# DEplain: A German Parallel Corpus with Intralingual Translations into Plain Language for Sentence and Document Simplification

**Regina Stodden**, **Omar Momen** and **Laura Kallmeyer**
Heinrich Heine University
Düsseldorf, Germany
{firstname.secondname}@hhu.de

## Abstract

Text simplification is an intralingual translation task in which documents, or sentences of a complex source text are simplified for a target audience. The success of automatic text simplification systems is highly dependent on the quality of parallel data used for training and evaluation. To advance sentence simplification and document simplification in German, this paper presents DEplain, a new dataset of parallel, professionally written and manually aligned simplifications in plain German (*"plain DE" or in German: 'Einfache Sprache"*). DEplain consists of a news-domain (approx. 500 document pairs, approx. 13k sentence pairs) and a web-domain corpus (approx. 150 aligned documents, approx. 2k aligned sentence pairs). In addition, we are building a web harvester and experimenting with automatic alignment methods to facilitate the integration of non-aligned and to be-published parallel documents. Using this approach, we are dynamically increasing the web-domain corpus, so it is currently extended to approx. 750 document pairs and approx. 3.5k aligned sentence pairs. We show that using DEplain to train a transformer-based seq2seq text simplification model can achieve promising results. We make available the corpus, the adapted alignment methods for German, the web harvester and the trained models here: https://github.com/rstodden/DEPlain.

## 1 Introduction

Automatic text simplification (TS) is the process of automatically generating a simpler version of complex texts while preserving the main information (Alva-Manchego et al., 2020b). Current TS research mostly focuses on English and on sentence-level simplification.

This paper contributes to TS research on German. Compared to other European languages, German is more difficult to read due to complex sentence structures and many compound words (Marzari, 2010). According to Buddeberg and Grotlüschen (2020), roughly 6.2 mio. adults in Germany (approx. 12.1%) have reading and writing problems on the character-level (0.6%), word-level (approx. 3.4%) or sentence-level (approx. 8.1%). To counteract and make texts accessible to more people, currently two dominant German variants for simplified language exist (Maaß, 2020):

1. *easy-to-read German (de: "Leichte Sprache")*: following strict rules the complexity of the language is maximally reduced (almost corresponds to CEFR level A1). The main target group is people with cognitive or learning disabilities or communication impairments.

2. *plain German (de: "Einfache Sprache")*: reduced complexity with a mild to a strong extent (almost corresponds to CEFR levels A2 and B1), which can be compared to texts for non-experts. The main target group is people with reading problems and non-native German speakers.

This is also reflected in a rise in research and application of manual and automatic German text simplification: i) Many German web pages are provided in standard German as well as in plain or easy-to-read German, e.g., Apotheken Umschau[1] or the German Federal Agency for Food[2], ii) News agencies are publishing their news in plain or easy–to-read German, e.g., Austrian Press Agency[3] or Deutschlandfunk[4].

Klaper et al. (2013) were the first who made use of these resources for supervised, automatic German TS. They created a small parallel corpus of approx. 250 web pages with intralingual translations from standard German to easy-to-read Ger-

[1] https://www.apotheken-umschau.de/einfache-sprache/
[2] https://www.bzfe.de/einfache-sprache/
[3] https://science.apa.at/nachrichten-leicht-verstandlich/
[4] https://www.nachrichtenleicht.de/

*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 16441–16463
July 9-14, 2023 ©2023 Association for Computational Linguistics

man. However, due to copyright issues, they (and also its extension by Battisti et al. (2020)) could not make their corpus publicly available. To avoid such problems, Hewett and Stede (2021); Aumiller and Gertz (2022) built TS corpora based on open accessible Wikipedia texts simplified for children. Due to the high cost of manual sentence-wise alignment or not applicable automatic alignment methods (Aumiller and Gertz, 2022), these corpora are only aligned on the document level. Furthermore, Spring et al. (2022) report results on experiments with some existing automatic alignment methods and show non-satisfying, error-prone results.

In this work, we tackle some of the named problems by proposing, DEPLAIN, a new parallel German corpus for text simplification with manual and automatic alignments on the document and sentence level. DEPLAIN contains intralingual translations mostly into plain German and includes "strong" as well as "mild" simplifications.

Overall, we propose 4 subcorpora with in total 1,239 document pairs, 14,968 manual sentence-wise alignments and 1,594 automatic sentence-wise alignments. One subcorpus is built from professionally simplified news articles in plain language of the Austrian Press Agency[5]. The resources of the other 3 subcorpora are compiled by a new web harvester, making use of publicly available parallel documents. We analyze these subcorpora based on human ratings and annotations to get more insights into the quality and the simplification processes within the data. We further show two use cases of our new TS corpus: i) evaluating automatic alignment methods, and ii) exemplifying TS training and evaluation with DEPLAIN. Our data, web harvester, code for alignment methods and models are publicly available (with some restrictions).

## 2 Related Works

**Text Simplification (TS)** is an NLG task in which mostly machine learning models learn from complex-simple pairs how to simplify texts for a specific target group. For a lot of languages, parallel TS corpora exist either on sentence-level or document-level.[6] Only a few corpora contain data on both levels, e.g., EW-SEW v2.0 (Kauchak, 2013), Newsela 2015 (Xu et al., 2015), or Wiki-Auto (Jiang et al., 2020). Newsela (Xu et al., 2015), furthermore, includes for each source text several simplified versions targeted to different audiences. Therefore, it contains "strong" simplifications (highest to lowest complexity level) and also "mild" simplifications (intermediate complexity levels) (Štajner et al., 2017). In this paper, we also introduce one corpus with rather "mild" simplifications (DEPLAIN-APA) and one with rather "strong" simplifications (DEPLAIN-WEB) which allows more analysis of the capabilities of TS models. Like most other existing TS corpora (see Trienes and Vásquez-Rodríguez 2023), our corpus contains only one golden simplification (reference) per simplification pair.

**German corpora** were also proposed in recent years, both on either document level (e.g., Lexika-corpus Hewett and Stede 2021, or 20Minuten Rios et al. 2021) or sentence-level (e.g., web-corpus, APA-LHA, capito-Corpus Ebling et al. 2022, or Simple-German-Corpus Toborek et al. 2022), but not focussing on both levels[7]. Unfortunately, many of the datasets cannot be used for training TS models or only with caution because, for example, i) they are too small for training (e.g., Klaper et al. (2013)), ii) they are automatically aligned with questionable quality (e.g., Spring et al. 2021), iii) are only available for evaluation (e.g., Mallinson et al. 2020; Naderi et al. 2019), iv) they are not truly parallel as the complex and simple versions are written independently (e.g., Aumiller and Gertz 2022), or v) are not available (e.g., Ebling et al. 2022) sometimes due to copyright issues (e.g., Battisti et al. 2020). DEPLAIN tackles all of the mentioned problems, i.e., size, alignment quality, simplification quality, and availability.

**Alignment Methods** and web scraping are already used to overcome some of these issues. For example, similar to our work, Toborek et al. (2022) present a web scraper to scrape parallel documents from the web and automatically align them. However, (Spring et al., 2022) have shown that automatic sentence alignment is still an open challenge for German by comparing some existing alignment methods. In this work, we will evaluate the alignment methods on our manually aligned data and will adapt them for our purpose, e.g., use German resources and incorporate $n$:$m$ alignments.

[5] https://science.apa.at/nachrichten-leicht-verstandlich/

[6] For an extensive overview see Štajner (2021) or Trienes and Vásquez-Rodríguez (2023).

[7] For an extensive overview of existing German datasets including meta-data see Table 6.

| Name | License | # Doc. Pairs | # Original Sents | # Simple Sents. | Alignment | # Sent. Pairs |
|---|---|---|---|---|---|---|
| **DEplain-apa** | upon request | 483 | 25,607 | 26,471 | manual | 13,122 |
| **DEplain-web** | open | 147 | 6,138 | 6,402 | manual | 1,846 |
| | open | 249 | 7,087 | 7,760 | auto | 652 |
| | closed | 360 | 12,847 | 18,068 | auto | 942 |
| **In total** | mixed | 1,239 | 51,681 | 58,701 | mixed | 16,562 |

Table 1: Overview of the corpora of DEplain including meta data.

## 3 Document-level TS Corpora

We present two new TS corpora on the document level, DEplain-apa and DEplain-web, containing parallel documents in standard German and plain German. Table 1 provides statistics of both corpora.

### 3.1 DEplain-apa

The Austrian Press Agency (APA) publishes everyday news in standard German and parallel, professionally simplified versions for German language learners of CEFR level B1 and A2 (both equivalent to plain language): DEplain-apa contains news text of APA of CEFR level A2 and B1 which were published between May 2019 and April 2021. Data from the same source was already used for experiments with TS (see for an overview Ebling et al. (2022)) and made available as APA-LHA (Spring et al., 2021).

However, the APA-LHA alignments have some issues that are problematic for training a TS system: The alignment format is unclear in the sense of not distinguishing between 1:1, 1:$m$, and $n$:1 sentence alignments. Furthermore, the documents were aligned automatically, which results in many misaligned sentence-level alignments. Some examples of these problems are presented in Appendix B.

