert complementary information** (the simplified version contains content that is not explicit in the original), **Implicit** (content is included implicitly in original), and **Identical** (sentences are identical). Often the majority of sentences could be described as being paraphrases, and so the label **Paraphrase** was only to be used when no other category was suitable. The inter-annotator agreement, calculated using Cohen's kappa, is .62 for the labels from original to B1 and .72 for B1 to A2, which compares to related work (.62 for five transformation categories in Laban et al. 2023).

[1]These articles do not have a language level but are assumed to be at C1/C2 level.

Label	OR⇒B1	B1⇒A2
Paraphrase	15%	46%
Simple split	1%	9%
Complex split	23%	13%
Join	4%	3%
Drop extra info	34%	13%
Insert complementary info	19%	9%
Implicit	2%	4%
Identical	2%	3%

Table 1: Distribution of transformation labels. Note that for OR⇒B1 78% of the sentences are dropped, for B1⇒A2 14% are dropped. The distribution of the labels amongst the remaining 22% and 86% are shown here.

The distribution of our labels can be seen in Table 1. For the rest of the study, we focus on the labels **Insert complementary information** and **Implicit**. Although these do not constitute the largest categories of transformations in a simplification, we choose to focus on them as choosing the right complementary information to insert requires high-level reasoning and is linked to the 'hallucinatory' nature of texts produced by LLMs.

3.2 Categories

We built a small typology of categories of inserted information, based on the transformation labels and their descriptions that were outlined in Section 2. Our categories and their descriptions can be seen in Table 2. We label all sentences that have the alignment label **Insert complementary information** or **Implicit**. In addition to this, we focus on the DEplain alignments labelled with **Additional**. We exclude any of the DEplain sentences which do not match with our alignment transformation guidelines, i.e. if a sentence is labelled as **Additional**, but would be labelled as a different category according to our guidelines, we exclude it. Note that

Name of category	Description	Example	%
Example	Provide an example to make a concept clearer.	For example, coloured pencils from the same company cost more in some shops than in others.	4.2%
Background	Provide information that is a prerequisite for understanding the rest of the text.	In the Spanish region of Catalonia, many people voted in favour of independence from Spain in 2017.	33.1%
Relation	Make a relation more clear/explicit between concepts.	The new virus variant emerged for the first time in South Africa. (Next sentence: All people who have returned from certain South African countries in the last few days should now take a PCR test).	32.2%
Definition	Provide a definition or summary of a concept.	Pub is the English word for a *Lokal*.	15.1%
Additional	Provide information that is new but is not necessarily required for understanding the main points.	Marcel Sabitzer won the vote last year.	15.5%

Table 2: The names, descriptions and distribution of our fine-grained labels for inserted information.

the APA texts often include glossaries in the simplifications, providing definitions on concepts and words. We do not include these in our analysis, as we focus on coherence within the main text.

The largest categories of inserted information are **Background** and **Relation**, which are both specific to the context of the text that is being simplified. **Examples** are the rarest kind of inserted information; we note however that this is not to say that examples are rare in the texts overall, it is often the case that the examples are present in the original texts and therefore do not constitute *inserted* examples. We note that additional information that seems to have no purpose other than providing more (non-prerequisite) knowledge also occurs (**Additional**), but that generally there is a balance between succinctness and level of simplification.

3.3 RST analysis

We look at the RST trees and the overall structure of the texts in APA-RST, and consider the individual properties of the inserted information, such as the position, the RST relation, the nuclearity status, and how this relates to the fine-grained category (i.e. the type of inserted information, as outlined in the previous section). Adding definitions and prerequisite information is done to contribute to making a text easier to understand, i.e. by making relations between concepts and facts more explicit and reducing the background knowledge needed to understand a text. However, adding this new information changes the structure and flow of texts, and also changes the way adjacent statements relate to one another. Analysing the RST annotations could help shed light on how the structure of texts change and how new information is used to 'facilitate connections between content in the original text' (Srikanth and Li, 2021).