We tackle these problems by making use of the provided manual document alignments of APA from Common European Framework of Reference for Languages (CEFR) level B1 to A2. As the document alignments are not available for all APA documents, our corpus is reduced to 483 parallel documents. In a further comparison, DEplain-apa focuses more on mild simplifications which might be easier to learn for a document TS system than strong simplifications as in APA-LHA (C2 to B1 and C2 to A2).[8]

Overall, DEplain-apa contains 483 document pairs (see Appendix A and Table 6c). On average, the complex documents (CEFR-level B1) have a German Flesch-Reading-Ease score (FRE) (Flesch, 1948; Amstad, 1978) of 61.05 ± 4.67 and simple documents of 66.48 ± 4.56, which can be both interpreted as *simple* (following Amstad (1978, p. 117)).[9]

### 3.2 DEplain-web

The second document-level corpus of DEplain, i.e., DEplain-web, is a dynamic corpus with parallel documents in standard German and plain German from the web. Similar to Battisti et al. (2020) and Toborek et al. (2022), we have built an open-source web harvester in Python to download, align and extract text of parallel documents of given web pages (including paragraphs). For reproducibility, we made the code and the list of web pages available.

However, the automatic extraction of the web data is not perfect as some recent changes in the HTML structure are not recognized by the crawler, and some layouts such as tables or lists might not be extracted correctly. Following this, the data might include some low-quality data.

DEplain-web currently contains 756 parallel documents crawled from 11 web pages and covering 6 different domains: fictional texts (literature and fairy tales), bible texts, health-related texts, texts for language learners, texts for accessibility, and public authority texts. The first three domains are not included in any other German TS corpus.

All simplified documents are professionally simplified by trained translators and often reviewed by the target group. The simplified documents of 5 of the 11 web pages are written in plain German, 6 in easy-to-read German. All complex documents are in standard German, except *Alumniportal Deutschland*, which contains data on CEFR level B2. Some of the fictional complex documents are only available in their original language from the 19th century and are therefore more difficult to

[8] Examples for strong and mild simplifications of DEplain can be found in Appendix C.

[9] We calculated the German variant of FRE with the Python package textstat. For criticism on traditional readability scores for TS see, e.g., Tanprasert and Kauchak (2021).

read. More details on the scraped web pages are given in Appendix E. We plan further extensions of DEPLAIN-WEB, e.g., by a political lexicon in plain German[10].

The corpus is dynamic for three reasons: i) it can be extended with new web pages, ii) the number of parallel documents of a web page can change, and iii) the content of the considered web pages can change over time. More details on the web crawler, reasoning for choosing the current web pages, and the document alignment process can be found in Appendix E.

On the one hand, some of these web documents are openly licensed and some data providers allowed us to use and share the data for academic purposes. Therefore, we can publicly share this data; this corpus contains 396 document pairs which are represented in the second and third rows in Table 1. On the other hand, we additionally provide the web crawler to download and use the parallel documents with restricted licenses (360 documents) which is represented in the last row in Table 1.

## 4 Sentence-level TS Corpora

We aligned both corpora, DEPLAIN-APA and DEPLAIN-WEB, also on the sentence level. All 483 available parallel documents of DEPLAIN-APA and 147 documents of DEPLAIN-WEB are manually aligned on the sentence level with the assistance of a TS annotation tool. Overall, 14,968 sentence pairs of 630 document pairs are manually aligned. We first describe the annotation procedure (see subsection 4.1) and the resulting statistics per subcorpus (subsection 4.2 and subsection 4.3).

### 4.1 Annotation Procedure

DEPLAIN-APA and DEPLAIN-WEB are both annotated following the same procedure. The sentence pairs are manually aligned by two German native speakers[11] using the TS annotation tool TS-anno (Stodden and Kallmeyer, 2022) which assist, for example, in splitting the documents into sentences, alignment of $n$:$m$ sentence pairs, automatic alignment of identical sentence pairs, and the annotation of simplification operations and manual evaluation. The annotators were instructed by the principal investigator and were also provided with instructions on how to use the annotation tool and with an annotation guideline.[12]

**Sentence-wise Alignment** The manual sentence-wise alignments reflect all possible alignment types: i) 1:1 (rephrase and copy), ii) 1:$m$ (split of a complex sentence), iii) $n$:1 (merge of complex sentences), iv) $n$:$m$ (where $n$ and $m > 1$, fusion of complex and simple sentences). Furthermore, all not annotated sentences of an annotated document are either treated as v) 1:0 (deletion of a complex sentence), and vi) 0:1 (addition of a simplified sentence).

In the alignments of DEPLAIN-APA and DEPLAIN-WEB, the complex documents are fully aligned with the simplified documents. This means the alignments also reflect deletions and additions.

The publication of the full document alignments, also enhance the option for example, i) to build a simplification plan for document-level simplification using sequence labeling (see Cripwell et al. 2023), ii) to include preceding and following sentences for context-aware sentences simplification (see Sun et al. 2020), or iii) to use identical pairs and additions as augmented data during training (see Palmero Aprosio et al. 2019).

**Agreement of Alignment and Data Cleaning** To compare the agreement of both annotators, we randomly sampled 99 documents over all domains which were annotated by both annotators. For calculating the inter-annotator-agreement we framed the alignment as a classification task in which a label (not aligned, partially aligned, or aligned) is assigned to each combination of complex sentences and all simple sentences per document. This format was proposed by Jiang et al. (2020) for training and evaluating a sentence-wise alignment algorithm. The inter-annotator-agreement (measured with Cohen's $\kappa$) for these documents is equal to approx. 0.85 (n=87645 sentence combinations) which corresponds to a strong level of agreement (following McHugh (2012, p. 279)). The lowest agreement is shown for the domain of health data ($\kappa$=0.52, n=13736, interpretation: weak) whereas the highest agreement is shown for the language learner data ($\kappa$=0.91, n=18493, interpretation: almost perfect).[13] The health data was strongly and independently written in plain language (not sen-

[10] https://www.bpb.de/kurz-knapp/lexika/lexikon-in-einfacher-sprache/

[11] Both annotators were paid for their work with at least the minimum wage of their country of residence.

[12] We adapted the annotation schema of Stodden and Kallmeyer (2022) to our needs.

[13] In Appendix D, we show all inter-annotator agreements per domain.

tence by sentence), including moving sentences from document beginning to ending or sentence fusion. Therefore, the manual alignment of strong simplifications seems to be less congruent than for the very mild simplifications of the language learner data which have a low edit distance.

As the texts of DEplain-web are automatically extracted from the websites, and the documents of both, DEplain-apa and DEplain-web, were automatically split into sentences, the sentence pairs can contain some sentences that are wrongly split. Therefore, we cleaned the dataset and removed too short sentences (e.g., "Anti-Semitismus.", engl.: "Antisemitism.") and too similar sentences with only one character changed (e.g., complex: "Das ist schön!", simple: "Das ist schön.", engl.: "That is nice."). Furthermore, some sentence pairs (especially term explanations in the news dataset, n=1398) are repeated several times in different documents, we decide to remove all duplicates to make sure that only unseen sentence pairs occur in the test data set.

**Linguistic Annotation** After cleaning the data, similar to (Cardon et al., 2022), some randomly selected sentence pairs are annotated with additional linguistic annotations to get more insights into the simplification process of the aligned sentence pairs. We follow the annotation guideline provided in Stodden and Kallmeyer (2022)[14]. We built a typology on linguistic-based operations, which are performed during the simplification process, following a literature review of existing typologies Bott and Saggion (2014); Brunato et al. (2015); Gonzalez-Dios et al. (2018); Koptient et al. (2019). Our typology includes 8 operations, i.e., i) delete, ii) insert, iii) merge, iv) reorder, v) split, vi) lexical substitution, vii) verbal changes, and viii) no changes of which each can be annotated on the paragraph-level, sentence-level, clause-level, or word-level. Furthermore, we also manually evaluated the sentence-wise pairs on a few aspects. As no standards for manual evaluation exist (Alva-Manchego et al., 2020b), we decided to evaluate on the following three most often used criteria, i) grammaticality$^\dagger$, ii) meaning preservation$^\ddagger$, and iii) overall simplicity$^\ddagger$, and the following additional aspects: iv) coherence$^\dagger$, v) lexical simplicity$^\ddagger$, vi) structural simplicity$^\ddagger$ (similar to Sulem et al. (2018b)), and vii) readability (or simplicity)$^\dagger$ (similar to Brunato et al. (2018)). All aspects marked with $^\ddagger$ are rated on the sentence pair whereas all aspects with $^\dagger$ are rated on the complex as well as the simplified part of the sentence pair. These aspects are rated on a 5-point Likert-scale, to be more clear in the meaning of the scale, the scale either range from -2 to +2 or 1 to 5 following Stodden (2021). Following Alva-Manchego et al. (2020a); Maddela et al. (2021), we provide a statement per aspect on which the annotators are asked to agree or disagree on. An overview of the statement per aspect is added to Appendix G.

### 4.2 DEplain-apa

**Alignment Statistics** For the sentence-level part of DEplain-apa, all 483 parallel documents are manually aligned following the annotation procedure described above. Overall the subcorpus contains 13,122 manually aligned sentence pairs with 14,071 complex aligned sentences (55.82% of all complex sentences), and 16,505 simple aligned sentences (63.38% of all simple sentences) (see Table 1, and Table 6c).

The largest part of the aligned sentence pairs are rephrasings, 75.54% of the pairs are 1:1 aligned (excluding identical pairs). 17.99% of the complex sentences are split into several simpler sentences (1:$m$ alignments) and 2.91% were merged into one simple sentence. The remaining 3.57% are a fusion of several complex and several simple sentences (see Appendix F). Overall, the average sentence length has increased during simplification (complex: 12.64, simple: 13.02) which might be due to splitting long compound words into several tokens.

**Manual Evaluation.** For manual evaluation of DEplain-apa, 46 randomly sampled sentences were rated. The ratings (see Table 2) confirm that the corpus contains rather mild simplifications: the original sentences are already simple (4.39±0.77, max=5) and they are only simplified a bit (0.57±0.86). Furthermore, the original and the simplified sentences are very grammatical (complex:1.96±0.29, simple: 2.0±0.0), rather coherent (complex:3.26±1.6, simple: 3.54±1.54), and preserve the meaning (4.33±0.97).

**Simplification Operations.** 184 sentence pairs were annotated with their transformations, for some sentence pairs more than one transformation was performed at the same time. 47.83% of the pairs are changed on the sentence level and in 84.24% a

[14] The full annotation guideline can be found here: https://github.com/rstodden/TS_annotation_tool/tree/master/annotation_schema

| | | Simplicity | LexSimp | StructSimp | MeaningP. | Coherence | | Grammaticality | | Simplicity | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | sent. pair | sent. pair | sent. pair | sent. pair | complex | simple | complex | simple | complex | simple |
| corpus | n | (-2 to +2) | (-2 to +2) | (-2 to +2) | (1 to 5) | (1 to 5) | (1 to 5) | (-2 to +2) | (-2 to +2) | (1 to 5) | (1 to 5) |
| APA | 46 | 0.57±0.86 | 0.28±0.54 | 0.5±0.81 | 4.33±0.97 | 3.26±1.6 | 3.54±1.54 | 1.96±0.29 | 2.0±0.0 | 4.39±0.77 | 4.72±0.46 |
| WEB | 384 | 1.04±0.82 | 0.67±0.75 | 0.95±0.87 | 4.29±0.93 | 2.82±1.48 | 3.08±1.4 | 1.72±0.79 | 1.96±0.26 | 3.48±1.18 | 4.46±0.69 |
| news | 46 | 0.57±0.86 | 0.28±0.54 | 0.5±0.81 | 4.33±0.97 | 3.26±1.6 | 3.54±1.54 | 1.96±0.29 | 2.0±0.0 | 4.39±0.77 | 4.72±0.46 |
| bible | 155 | 1.39±0.68 | 0.98±0.78 | 1.28±0.77 | 4.34±0.84 | 2.12±1.22 | 2.63±1.22 | 1.45±1.06 | 1.92±0.35 | 2.97±1.27 | 4.44±0.72 |
| lang. | 157 | 0.67±0.74 | 0.36±0.57 | 0.57±0.73 | 4.46±0.73 | 3.83±1.27 | 3.82±1.27 | 1.96±0.22 | 1.97±0.21 | 4.01±0.81 | 4.43±0.71 |
| fiction | 72 | 1.1±0.95 | 0.69±0.78 | 1.08±1.02 | 3.82±1.29 | 2.08±1.06 | 2.42±1.33 | 1.75±0.71 | 2.0±0.0 | 3.42±1.16 | 4.56±0.58 |

Table 2: Results (mean and standard deviation) of the manual evaluation of the manually aligned sentence pairs per subcorpus (upper) and domain (lower). The left part contains results of aspects on the sentence pair (simplicity, lexical simplicity (LexSimp), structural simplicity (StructSimp), and meaning preservation (MeaningP.)) and the right part for the original and simplified sentences (coherence, grammaticality, and simplicity).

change was performed on the word level. On the sentence level, most often a sentence was reordered (48.86%), split (35.23%), or rephrased (12.5%). On the word level, most often a lexical substitution was performed (84.24%), a word added (46.45%) or a word deleted (35.48%).

**Interpretation & Summary** This analysis shows that the simplifications of DEPLAIN-APA are of a high quality (grammaticality, meaning preservation, coherence) and that they contain a lot of different simplification strategies (e.g., reordering, splitting, lexical substitution). So even if they are labeled as "mild" simplifications due to their close language levels (B1 to A2), they seem to be very valuable for training a TS corpus.

### 4.3 DEPLAIN-WEB

For the sentence-level part of DEPLAIN-WEB, 147 of the 456 parallel documents are manually sentence-wise aligned. The manual alignment process resulted in 1,846 sentence pairs (see Table 1 and Appendix A).

**Alignment Statistics.** In contrast to DEPLAIN-APA, both the complex sentences (avg$_{web}$=22.59, avg$_{APA}$=12.64) and the simplified sentences (avg$_{web}$=19.76, avg$_{APA}$=13.02) are longer on average, which is due to the different complexity levels (in web, complex is comparable to CEFR level C2, and A2-B2 for simple documents). Following that, the sentence pairs of DEPLAIN-WEB are more often split (43.12%) than DEPLAIN-APA (17.99%). However, still the most often alignment type is the 1:1 alignment (46.86%). Only 4.06 % of the complex sentences are merged and 5.96% are fused.For more statistics on DEPLAIN-WEB see Table 1, Table 6c and Appendix F.

**Manual Evaluation.** 384 randomly sampled sentence pairs are rated regarding simplification aspects (see Table 2). Overall during the simplification process, the sentences were improved in coherence (complex: 2.82±1.48, simple: 3.08±1.4), grammaticality (complex: 1.72±0.79, simple: 1.96±0.26) and simplicity (complex: 3.48±1.18, simple: 4.46±0.69). As presumed before, the original sentences of the bible (2.97±1.27) and the fictional literature (3.42±1.16) are more complex than the other original texts (even if not reflected in the FRE)[15]. However, their simplicity scores of the simple sentence (bible: 4.44±0.72, fiction: 4.56±0.58) are comparable to the scores of the other domains, therefore, these alignments can be seen as "strong" simplifications. Furthermore, they also have a higher average for structural and lexical simplifications than the other domains. Overall, DEPLAIN-WEB is comparable to DEPLAIN-APA in terms of meaning preservation (WEB: 4.29±0.93, APA: 4.33±0.97) and grammaticality (WEB simple: 1.96±0.26, APA simple: 2.0±0.0), contains stronger simplifications (WEB: 1.04±0.82, APA: 0.57±0.86) but the simplified web texts are less coherent than the simplified news (WEB simple: 3.08±1.4, APA simple: 3.54±1.54).

**Simplification Operations.** 350 pairs were rated, 50.57% on the sentence level and 69.43% on the word level. The manual annotation corresponds to the automatic calculation of alignment types: the most often change on a sentence level is the split of the sentence (50.57%). 30.51% of the pairs are rephrased and 14.69% are reordered. Interestingly also a high percentage of verbal changes (7.91%) which include changes from passive to active or subjunctive to indicative.

[15] This might due to the fact that FRE is build for text level and not sentence level. The calculation seems to fail for some sentences, e.g., "Anti-Semitismus." (engl: "Antisemitism.") got a score of -172 (extremely difficult to understand) whereas "Tom!" is scored with 120.5 (extremely easy to understand). Therefore, these scores must be interpreted with caution.

On the word level, again lexical substitution is performed most often (86.42%), in 24.28% at least one word is deleted, and in 11.52 % one word is added.

**Interpretation & Summary** This analysis again shows a mix of different simplification strategies, including lexical changes as well as syntactical changes. The manual ratings also lead to the assumption that the simplifications are strong and of high quality. Therefore, this corpus can also be a great benefit for German TS.

Furthermore, the manually aligned sentence pairs and the document pairs of DEPLAIN-WEB can be used for evaluating alignment algorithms across different domains. The alignment algorithm can then be used to automatically align the not-aligned documents of DEPLAIN-WEB. We are showing this process in the next section.

## 5 Automatic Sentence-wise Alignment

To exemplify the usage of the manual alignments and to provide sentence-wise alignments for the unaligned documents of DEPLAIN-WEB we evaluate different alignment algorithms on the manually aligned data.