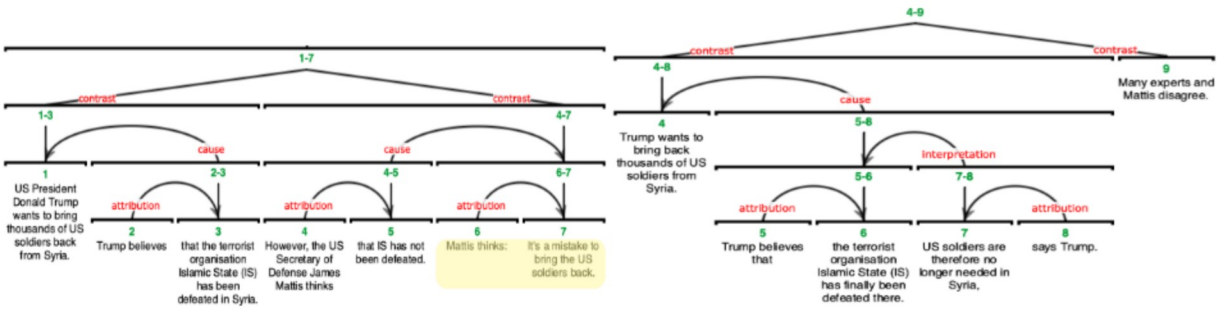
Relation. Overall we find that when the function of the inserted information is annotated as 'relation', i.e. making the link between two concepts more explicit, the inserted information is part of an RST relation broadly belonging to the causal category, such as *cause*, *motivation* or *evidence*. This can for example be seen in Figure 1a, where segments 6 and 7 are inserted information which have been annotated as 'relation'; they serve as the consequence of the causal relation, which is left more implicit in the original and the B1 text. This inserted information also makes the contrast relation, which connects a large amount of segments in the text, even more apparent, as it evens out the amount of sentences on each side of the contrast relation (2 vs. 2 in the A2 text, 3 vs. 1 in the B1 text).

Background. The inserted 'background' information is often at the beginning of the text; either directly at the beginning, as in:

> **After a week of lockdown in Austria, the government started discussing the Corona situation on Monday.** (*N elaboration*, 2-29-11-21-b1)[2]

Or after an initial sentence that has been paraphrased from the original article. In some cases this summarising background sentence at the beginning of the original articles is suitable to start a simplification with, and in other cases it is necessary to add

[2]The whole texts can be viewed here: https://github.com/fhewett/apa-rst. Sentences in **bold** represent inserted information.



(a) A2 (b) B1

Figure 1: Extracts of the RST annotations for the text 1-21-2-18. The new information is highlighted in yellow. The trees were created using rstWeb (Zeldes, 2016).

information before this first sentence. This background information is often elaborated upon in the article and therefore often has the relation *elaboration* or *background*. In the A2 versions, the content added in the B1 versions is expanded upon with more additional content, to make relations even clearer or to reduce the amount of presupposed background knowledge:

> **Because the hailstones were so large, they caused a lot of damage.** (*N evidence*, 3-21-2-18-b1)

This is expanded in the A2 text with two additional sentences preceding it:

> **When it hails, icy stones fall from the sky. Normally the hailstones are as small as peas.** (*S background*, *S concession*, 3-21-2-18-a2)

This indicates that when creating simplified texts at different levels, the same content that has been added for a more complex level can be expanded upon for a less complex level (as opposed to adding new content which covers a different topic than the previously added content).

Definitions. When definitions are added to the text directly (as opposed to glossary entries, which are displayed outside of the text), they are often used for conversions, or for translations:

> **That [23%] is almost a quarter more expensive than last year.** (*S elaboration*, 3-29-11-21-b1)

Inserting new information does create more "distance" between some entities:

> In New York, the city in the US, a painting has been sold at auction for around 45 million dollars. **That is around 40 million euros.** The picture originates from the Italian painter Sandro Botticelli. (*S e-elaboration*, 5-freitag-28-1-22-a2)

In this text, the information about the equivalent euros amount is added, and the third sentence then goes on to talk about the painting again (i.e. the entity introduced in the first sentence). It is not clear if this added distance makes comprehension more difficult. It seems that, at least in the articles published by the APA, longer definitions are not favoured in the main text, instead being given in a separate glossary. On the one hand, this ensures that the added definition does not cause too much distance between information on the same entity, on the other hand, it requires the reader to move between the main text and the additional glossary, interrupting a normal reading flow.

We note that there are no clear trends regarding the local (the importance of a segment within a segment-level relation) or global nuclearity (the importance of a segment within the overall tree) of the inserted information, indicating that it has many roles within a text.

Inserted **Examples** do not occur in the APA-RST, and as **Additional** inserted information may in fact be undesirable in a simplification (the information is unnecessary and increases the length of the text), we do not go into detail on this category.

3.4 Summary of corpus analysis

Our transformation labels show that the insertion of information does occur at both simplification levels, and whilst not as common as dropping information or splitting sentences, it still is frequent, particularly in simplification of original texts to B1.

Our fine-grained categories show that **Background** and **Relation** are the most common types

Prompt ID	Prompt text
Basic	Can you please summarise and simplify the following text to a B1/A2 level in German? Write a maximum of N sentences.
Background	Basic + add 1-2(B1)/2-3(A2) sentences at the beginning to give the user an overview of the topic. The text should have a clear introduction and information should be presented in a logical order.
Relation	Basic + add more contextual information to make the text easier to understand.

Table 3: The different prompts we use in our experiments. N is changed dynamically to reflect the amount of sentences in the reference simplification, and B1 or A2 used depending on the test set.

of inserted information, indicating that effective text simplification also involves conceptual simplification, i.e. decreasing the amount of background knowledge needed by the reader and therefore making relations more explicit. These transformations are more contextual than simply providing a definition.

Our RST analysis shows that background information is often at the beginning of a text, and often has the relation *elaboration* or *background*. Definitions that are added to the text could create 'distance' between related concepts, i.e. they add information that only attaches to one segment in the annotation, which may be why definitions only occur fairly rarely. In other texts, summarising sentences are used at the beginning or end of a sub-tree, so before the topic is changed slightly. When comparing simplifications from B1 to A2, the inserted content expands on the content that has already been inserted for the B1 text. Inserted content which makes a relation more clear often has a causal relation, so is making a cause or a consequence more explicit.

4 Automatic simplification models

We use `Meta-Llama-3-8B-Instruct` for our experiments as it is one of the most capable open-weight LLMs at the time of writing and performs well in benchmarks.[3] Additionally, LLMs that have been trained using strategies such as instruction-tuning and RLHF (as is the case for `Llama-3`) have been found to perform well in the task of automatic sentence simplification (Kew et al., 2023). We use `Meta-Llama-3-8B-Instruct` out-of-the-box, and also use this base model to fine-tune on B1 texts and A2 texts. We then explore using different prompts which are influenced by the findings from our corpus analysis.

4.1 Experimental setup

For fine-tuning, we use the same kind of texts found in the APA-RST, but in an extended version, and with no annotations.[4] The original texts are aligned with versions at B1 and A2. We use 2000 articles for training, and 296 for testing. We remove headlines and any glossary entries. We use the Basic prompt in Table 3 for fine-tuning; we include the word 'summarise' in the prompt as the simplified texts in our corpus are approximately a third of the length of the original texts. We separately fine-tune a model with A2 texts (FT-A2) and with B1 texts (FT-B1). As we use the 25 texts from APA-RST for our prompting experiments for inference, these texts are neither in the training nor the test set. Information on hyperparameters can be found in the Appendix A.2.