### 5.1 Alignment Methods

We evaluated the following alignment methods: i) *LHA* (Nikolov and Hahnloser, 2019), ii) *SentenceTransformer* (Reimers and Gurevych, 2020) with *LaBSE*[16] (Feng et al., 2022) and *RoBERTa*[17] (Conneau et al., 2020) iii) *VecAlign* (Thompson and Koehn, 2020), iv) *BertAlign* (Liu and Zhu, 2022), v) *MASSAlign* (Paetzold et al., 2017), and vi) *CATS* (Štajner et al., 2018). Before testing any of these alignment methods, we investigated the implementation of their algorithms and checked for any room for adaptation to benefit our purpose.[18]

### 5.2 Evaluation of Alignment Methods

We chose the subcorpus of DEPLAIN-WEB that has manual alignments and is open for sharing (second row in Table 1) for evaluating the methods, as it has a sufficient number of alignments representing different domains and different types of alignments (1:1 and $n$:$m$). The dataset comprises 147 aligned pairs of documents, these complex-simple document pairs were split into 6,138 and 6,402 sentences respectively. The manual alignment of these sentences resulted in 2,741 alignments, comprising 1,750 1:1 alignments (out of which are 887 identical pairs), 804 1:$m$ alignments, 77 $n$:1 alignments, and 110 $n$:$m$ alignments.[19]

For evaluation, we treat the alignment task as a binary classification problem (either aligned or not aligned) and report precision, recall, and F1-score. We do not consider partial alignments within the evaluation. We argue that for curating a finetuning dataset for automatic text simplification systems, having an accurate alignment is more important than missing an accurate one, therefore we value precision over recall. Hence, we also measured the F$_\beta$ score with $\beta = 0.5$ which weighs precision more than recall.

| | 1:1 | | | | $n$:$m$ | | | |
|---|---|---|---|---|---|---|---|---|
| **name** | **P** | **R** | **F$_1$** | **F$_{0.5}$** | **P** | **R** | **F$_1$** | **F$_{0.5}$** |
| LHA | .94 | .41 | .57 | .747 | - | - | - | - |
| Sent-LaBSE | **.961** | .444 | .608 | **.780** | - | - | - | - |
| Sent-RoBERTa | .960 | .444 | .607 | .779 | - | - | - | - |
| CATS-C3G | .247 | **.553** | .342 | .278 | - | - | - | - |
| VecAlign | .271 | .404 | .323 | .290 | .260 | .465 | .333 | .285 |
| BERTalign | .743 | .465 | .572 | .664 | .387 | .561 | .458 | .412 |
| MASSalign | .846 | .477 | **.610** | .733 | **.819** | .509 | **.628** | **.730** |

Table 3: Results of the alignment methods with 1:1 (upper part) and $n$:$m$ capabilities (lower part) on sentence-pairs with 1:1 (n=1750, left part) and $n$:$m$ alignments (n=991, right part).

### 5.3 Results

Three of our studied alignment methods can produce only 1:1 alignments (LHA, SentenceTransformer, CATS), and the other three methods can produce additionally $n$:$m$[20] alignments (VecAlign, BertAlign, MASSAlign)

Theoretically, our ideal aligner should be able to produce $n$:$m$ alignments with high precision as splitting and merging are frequent in TS corpora. However, in our experiments, we observed that producing $n$:$m$ alignments is a difficult task. We found that SentenceTransformer using the multilingual model LaBSE (Feng et al., 2022) got very high precise 1:1 results with a fair recall as well (see Table 3). On the other hand, MASSAlign performed the best on $n$:$m$ results, and also with totally acceptable 1:1 results (see Table 3). Hence, we concluded that MASSAlign is the most suitable aligner for our use case as it i) produces $n$:$m$

[16] https://huggingface.co/sentence-transformers/LaBSE

[17] https://huggingface.co/T-Systems-onsite/cross-en-de-roberta-sentence-transformer

[18] More details on our adaptations can be found in Appendix H.

[19] $n$ and $m$ are $> 1$ in this context

[20] where $n$ or $m$ equals to 1 in this context, but not both

alignments and ii) has fairly high scores for 1:1 and $n$:$m$ alignments. Therefore, we recommend MASSAlign to be used to automatically align the documents which the web crawler can scrape.

### 5.4 Corpus Statistics

Running MASSAlign on our unaligned corpus of DEPLAIN-WEB results in 1,594 sentence alignments. Following statistics of the manually aligned part of DEPLAIN-WEB (1,846 aligned pairs of 6,138 complex sentences), theoretically, a perfect aligner should get on average a maximum of 30% alignments of the complex sentences, which corresponds to 5,980 sentence pairs on the not-aligned documents that we posses (with 19,934 complex sentences). However, as we set our experiments with the aim of getting a precise aligner that values quality over quantity, these expected numbers were much reduced in reality (to approx 8%).

## 6 Automatic Text Simplification

To exemplify the usage of DEPLAIN for training and evaluating TS models, we are presenting results on finetuning *long-mBART* on our document-level corpus as well as finetuning *mBART* on our sentence-level corpus, using code provided by Rios et al. (2021)[21].

### 6.1 Data

We have split the document and sentence pairs of DEPLAIN-APA and DEPLAIN-WEB into training, development, and testing splits, the sizes of all splits are provided in Appendix I.[22]

We are reporting evaluation metrics on the test sets of DEPLAIN-APA, and DEPLAIN-WEB for both document- and sentence-level systems. More results on other test data sets can be found in Appendix J.

### 6.2 Evaluation of Text Simplification

For evaluation, we use the following automatic metrics provided in the evaluation framework EASSE (Alva-Manchego et al., 2019): for simplification, SARI (Xu et al., 2015), for quality and semantic similarity to the target reference, BERTScore Precision (BS-P) is reported (Zhang et al., 2019), and BLEU for meaning preservation (Papineni et al., 2002), following the recommendations of Alva-Manchego et al. (2021) regarding TS evaluation on English texts. As our corpus has just one reference and not multiple as English TS corpora, e.g., Newsela (Xu et al., 2015) or ASSET (Alva-Manchego et al., 2020a), SARI might not work as good as expected. Instead of Flesch-Kincaid Grading Level (FKGL) which is built for only English data, we are reporting the German version of Flesch-Reading-Ease (for readability).[23]

As baseline we use a src2src-baseline or identity-baseline, which As baseline we report the results of a src2src-baseline or identity-baseline in which the complex source sentences are just copied and, hence, the complex sentences are used as original and potential simplification data. We cannot use a reference baseline (tgt2tgt) as used in related works, because DEPLAIN has only one reference to evaluate against, hence, the scores would always result in the highest scores, e.g., 100 for SARI.

### 6.3 Results

#### 6.3.1 Document-level Text Simplification System

The evaluation results of the document simplification systems are summarized in Table 4, In terms of SARI, all the fine-tuned models are outperforming the Identity baseline src2src on both test sets.[24]

However, against our hypothesis, the strong simplifications of the web data seems to be easier to be simplified (SARI>43) than the mild simplifications of the APA data (SARI>35).

For BLEU score, the higher the score, the more of the content was copied (Chatterjee and Agarwal, 2021). So, if the BLEU score is less for the system outputs than for the identity baseline that means that the simplification system has changed something and not only copied. Therefore, it is reasonable that the BLEU score is higher for the src2src baselines than for some system outputs (see Xu et al. 2016, Sulem et al. (2018c), (Chatterjee and Agarwal, 2021)). A human evaluation is required in order to obtain a more reliable assessment of

[21] For our experiments, we use all default parameters as reported on GitHub https://github.com/a-rios/ats-models. We finetuned all of the TS models only once, and therefore reporting only one result per TS model.

[22] We also provide the data split here: https://github.com/rstodden/DEPlain.

[23] We are using metrics even if it is shown (e.g., for SARI (Alva-Manchego et al., 2021), BLEU (Sulem et al., 2018a) and FKGL (Tanprasert and Kauchak, 2021)) that they are not perfect for measuring simplification as no other more reliable metrics exist yet. A manual evaluation would be most reliable, but this is out of the scope of this work as we are evaluating the validity of the dataset and not the TS system.

[24] Further results on another test dataset are added to the subsection J.1).

| train data | n | SARI ↑ | BLEU ↑ | BS-P ↑ | FRE ↑ |
|---|---|---|---|---|---|
| DEplain-APA | 387 | **44.56** | **38.136** | **0.598** | **65.4** |
| DEplain-web | 481 | 35.02 | 12.913 | 0.475 | 59.55 |
| DEplain-APA+web | 868 | 42.862 | 36.449 | 0.589 | 65.4 |
| src2Src-baseline | | 17.637 | 34.247 | 0.583 | 58.85 |

(a) DEPLAIN-APA test (n=48)

| train data | n | SARI ↑ | BLEU ↑ | BS-P ↑ | FRE ↑ |
|---|---|---|---|---|---|
| DEplain-APA | 387 | 43.087 | 21.9 | 0.377 | **64.7** |
| DEplain-web | 481 | 49.584 | 23.282 | **0.462** | 63.5 |
| DEplain-APA+web | 868 | **49.745** | **23.37** | 0.445 | 57.95 |
| src2Src-baseline | | 12.848 | 23.132 | 0.432 | 59.4 |

(b) DEPLAIN-WEB test (n=147)

Table 4: Results on Document Simplification using fine-tuned long-mBART. n corresponds to the length of the training data.

| train data | n | SARI ↑ | BLEU ↑ | BS-P ↑ | FRE ↑ |
|---|---|---|---|---|---|
| DEplain-APA | 10660 | 34.818 | 28.25 | 0.639 | **63.072** |
| DEplain-APA+web | 11941 | **34.904** | **28.506** | **0.64** | 62.669 |
| src2src-baseline | | 15.249 | 26.893 | 0.627 | 59.23 |

(a) DEPLAIN-APA test (n=1231)

| train data | n | SARI ↑ | BLEU ↑ | BS-P ↑ | FRE ↑ |
|---|---|---|---|---|---|
| DEplain-APA | 10660 | 30.867 | 15.727 | 0.413 | 64.516 |
| DEplain-APA+web | 11941 | **34.828** | **17.88** | **0.436** | **65.249** |
| src2src-baseline | | 11.931 | 20.85 | 0.423 | 60.825 |

(b) DEPLAIN-WEB test (n=1846)

Table 5: Results on Sentence Simplification using fine-tuned mBART. n corresponds to the length of the training data.

the quality of the results, as although those metrics are the traditionally used metrics in this area, they were originally designed for sentences evaluation, and not documents evaluation.

### 6.3.2 Sentence-Level Text Simplification System

The evaluation results of the sentence level systems are summarized in Table 5.[25] These results can be seen as baselines for further experiments with the DEPLAIN corpus.

Comparing the FRE of DEPLAIN-APA and DEPLAIN-APA+WEB on the two test sets, the DEPLAIN-APA test always achieves a higher FRE. Also the model that was trained on APA+web data did not make a big difference from the one trained only on APA when tested on the APA test set; the additional web data does not affect the model much in this case. However, when tested on the web test set, adding the web data to the training data has improved all the measured metrics. This supports that adding training data from different contexts leads to a better generalization of the model. The combination of DEPLAIN-APA+WEB achieves the highest scores in terms of all metrics.

Although our data doesn't include simplifications for children, the SARI scores on the ZEST-test set are better than the reported models (increase by approx. 5 points on SARI, see appendix J.2, Table 16). This result might be due to issues with the automatic TS metrics. A manual evaluation is required to justify if finetuning mBART on DEPLAIN-APA+WEB can really simplify texts for children. The target group of the texts a model is trained on should always be considered in the interpretation of model evaluation as each target group requires different simplification operations (Gooding, 2022).

## 7 Conclusion and Future Work

In this paper, we have introduced a new German corpus for text simplification, called DEplain. The corpus contains data for document simplification as well as sentence simplification of news (DEPLAIN-APA) and web data (DEPLAIN-WEB). The major part of the sentence-wise alignments are manually aligned and a part of it is also manually analyzed. The analysis shows that the subcorpus DEPLAIN-APA contains rather mild simplifications whereas DEPLAIN-WEB contains rather strong simplifications. However, for both corpora, a large variety of simplification operations were identified.
Furthermore, we evaluated automatic sentence alignment methods on our manually aligned data. In our experiments, MASSalign got the best results but it has only identified a few $n$:$m$ alignments (where $n > 1$ or $m > 1$). One direction for future work is to further investigate $n$:$m$ alignments algorithms for TS corporaand include paragraphs into the automatic alignment process. We also showed first promising experiments on sentence and document simplification. However, these are just simple benchmarks and have only been evaluated with automatic metrics yet. In future work, the results should be verified by manual evaluation and could be improved by using more sophisticated approaches. Finally, we think that DEPLAIN can boost and improve the research in German text simplification, to make more complex texts accessible to people with reading problems.

[25] Further results on other test data are added to Appendix J.

## Acknowledgments

We gratefully acknowledge the support of the Austria Presse Agentur (Austrian Press Agency, APA) for providing us with the news documents and Maya Kruse for her work on annotating the corpus. We further thank NVIDIA for the access to GPUs. This research is part of the PhD-program "Online Participation", supported by the North Rhine-Westphalian (German) funding scheme "Forschungskolleg".

## Limitations

However, our work shows some limitations. A major restriction of our data is the different licenses of our proposed dataset, i.e., i) DEPLAIN-APA can be obtained for free for research purposes upon request, ii) the manual aligned part of DEPLAIN-WEB and the smaller part of the automatically aligned sentences are available under open licenses, e.g., CC-BY-4.0, iii) the other automatically aligned sentence pairs and their documents are not allowed to be shared. However, the web harvester and an automatic alignment method can be used to reproduce the document and sentence pairs.

But, the web crawler does not perform well for all web pages and extracts one-token sentences. While this is not reflected in the manually aligned sentence pairs due to manual quality checks before alignment, this does not happen for the automatic alignment. As a consequential error, the text simplification model would also learn wrong sentence structures and simplifications from this data. Therefore, more time in automatic data cleaning is required. In addition, the web crawler is currently mostly extracting HTML documents and only a few PDF files. The corpus could be increased if existing parallel PDF files would be crawled and correctly extracted. Furthermore, the web crawler just harvest a given set of web pages and do not search in the whole web for parallel German complex-simplified documents such as a general web crawler. Currently, a general web crawler wouldn't add much more parallel data than the proposed one, as this data is scarce at the moment and, if parallel data is available, there is no link between complex and simplified documents and the title of the pages are often that different that they cannot be aligned automatically. However, if in future more parallel texts are available we would like to extend our corpus with a general web crawler to also include more variance within the domains.

Compared to TS corpora in English, e.g., Wiki-Auto or Newsela-Auto (Jiang et al., 2020), DEPLAIN is smaller and is not (fully) balanced in terms of domains. However, we believe that the current size of the dataset is already large enough due to the high proportion of professionally simplified texts and the high-quality of manual alignments. Further, we are aiming at increasing the corpus following our dynamic approach, e.g., by extending the capabilities of the web harvester or the alignment algorithms.

Furthermore, the automatic alignment methods currently align mostly 1:1 alignments but $n{:}m$ alignments are important for text simplification and should be considered for training. Therefore, the automatic alignment methods should be improved in the direction of $n{:}m$ alignments. Until then, we would recommend using the automatically aligned sentence pairs only for additional training data.

For document simplification, a GPU with at least 24 GB of memory is required to reproduce our results with mBART and at least 16 GB for sentence simplification respectively. However, for other approaches, e.g., unsupervised learning or zero-shot approaches, the experiments with DEPLAIN-APA and DEPLAIN-WEB on the document- and sentence-level could be performed with less memory.

For the evaluation of our text simplification models, we are just reporting automatic alignments, although they are mainly built for English TS and their quality is not evaluated on German yet. Our datasets has just one reference and not multiple as in ASSET or TurkCorpus, therefore SARI might work not as good as with several references. In addition, in some work (e.g., Alva-Manchego et al. (2021)) it was already shown that the automatic metrics do not perfectly correlate with human judgments, hence, the results of the automatic metrics should be interpreted with caution. It is recommended to manually evaluate the results, but this was out of the scope of this work which mainly focuses on proposing a new dataset and not new TS models.

## Ethics & Impact Statement

### Data Statement

The data statement for DEPLAIN is available here: https://github.com/rstodden/DEPlain.

### NLP application statement

**Intended use.** In this work, we propose a new corpus for training and evaluating text simplification models. It is intended to use this corpus for training text simplification models or related works, e.g., text style transfer. The resulting text simplification system is intended to be used to simplify texts for a given target group (depending on the training data). However, the generated simplifications of the TS model might have some errors, therefore they shouldn't be shown to a potentially vulnerable target group before manually verifying their quality and possibly fixing them. The text simplification system could be provided to human translators who might improve and timely reduce their effort in manually simplifying a text.

Furthermore, the dataset can be reused for related tasks to TS, this includes, but is not limited to, text leveling, evaluation of alignment methods, evaluation of automatic TS metrics, and analysis of intralingual translations.

**Misuse potential & Failure modes.** One potential misuse of DEPLAIN is to reverse the input order of the texts into a deep learning model. The resulting system would be able to make the texts even more complex than simplifying them. Although a system that would be developed for producing complex texts can be used for beneficial use cases (e.g., generation of more challenging texts for language learners), however, it could be used to obscure pieces of information from some kind of audience on purpose. Furthermore, the TS system could generate content with low similarity to the complex sentence given as input and, hence, change the meaning of the original text. Due to this, it should always be stated that the simplification is generated automatically and might not reflect the original meaning of the source text.

**Biases.** No biases are known yet.

**Collecting data from users.** When a researcher requests access to the DEPLAIN-APA corpus, the name, the institution, and the email address of the researcher are saved by the authors of the paper. This is required to make the use of the dataset transparent to the data provider, i.e., the Austrian Press Agency.

**Environmental Report.** For the manual alignment and annotation of the corpus, a server with the text simplification tool has run all time during the annotation duration. For the evaluation of all alignment methods, we required less than 1 GPU hour on an NVIDIA RTX A5000 with 24 GB. The experiments with document and sentence simplification, overall, took less than 18 GPU hours on a NVIDIA RTX A5000 with 24 GB.

## References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020a. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020b. 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.

Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Studenten-Schreib-Service. PhD Thesis.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Ben Athiwaratkun, Andrew Wilson, and Anima Anandkumar. 2018. Probabilistic FastText for multi-sense word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–11, Melbourne, Australia. Association for Computational Linguistics.

Dennis Aumiller and Michael Gertz. 2022. Klexikon: A German dataset for joint summarization and simplification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2693–2701, Marseille, France. European Language Resources Association.

Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. 2020. A corpus for automatic readability assessment and text simplification of German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3302–3311, Marseille, France. European Language Resources Association.

Stefan Bott and Horacio Saggion. 2014. Text Simplification Resources for Spanish. *Language Resources and Evaluation*, 48(1):93–120.

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699, Brussels, Belgium. Association for Computational Linguistics.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first Italian corpus for text simplification. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 31–41, Denver, Colorado, USA. Association for Computational Linguistics.

Klaus Buddeberg and Anke Grotlüschen. 2020. *LEO 2018: Leben mit geringer Literalität*. wbv.