At inference, in addition to a basic prompt, we try out two other prompts (per model) which target the aspects **Background** and **Relation**.

We focused on these two categories as they were found to be most prominent in our corpus analysis. We leave experiments with the other categories for future work, but note that examples which are inserted in the simplification (i.e. the category **Example**) were rare in our corpus analysis and that additional information (i.e. the category **Additional**) could be difficult to evaluate and is potentially also not desirable even in a gold simplification, as it increases the complexity of a text and introduces potentially unnecessary information.

The prompts can be seen in Table 3. We use the texts from APA-RST as part of the prompts, for in-context learning. We used the following template for the **Background** and **Relation** prompts[5]:

[3] More information can be found on the model card on HuggingFace: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[4] A version of this dataset is also used by Rios et al. (2021) and Stodden et al. (2023).

[5] The exact format can be found in our repository: https://github.com/fhewett/GermanElabSimplification

Model	Prompt	Test set	SARI ↑	FRE ↑	M.P. ↑	S ↑	C ↑	F ↑	Avg. ↑
Baseline	Basic^{A2}	A2	41.2	59.4	**.89**	.38	**.96**	**.84**	**.77**
FT-A2	Basic^{A2}	A2	**44.0**	**70.6**	.49	**.82**	.56	.64	.63
Baseline	Basic^{B1}	B1	42.3	56.8	**.85**	.4	**.9**	**.9**	**.76**
FT-B1	Basic^{B1}	B1	**42.4**	**60.0**	.75	**.55**	.6	.75	.66

Table 4: Comparing `Llama-3` out-of-the-box and fine-tuned. The test set consists of 296 articles. The mean FRE score for the reference simplifications is 63.2 for the B1 texts, 69.1 for the A2. FT stands for fine-tuned. The right hand side shows the results of the manual evaluation, done on the outputs from each model for 10 texts. M.P. stands for meaning preservation, S for simplification, C for coherence, F for factuality; the score represents the percentage of 'yes' answers.

Model	Prompt	Test set	SARI ↑	FRE ↑	M.P. ↑	S ↑	C ↑	F ↑	Avg. ↑
FT-A2	Basic^{A2}	A2	44.0	70.6	.48	.8	.58	**.63**	.62
FT-A2	Background	A2	44.2	**70.8**	.51	.8	**.59**	.54	.61
FT-A2	Relation	A2	**44.5**	70.7	**.55**	**.95**	.57	.55	**.65**
FT-B1	Basic^{B1}	B1	42.4	60.0	**.75**	.55	.6	**.75**	**.66**
FT-B1	Background	B1	42.6	**64.7**	.47	**.79**	**.63**	.32	.55
FT-B1	Relation	B1	**43.0**	64.0	.58	.68	.47	.68	.61

Table 5: Results for prompting experiments. The test set consists of 296 articles. The mean FRE score for the reference simplifications is 63.2 for the B1 texts, 69.1 for the A2. FT stands for fine-tuned. The right hand side shows the results of the manual evaluation, done on the outputs from each model for 10 texts. M.P. stands for meaning preservation, S for simplification, C for coherence, F for factuality; the score represents the percentage of 'yes' answers.

system You are a helpful assistant and help the user to understand texts.
user {basic prompt} {original article}
assistant {text without inserted information}
user Thank you, that is good, but {additional prompt}
assistant {text with inserted information}
user Great, {additional prompt} {next original article}
assistant

To determine the wording for the prompts in Table 3, we first prompt `Llama-3`, asking it to tell us which of two texts are easier to understand and why; one text is an A2/B1 text from APA-RST, the other text is the same but with the inserted information removed (and small changes made to keep the text coherent). An example of this can be found in Appendix A.3. The overall approach was inspired by the concept of prompt chaining, where splitting up a task into subtasks potentially allows for more controllability (cf. Wu et al. 2022); as we want to encourage elaborative simplification, our approach involves making this explicit in the prompt by differentiating between simplification with and without elaboration.

4.2 Results

Model	Prompt	Test set	Elab
Baseline	Basic^{A2}	A2	.04
FT-A2	Basic^{A2}	A2	.64
FT-A2	Background	A2	.65
FT-A2	Relation	A2	**.67**

Table 6: Manual evaluation of the models' ability to perform elaborative simplification. The 'Elab' column refers to percentage of output texts in which the annotators said that elaborative simplification was performed.

We perform automatic evaluation using the metrics SARI (Xu et al., 2016) and Flesch Reading Ease (FRE). SARI is often used to evaluate simplification[6], and FRE is used to evaluate readability. FRE takes the average sentence length and average amount of syllables per word into account. The mean FRE score for the reference simplifications in our dataset is 63.2 for the B1 texts and 69.1 for the A2 texts. SARI is based on n-grams added, kept, and deleted between a reference simplification and a generated simplification. We report these metrics to be comparable with related work, but note that they have been criticised for being insuffi-

[6]We are aware of D-SARI (Sun et al., 2021), a version of SARI specifically adapted for document-level simplification evaluation; however we choose to use SARI to be comparable with other related work on document-level simplification (Stodden et al., 2023; Laban et al., 2023).

Model/Prompt	Simplification	**Baseline simplification** (FT-A2/B1, basic prompt)
FT-A2 Background	**Salzburg is a city in Austria. It is located in the Alps. Salzburg is famous for its festival.** [...] That's why Salzburg is the best city to visit in the coming year.	According to the travel book publisher "Lonely Planet", the city of Salzburg in Austria is the best city to visit in the coming year. [...]
FT-A2 Relation	**There are 14 mountains in Nepal that are very high. They are called eight-thousanders.** The highest mountain is Mount Everest. The Nepalese Nirmal Purja has climbed all 14 mountains. [...]	Nirmal Purja is a mountaineer from Nepal. He has set a record. He has climbed all 14 eight-thousanders in just 7 months. [...]
FT-B1 Background	**Ursula Stenzel is a politician from the FPÖ. She is a city councillor in Vienna.** [...]	Vienna City Councillor Ursula Stenzel (FPÖ) has not withdrawn after her appearance at a rally organised by the far-right *Identitären*. [...]
FT-B1 Relation	[...] **This is a problem because cars emit a lot of carbon dioxide. This is harmful for the environment.** The Austrian Transport Club (VCÖ) is **therefore** calling for more buses and trains. [...]	In Austria, car traffic has risen sharply since 2010. [...] The VCÖ is calling for a denser public transport network with more frequent train and bus connections.

Table 7: Examples of texts generated with different models and different prompts, compared to the basic prompt. The texts have been translated from German. The desired inserted information is in bold. We note that the FT-A2 **Relation** output contains a factual error, which is reflected in our manual evaluation.

cient measures of the quality of a simplification (cf. Alva-Manchego et al. 2021).

We extract 35 input texts and manually evaluate the outputs of our different models and prompts. We annotate the model outputs manually according to four criteria: meaning preservation, simplicity, coherence, factuality. Each criterion is a binary yes/no question. In addition to this, for a subset of 20 of these input texts we additionally annotate if the A2 models performed elaborative simplification. We only include the A2 models in this second evaluation as we use reference annotations to guide the evaluation and the majority of the instances in our corpus analysis were from A2 texts, due to the structure of DEplain. In total, three annotators evaluated 260 output texts. For 60 of these texts we have double annotations. The inter-annotator agreement for these texts across all criteria is .37 calculated using Cohen's kappa or .8 using the F1 score.