Rémi Cardon, Adrien Bibal, Rodrigo Wilkens, David Alfter, Magali Norré, Adeline Müller, Watrin Patrick, and Thomas François. 2022. Linguistic corpus annotation for automatic text simplification evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1842–1866, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Niladri Chatterjee and Raksha Agarwal. 2021. Depsym: A lightweight syntactic text simplification approach using dependency trees. In *Proceedings of the First Workshop on Current Trends in Text Simplification (CTTS 2021)*, pages 42–56. CEUR-WS.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006, Dubrovnik, Croatia. 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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. 2018. The corpus of Basque simplified texts (CBST). *LANGUAGE RESOURCES AND EVALUATION*, 52(1):217–247.

Sian Gooding. 2022. On the ethical considerations of text simplification. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57, Dublin, Ireland. Association for Computational Linguistics.

Silvia Hansen-Schirra, Jean Nitzke, and Silke Gutermuth. 2021. *An Intralingual Parallel Corpus of Translations into German Easy Language (Geasy Corpus): What Sentence Alignments Can Tell Us About Translation Strategies in Intralingual Translation*, pages 281–298. Springer Singapore, Singapore.

Freya Hewett and Manfred Stede. 2021. Automatically evaluating the conceptual complexity of German texts. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 228–234, Düsseldorf, Germany. KONVENS 2021 Organizers.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria. Association for Computational Linguistics.

Joongwon Kim, Mounica Maddela, Reno Kriz, Wei Xu, and Chris Callison-Burch. 2021. BiSECT: Learning to split and rephrase sentences with bitexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6193–6209, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Klaper, Sarah Ebling, and Martin Volk. 2013. Building a German/simple German parallel corpus for automatic text simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria. Association for Computational Linguistics.

Anaïs Koptient, Rémi Cardon, and Natalia Grabar. 2019. Simplification-induced transformations: typology and some characteristics. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 309–318, Florence, Italy. Association for Computational Linguistics.

Lei Liu and Min Zhu. 2022. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*. Fqac089.

Christiane Maaß. 2020. *Easy Language - Plain Language - Easy Language Plus. Balancing Comprehensibility and Acceptability*. Easy – Plain – Accessible. Frank & Timme, Berlin.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126, Online. Association for Computational Linguistics.

Robert Marzari. 2010. *Leichtes Englisch, schwieriges Französisch, kompliziertes Russisch: Evaluation der Schwierigkeiten des Englischen, Deutschen, Französischen, Italienischen, Spanischen, Russischen und Polnischen als Fremdsprache*. Schiler.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective assessment of text complexity: A dataset for german language.

Nikola I. Nikolov and Richard Hahnloser. 2019. Large-scale hierarchical alignment for data-driven text rewriting. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 844–853, Varna, Bulgaria. INCOMA Ltd.

Gustavo Paetzold, Fernando Alva-Manchego, and Lucia Specia. 2017. MASSAlign: Alignment and annotation of comparable documents. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 1–4, Tapei, Taiwan. Association for Computational Linguistics.

Gustavo Henrique Paetzold and Lucia Specia. 2016. Vicinity-driven paragraph and sentence alignment for comparable corpora. *CoRR*, abs/1612.04113.

Alessio Palmero Aprosio, Sara Tonelli, Marco Turchi, Matteo Negri, and Mattia A. Di Gangi. 2019. Neural text simplification in low-resource conditions using weak supervision. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 37–44, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stefan Paun. 2021. Parallel text alignment and monolingual parallel corpus creation from philosophical texts for text simplification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 40–46, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

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, Online and in Dominican Republic. Association for Computational Linguistics.

Andreas Säuberli, Sarah Ebling, and Martin Volk. 2020. Benchmarking data-driven automatic text simplification for German. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 41–48, Marseille, France. European Language Resources Association.

Thorben Schomacker. 2023. Aligned narrative documents. `https://github.com/tschomacker/aligned-narrative-documents`.

Melanie Siegel, Dorothee Beermann, and Lars Hellan. 2019. Aspects of linguistic complexity: A german - norwegian approach to the creation of resources for easy-to-understand language. In *11th International Conference on Quality of Multimedia Experience QoMEX 2019, Berlin, Germany, June 5-7, 2019*, pages 1–3. IEEE.

Nicolas Spring, Marek Kostrzewa, Annette Rios, and Sarah Ebling. 2022. Ensembling and score-based filtering in sentence alignment for automatic simplification of german texts. In *Universal Access in Human-Computer Interaction. Novel Design Approaches and Technologies*, pages 137–149, Cham. Springer International Publishing.

Nicolas Spring, Annette Rios, and Sarah Ebling. 2021. Exploring German multi-level text simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1339–1349, Held Online. INCOMA Ltd.

Sanja Štajner. 2021. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652, Online. Association for Computational Linguistics.

Sanja Štajner, Marc Franco-Salvador, Simone Paolo Ponzetto, Paolo Rosso, and Heiner Stuckenschmidt. 2017. Sentence alignment methods for improving text simplification systems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 97–102, Vancouver, Canada. Association for Computational Linguistics.

Sanja Štajner, Marc Franco-Salvador, Paolo Rosso, and Simone Paolo Ponzetto. 2018. CATS: A tool for customized alignment of text simplification corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Regina Stodden. 2021. When the scale is unclear – analysis of the interpretation of rating scales in human evaluation of text simplification. In *Proceedings of the First Workshop on Current Trends in Text Simplification (CTTS 2021)*, pages 84–95. CEUR-WS.

Regina Stodden and Laura Kallmeyer. 2022. TS-ANNO: An annotation tool to build, annotate and evaluate text simplification corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 145–155, Dublin, Ireland. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018a. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018b. Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 685–696, New Orleans, Louisiana. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018c. Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 162–173, Melbourne, Australia. Association for Computational Linguistics.

Renliang Sun, Zhe Lin, and Xiaojun Wan. 2020. On the helpfulness of document context to sentence simplification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1411–1423, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2020. Exploiting sentence order in document alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5997–6007, Online. Association for Computational Linguistics.

Vanessa Toborek, Moritz Busch, Malte Boßert, Christian Bauckhage, and Pascal Welke. 2022. A new aligned simple german corpus.

Jan Trienes, Jörg Schlötterer, Hans-Ulrich Schildhaus, and Christin Seifert. 2022. Patient-friendly clinical notes: Towards a new text simplification dataset. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 19–27, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Jan Trienes and Laura Vásquez-Rodríguez. 2023. Text simplification datasets. `https://github.com/jantrienes/text-simplification-datasets`.

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. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

# Appendix

## A Overview of Existing German TS Corpora

In Table 6, an overview of existing German text simplification corpora is shown. For Hewett and Stede (2021) we report the numbers from the updated version of March 2022.

## B Worse Examples of APA-LHA

In the APA-LHA corpus (Spring et al., 2021) we found original sentences repeated several times in the training data aligned to multiple different simplified sentences (in Table 7 called "simplifications"). This format can either comprises different simplifications for the same complex sentence or a split of one complex sentence into several simple sentences.

If we see these simplifications as alternative simplifications, some original-simple pairs seems to be wrongly aligned. In Table 7, we show two alignment pairs in which the meaning is heavily changed. In the first example, the original and all simplifications are related to career but at different states, i.e., looking back at the career, starting the career and quitting the career. In the second example, the terms of unemployment and short time are mixed and also the numbers are totally different.

In row three and four of Table 7, we provide more examples regarding the unclear format, it is not clear whether pairs with identical complex sentences are alternative simplifications (references) (see row 3) or 1:$m$ alignments (see row 4).

## C Examples of Mild and Strong Simplifications

In Table 8, some examples of strong and mild simplifications of DEPLAIN are provided including English translations.

## D Inter-Annotator Agreement

In Table 9, we show an overview of the inter-annotator agreement per domain.

## E Details on DEPLAIN-WEB

In this section, we will describe more details on the web harvester and the process of creating the dataset.

### E.1 Overview of Web Pages

The web pages in Table 10 were crawled for generating DEPLAIN-WEB. We selected these pages based on a web research regarding web pages in German plain language ("Einfache Sprache"). We further checked the references of translation offices, e.g., which web pages are simplified by them and if they contain parallel alignments.

### E.2 Document Alignment

The documents are aligned with three strategies in the following order: i) automatic alignment by the reference to the simple document within the complex documents, ii) automatically matching the titles of the documents on the website, and iii) aligning the documents manually. All the books in the fiction domain were manually aligned on the document level as the complex data is provided on another web page (i.e. Projekt Gutenberg[26]) than the simplified data (i.e., Spaß am Lesen Verlag[27], Passanten Verlag[28], or NDR[29]). For the simple books, only a preview was available, therefore we only added the first section of the complex book. If the simplified book summarizes parts of more than the first chapter, the documents might be not comparable. We haven't checked that manually.

In addition, to download and align the HTML files, the web crawler also extracts some metadata, the plain text of the documents, and the plain text including paragraph endings.

To further align the documents on paragraph or sentence-level the crawler can be integrated into existing alignment tools, e.g., TS-anno (Stodden and Kallmeyer, 2022).

### E.3 Technical Details.

We used Python 3 and the Python Package Beautiful Soup for the HTML data and pymupdf for the PDF data. The code of the web crawler is freely available under the CC-BY-4.0 license and can be accessed via https://github.com/rstodden/DEPlain.