`Llama-3` out-of-the-box vs. fine-tuned. As can be seen in Table 4, our fine-tuned models only slightly outperform `Llama-3` out-of-the-box (referred to as baseline) for the B1 texts, but for A2 texts the improvement is more pronounced, particularly in terms of readability, as reflected by the FRE score. Our results are higher than (Rios et al., 2021), who report a highest SARI score of 32.9 using APA data, and compare to (Stodden et al., 2023), who report a highest SARI score of 44.6 when simplifying from B1 to A2 (not from standard to A2/B1, as we do in this study). We note that this improvement is rather due to the improvements that LLMs have made, rather than our method. The manual evaluation shows that the baseline model produces coherent, factual texts that cover the main points of the article, but are not necessarily written in a simpler way. As our main goal is simplification, we use our fine-tuned models for our prompting experiments.

Prompting experiments. As can be seen in Table 5, our prompts do result in slightly higher SARI and FRE scores. However, according to our manual evaluation, the prompts lead to a drop in factuality, meaning preservation and coherence. Overall, our prompts do lead to more simplification, and most importantly for this study, more elaborative simplification (cf. Table 6). Table 7 shows some examples where our prompts have had the intended effect, as compared to the basic prompt. The last example in Table 7 contains a factual error, which is a typical example of the nature of the factual errors we observed. The insertion of irrelevant or non-factual information is particularly problematic in the context of text simplification, where target users of a simplification will typically have difficulties comprehending the input text and may be less able to discern if the inserted information is factual or not (cf. Devaraj et al. 2022).

5 Conclusion and outlook

We have presented an in-depth analysis of elaborative simplification in German-language texts, using RST annotations and more fine-grained categories. We have experimented with using these insights to improve an LLM's ability to produce

elaborative simplifications. Our fine-tuned model and our different prompts do encourage the model to insert additional information (see Table 6), increase the level of simplification, and also result in marginal improvements on the SARI and FRE scores. However, the coherence and factuality seem to be adversely affected, indicating that these outputs contain repetitions or so-called hallucinations. This confirms results from related work, where conservative models may produce output texts that preserve the meaning of the input text, but fail to produce simplifications (cf. Cripwell et al. 2024).

As our analysis showed, not all simplified texts contain additional information and certainly not all types of additional information. In most cases, just one type is necessary, i.e. for texts of a political nature, more background knowledge and the relation between the entities in the text may be more important for understanding the text. Future work could investigate on selecting a prompt dependent on the input text.

Adding new information is not trivial; as can be seen in Figure 1, making relations more explicit, for example, can also slightly change the content of a text. In Figure 1b, segment 9 leaves some room for interpretation, as 'disagreeing' is not specific, whereas segments 4 to 7 in Figure 1a make this 'disagreement' very concrete. By keeping content more open and vague, it is easier to stay 'factual', showing that it is a fine line between making relations explicit and staying factual. Overall, elaborative or additive simplification remains a challenging sub-task of automatic simplification.

As shown by our manual evaluation, factuality and meaning preservation seem to represent separate requirements. We therefore advocate for factuality being included as a separate and additional evaluation criterion for text simplification, as up until now faithfulness and factuality seem to have been used interchangeably in the simplification literature, and simplifications are often (manually) evaluated for their meaning preservation (i.e. faithfulness). Our experiments have been limited to fine-tuning and prompting approaches, but experiments which alter the training/fine-tuning paradigm and loss function could also be promising, as at the moment 'most summarization [and simplification] systems are trained to maximize the log-likelihood of the reference summary at the word-level, which does not necessarily reward models for being faithful' (Maynez et al., 2020).

Acknowledgments

We would like to thank Birte Lübbert and Irina Kühnlein for their support with annotating our models' outputs. We are grateful to the anonymous reviewers for their helpful feedback. 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. 2019. Cross-sentence transformations in text simplification. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 181–184, Florence, Italy. Association for Computational Linguistics.

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

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4):861–889.

Marcelo Amancio and Lucia Specia. 2014. An Analysis of Crowdsourced Text Simplifications. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL 2014*, pages 123–130.