Because the texts are on editable websites the content may change. Therefore we note the date when we crawled the data, the data can then

[26] https://www.projekt-gutenberg.org/

[27] https://einfachebuecher.de/

[28] https://www.passanten-verlag.de/

[29] https://www.ndr.de/fernsehen/barrierefreie_angebote/leichte_sprache/Maerchen-in-Leichter-Sprache,maerchenleichtesprache100.html

| Reference | Name | Target Simple | Domain | Availability | # Docs | # Sent. Complex | # Sent. Simple | # Aligned Pairs | Alignment |
|---|---|---|---|---|---|---|---|---|---|
| Siegel et al. (2019) | leichte-sprache-corpus | mixed | web | https://github.com/hdaSprachtechnologie/easy-to-understand_language | 351 | | | | |
| Hewett and Stede (2021) | Lexica-corpus-klexikon | children between 6-12 | wikipedia | https://github.com/fhewett/lexica-corpus | 1090 | | | | auto |
| Hewett and Stede (2021) | Lexica-corpus-miniklexikon | children younger than 6 | wikipedia | https://github.com/fhewett/lexica-corpus | 1090 | | | | auto |
| Rios et al. (2021)* | 20Minuten | general | news | https://github.com/ZurichNLP/20Minuten | 18305 | | | | ? |
| Aumiller and Gertz (2022) | Klexikon | children between 6-12 | wikipedia | https://github.com/dennlinger/klexikon | 2898 | 701577 | 94214 | | auto |
| Ebling et al. (2022) | Wikipedia-Corpus | A2 | wikipedia | - | 106126 | 6933192 | 1077992 | | ? |
| Trienes et al. (2022) † | simple-patho | laypeople | medical | https://github.com/jantrienes/simple-patho (not yet available) | 851 | 23,554 | 28,155 | 2,280 (paragraphs) | manual |
| Schomacker (2023) | MILS+EB+PV+KV | mixed | fiction | https://github.com/tschomacker/aligned-narrative-documents (not yet available) | | | | | |

(a) Overview of German simplification corpora on document-level.

| Reference | Name | Target Simple | Domain | Availability | # Docs | # Sent. Complex | # Sent. Simple | # Aligned Pairs | Alignment |
|---|---|---|---|---|---|---|---|---|---|
| Klaper et al. (2013) | Klaper | Leichte Sprache | web | upon request | 256 | | | approx. 2000 | manual&auto |
| Naderi et al. (2019) | TextComplexityDE19 | Non-native speaker (written by non-native speaker) | wikipedia | https://github.com/babaknaderi/TextComplexityDE | 23 | | | 250 | manual |
| Battisti et al. (2020); Ebling et al. (2022) | Web | A2 | web | - | 378 | 17121 | 21072 | | CATS |
| Mallinson et al. (2020) | ZEST-data | children between 5-7 | science for children | https://github.com/Jmallins/ZEST-data | 20 | | | 1198 | manual |
| Säuberli et al. (2020); Ebling et al. (2022) ‡ | APA-benchmark | B1 | news | - | | | | 3616 | CATS-WAVG |
| Kim et al. (2021) | BiSECT | | web & politics | https://github.com/mounicam/BiSECT | | | | 186,237 | |
| Hansen-Schirra et al. (2021) | GEASY | Leichte Sprache | mixed | - | 93 | 1596 | 4090 | | memsource & (commercial) |
| Spring et al. (2021); Ebling et al. (2022) ‡ | APA-LHA-or-a2 | A2 | news | https://zenodo.org/record/5148163 | 2426 | 60732 | 30432 | 9456 | LHA |
| Spring et al. (2021); Ebling et al. (2022) ‡ | APA-LHA-or-b1 | B1 | news | https://zenodo.org/record/5148163 | 2426 | 60732 | 30328 | 10268 | LHA |
| Spring et al. (2021); Ebling et al. (2022) | capito | B1 | news | - | 1055 | 183216 | 68529 | 54224 | LHA |
| Spring et al. (2021); Ebling et al. (2022) | capito | A2 | news | - | 1546 | 183216 | 168950 | 136582 | LHA |
| Spring et al. (2021); Ebling et al. (2022) | capito | A1 | news | - | 839 | 183216 | 24243 | 10952 | LHA |
| Toborek et al. (2022) | Simple German Corpus | A1 | web | https://github.com/buschmo/Simple-German-Corpus | 530 | | | 5889 | CATS |

(b) Overview of German simplification corpora on sentence-level.

| Reference | Name | Target Simple | Domain | Availability | # Docs | # Sent. Complex | # Sent. Simple | # Aligned Pairs | Alignment |
|---|---|---|---|---|---|---|---|---|---|
| | DEplain-apa ‡ | A2 | news | https://github.com/rstodden/DEPlain | 483 | 14,071 (aligned) | 16,505 (aligned) | 13122 | manual |
| | DEplain-web | mixed | web | https://github.com/rstodden/DEPlain | 147 (+609) | 2,287 (aligned) | 4,009 (aligned) | 1856 (+1594) | manual&auto |

(c) Overview of DEplain, the proposed German simplification corpus on document- and sentence-level.

Table 6: Overview of German simplification corpora. All corpora contain German languages (no dialect specified) except 20Minuten (see *, Swiss German, de-CH), APA-benchmark, APA-LHA, DEplain-apa (see ‡, Austrian German, DE-AT). All source texts from all corpora address a general audience, except simple-path (see †).

| complex-ids | Error Description | Original | Simplifications |
|---|---|---|---|
| A2#35;A2#235 | Original and all simplifications are related to “Karriere” (career) but with different meaning. | “Auf seine Karriere blickte er gerne zurück .”, | “Er begann seine Karriere mit 18 Jahren .”,<br>“Er beendet jetzt seine Karriere .” |
| A2#2621;A2#265;<br>A2#6530;A2_dev#85;<br>A2_dev#417 | The simplifications contain a mix of i) “unemployment” and “short time work”, ii) different digits, and iii) misalignment across documents. | “Derzeit sind damit aber noch immer über 123.000 Personen mehr arbeitslos als vor der Coronakrise .” | “Derzeit sind in Österreich auch noch 1,3 Millionen Menschen in Kurz-Arbeit .”,<br>“Die Kurz-Arbeit gibt es , damit nicht noch mehr Menschen ihre Arbeit verlieren .”,<br>“Dort gibt es jetzt 5.000 Arbeitslose weniger als vor einer Woche .”,<br>“Vor einer Woche waren noch 9.000 Menschen mehr arbeitslos .”,<br>“Ihr Geld bekommen sie aber nicht mehr von den Firmen , sondern vom Staat .” |
| B1#3392;B1#6763 | The simplifications can be interpreted as a split of the original sentence into two sentences (one 1:m simplification). | “Laut Polizei dürfte das Kind nach dem im Garten für den Vierbeiner abgelegten Futter gegriffen haben , als es zu der Attacke kam .” | “Laut Polizei griff der Bub im Garten nach dem Hundefutter .”,<br>“Dabei kam es zu der Attacke .” |
| A2#246; A2#5966 | The simplifications can be interpreted as alternative simplifications (two 1:1 simplifications). | “Menschen können Hunde und Katzen mit Coronavirus anstecken” | “Menschen können Hunde und Katzen mit dem Corona-Virus anstecken .”,<br>“Hunde und Katzen können mit dem Corona-Virus angesteckt werden .’ |

(a) Original German version.

| complex-ids | Error Description | Original | Simplifications |
|---|---|---|---|
| A2#35;A2#2352 | Original and all simplifications are related to “Karriere” (career) but with different meaning. | “He looks back on his career with pleasure .” | “He began his career at the age of 18 .”,<br>“He is now ending his career .” |
| A2#2621;A2#265;<br>A2#6530;A2_dev#85;<br>A2_dev#417 | The simplifications contain a mix of i) “unemployment” and “short time work”, ii) different digits, and iii) misalignment across documents. | “At present , however , this still leaves over 123,000 more people unemployed than before the Corona crisis .” | “Currently, 1.3 million people in Austria are also still in short-time work .”,<br>“The short-time work exists so that more people do not lose their jobs .”,<br>“There are now 5,000 fewer unemployed there than a week ago .”,<br>“A week ago , 9,000 more people were unemployed .”,<br>“But they no longer get their money from the companies , but from the state.” |
| B1#3392;B1#6763 | The simplifications can be interpreted as a split of the original sentence into two sentences (one 1:m simplification). | ‘According to police , the child may have reached for the food placed in the garden for the quadruped , when it came to the attack .” | “According to police , the boy reached for the dog food in the garden .”,<br>In the process, the attack occurred .” |
| A2#246;A2#5966 | The simplifications can be interpreted as alternative simplifications (two 1:1 simplifications). | “People can infect dogs and cats with coronavirus” | “Humans can infect dogs and cats with the Corona virus . ”,<br>“Dogs and cats can be infected with the Corona virus .” |

(b) Translated English version.