Ursula Bredel. 2016. *Leichte Sprache: theoretische Grundlagen, Orientierung für die Praxis*. Sprache im Blick. Dudenverlag, Berlin.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2024. Evaluating Document Simplification: On the Importance of Separately Assessing Simplicity and Meaning Preservation. In *3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)*, pages 1–14, Torino, Italy.

Debopam Das and Maite Taboada. 2018. Signalling of Coherence Relations in Discourse, Beyond Discourse Markers. *Discourse Processes*, 55(8):743–770.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland. Association for Computational Linguistics.

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.

Ioana Hulpuș, Sanja Štajner, and Heiner Stuckenschmidt. 2019. A spreading activation framework for tracking conceptual complexity of texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3878–3887, Florence, Italy. Association for Computational Linguistics.

Sarah Jablotschkin, Elke Teich, and Heike Zinsmeister. 2024. DE-Lite – a New Corpus of Easy German: Compilation, Exploration, Analysis. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 106–117, Malta. Association for Computational Linguistics.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

Philippe Laban, Jesse Vig, Wojciech Kryscinski, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. SWiPE: A dataset for document-level simplification of Wikipedia pages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10674–10695, Toronto, Canada. Association for Computational Linguistics.

Christiane Maaß. 2015. *Leichte Sprache: das Regelbuch*. Number 1 in Barrierefreie Kommunikation. Lit, Münster.

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.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Markus Nussbaumer. 1993. Textbegriff und Textanalyse. In Peter Eisenberg and Peter Klotz, editors, *Sprache gebrauchen – Sprachwissen erwerben*, pages 63–84. Klett, Stuttgart.

Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A New Dataset and Efficient Baselines for Document-level Text Simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161. Association for Computational Linguistics.

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative Simplification: Content Addition and Explanation Generation in Text Simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137, Online. Association for Computational Linguistics.

Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16441–16463, Toronto, Canada. Association for Computational Linguistics.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *CHI Conference on Human Factors in Computing Systems*, pages 1–22, New Orleans LA USA. ACM.

Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. 2023. Elaborative Simplification as Implicit Questions Under Discussion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5525–5537, Singapore. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 401–415.

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.

A Appendix

A.1 Limitations

In our study we have worked with one fairly homogeneous dataset. Different target groups and different genres will require different kinds of elaborative simplification. For example, Wu et al. (2023), find that definitions are the most common form of elaboration; the target group of their dataset is children.

As parts of our dataset are available online, we do not know if the data was part of the dataset used to pre-train `Llama-3`.

A.2 Hyperparameters

We use an NVIDIA V100S with 32GB VRAM for training and inference. Our hyperparameters can be found in Table 8. Note we also ran inference with a temperature of 0.4; the evaluation metrics were lower and so we only include the evaluation of models with this lower temperature.

temperature	0.0001
batch size per device	1
gradient accumulation steps	4
learning rate	3e-5
no. epochs	1
learning rate scheduler type	cosine
adam $\beta 1$	0.9
adam $\beta 2$	0.95

Table 8: Hyperparameters

A.3 Determining wording for prompts

To determine the wording for the **Background** and **Relation** prompts, we give the following input text and replace the {text with/out inserted information} with either a text from our corpus analysis that has inserted information from the category **Background** or **Relation**, respectively.

> **system** You are a helpful assistant and help the user to understand texts.
> **user** Can you tell me which text is simpler? Text 1: {text without inserted information} or Text 2: {text with inserted information}
> **assistant**

Example {text with background}, the first sentence in bold is removed for the {text without background}:

> **Energy has become much more expensive in the past year.** Many households are struggling to pay their energy bills. This is why the Austrian government has decided to introduce a so-called energy cost equalisation scheme. Almost all Austrian households will receive a one-off payment of 150 euros. Households in need will receive an additional 150 euros. This applies, for example, to the unemployed and people who receive benefits or a very low pension. In this way, the government wants to prevent households from falling into hardship in winter. (3-freitag-28-1-22-b1)