Table 7: Excerpt of worse automatically aligned sentence-level pairs in the training data of APA-LHA (Spring et al., 2021). All examples are contained in the training data. The complex-ids are a concatenation of the name of the dataset (either C2 to A2 or to B1) and the line number.

| | original | simplification | original (English) | simplification (English) | domain | OL | SL | source |
|---|---|---|---|---|---|---|---|---|
| **mild** | Innenminister Herbert Kickl gab am Donnerstag bekannt, dass im Jahr 2018 jedes 2. Verbrechen aufgeklärt wurde. | Außerdem wurde im Jahr 2018 jedes 2. Verbrechen aufgeklärt. Das gab Innenminister Herbert Kickl am Donnerstag bekannt. | Interior Minister Herbert Kickl announced Thursday that every 2nd crime was solved in 2018. | In addition, every 2nd crime was solved in 2018. This was announced by Interior Minister Herbert Kickl on Thursday. | news | B1 | A2 | Austria Press Agency |
| **mild** | Da entstand Helligkeit. | Und es wurde hell. | That's when brightness arose. | And it became bright. | bible | C2 | A1 | Offene Bibel |
| **strong** | Über dem Tisch, auf dem eine auseinandergepackte Musterkollektion von Tuchwaren ausgebreitet war – Samsa war Reisender – hing das Bild, das er vor kurzem aus einer illustrierten Zeitschrift ausgeschnitten und in einem hübschen, vergoldeten Rahmen untergebracht hatte. | Auf dem Tisch sind noch immer die Stoffe ausgebreitet. Gregor ist von Beruf Vertreter. Seine Aufgabe ist es, Stoffe zu verkaufen. Dafür reist er umher. Gregor sieht sich weiter in seinem Zimmer um. Über dem Tisch hängt immer noch das Bild. Das Bild, das er vor ein paar Tagen aus einer Zeitschrift ausgeschnitten hat. Gregor hat es in einem schönen Rahmen aufgehängt. In einem goldenen Bilder -Rahmen. | Above the table on which was spread an unpacked sample collection of drapery - Samsa was a traveler - hung the picture he had recently cut out of an illustrated magazine and placed in a handsome gilt frame. | On the table, the fabrics are still spread out. Gregor is a salesman by profession. His job is to sell fabrics. For that, he travels around. Gregor continues to look around his room. Above the table still hangs the picture. The picture he cut out of a magazine a few days ago. Gregor has hung it in a beautiful frame. In a golden picture frame. | fiction | C2 | A2 | Spaß am Lesen Verlag |
| **strong** | Solange keine vollständige Belastung möglich ist, muss eine Thromboseprophylaxe durch die Gabe von niedermolekularem Heparin durch Spritzen erfolgen. | Der Arzt gibt Ihnen alle wichtigen Informationen. Sprechen Sie deshalb mit Ihrem Arzt. Eine Thrombose ist gefährlich. Die Spritzen sind gegen eine Thrombose. Deshalb müssen Sie vielleicht Spritzen bekommen. Dann dürfen Sie nicht mit dem Fuß auftreten. Manchmal müssen Sie den Fuß mehrere Wochen schonen. | As long as complete weight-bearing is not possible, thrombosis prophylaxis must be given by administration of low-molecular-weight heparin by injection. | The doctor will give you all the important information. Therefore, talk to your doctor. Thrombosis is dangerous. The injections are against thrombosis. Therefore, you may have to get injections. Then you must not step with your foot. Sometimes you need to rest the foot for several weeks. | health | C2 | A2 | Wort & Bild Verlag Konradshöhe GmbH & Co. KG |

Table 8: Examples of mild (upper part) and strong simplifications (lower part) in different domains including the CEFR level of the original (OL) and the simplification (SL).

| domain | avg. | std. | interpretation | # sents | # docs |
|---|---|---|---|---|---|
| bible | 0.7011 | 0.31 | moderate | 6903 | 3 |
| fiction | 0.6131 | 0.39 | moderate | 23289 | 3 |
| health | 0.5147 | 0.28 | weak | 13736 | 6 |
| language learner | 0.9149 | 0.17 | almost perfect | 18493 | 65 |
| news | 0.7497 | 0.28 | moderate | 25224 | 10 |
| all | 0.8505 | 0.23 | strong | 87645 | 87 |

Table 9: Inter-annotator agreement per domain including average, standard deviation, number of sentence combinations (# sents), and number of documents (# docs).

be extracted using a web archieve, e.g., https://archive.org/web/. It might be possible that the data is updated or that some sources are not available anymore.

For each source, we aimed at downloading the complete relevant content of the websites. We removed parts such as navigation, advertisement, contact data, and other unnecessary stuff.

## F DEplain Alignment Statistics

In this section, we show statistics of DEPLAIN regarding the alignment of the manual aligned documents. Table 11 summarize numbers of DEPLAIN-APA and DEPLAIN-WEB with respect to $n : m$ alignments, where $n$ and $m$ are $> 0$, including rephrasing, splitting, merging, and fusion. Table 12 summarize numbers of DEPLAIN-APA and DEPLAIN-WEB with respect to $n : m$ alignments, where $n$ and $m$ are $\leq 1$, including copied sentences from the complex to the simplified document as well as deletions in the complex documents and additions in the simplified documents. These oddment pairs were automatically extracted after the documents were manually aligned.

## G Manual Evaluation Rating Aspects

In this section, we summarize the aspects used for manual evaluation as well as the accompanied statements. The statements are translated to English, they were shown to the annotators in German. All of the aspects were rated on a 5-point Likert scale, either from -2 to +2 or 1 to 5. An overview of the aspects is shown in Table 13.

## H Description of Adaptations of Alignment Methods

In this section, we are describing the alignment methods and the adaptations we made.

**LHA** (Nikolov and Hahnloser, 2019) is an unsupervised method that finds 1:1 sentence alignments in monolingual parallel corpora where documents don't need to be aligned beforehand. It works with a hierarchical strategy by aligning documents on the first level and then aligning sentences within these documents. This method was recommended by Ebling et al. (2022) for aligning German parallel documents.

| subcorpus | website simple | website complex | simple | complex | domain | description | # doc. |
|---|---|---|---|---|---|---|---|
| **EinfacheBücher** | https://einfachebuecher.de/ | https://www.projekt-gutenberg.org/ | PG | SG/OG | fiction | Books in plain German | 15 |
| **EinfacheBücherPassanten** | https://www.passanten-verlag.de/ | https://www.projekt-gutenberg.org/ | PG | SG/OG | fiction | Books in plain German | RS: 4 |
| **ApothekenUmschau** | https://www.apotheken-umschau.de/einfache-sprache/ ‡ | https://www.apotheken-umschau.de/einfache-sprache/ | PG | SG | health | Health magazine in which diseases are explained in plain German | 71 |
| **BZFE** | https://www.bzfe.de/einfache-sprache/ † | https://www.bzfe.de | PG | SG | health | Information of the German Federal Agency for Food on good nutrition | 18 |
| **Alumniportal** | https://www.alumniportal-deutschland.org/services/sitemap/ † | https://www.alumniportal-deutschland.org/services/sitemap/ | PG | PG | language learner | Texts related to Germany and German traditions written for language learners. | 137 |
| **Lebenshilfe** | https://www.lebenshilfe-main-taunus.de/inhalt/ | https://www.lebenshilfe-main-taunus.de/inhalt/ | ETR | SG | accessibility | | 49 |
| **Bibel** | https://offene-bibel.de/ †‡ | https://offene-bibel.de/ | ETR | SG | bible | Bible texts in easy-to-read German | 221 |
| **NDR-Märchen** | https://www.ndr.de/fernsehen/barrierefreie_angebote/leichte_sprache/Maerchen-in-Leichter-Sprache,maerchenleichtesprache100.html ‡ | https://www.projekt-gutenberg.org/ | ETR | SG/OG | fiction | Fairytales in easy-to-read German | 10 |
| **EinfachTeilhaben** | https://www.einfach-teilhaben.de/DE/LS/Home/leichtesprache_node.html | https://www.einfach-teilhaben.de | ETR | SG | accessibility | | 67 |
| **StadtHamburg** | https://www.hamburg.de/hamburg-barrierefrei/leichte-sprache/ | https://www.hamburg.de | ETR | SG | public authority | Information of and regarding the German city Hamburg | 79 |
| **StadtKöln** | https://www.stadt-koeln.de/leben-in-koeln/soziales/informationen-leichter-sprache | https://www.stadt-koeln.de | ETR | SG | public authority | Information of and regarding the German city Cologne | 85 |

Table 10: This table summarizes the web pages (including metadata) which can be extracted with the web crawler. The line separates the documents in plain German from those in easy-to-read German. *simple* correspond to the language level of the simplified documents, and *complex* of the complex documents, where PG=plain German, ETR=easy-to-read German, SG=standard German, OG=old German. The documents marked with † are openly licensed and therefore part of DEPLAIN-WEB (row 2 and row 3 in Table 1). All other documents are part of DEPLAIN-WEB (row 4 in Table 1). The data provider of the documents marked with ‡ explicitely state that their documents are professionally simplified and reviewed by the target group.

| Name | # pairs | 1:1 (rephrase) | 1:n (split) | n:1 (merge) | n:m (fusion) |
|---|---|---|---|---|---|
| **DEPLAIN-APA** | 13122 | 9912 | 2360 | 382 | 468 |
| **DEPLAIN-WEB** | 1846 | 863 | 796 | 77 | 110 |

Table 11: Statistics on $n : m$ alignments on manual aligned documents, where $n$ and $m$ are $> 0$.

| Name | # pairs | 1:1 (identical) | 0:1 (addition) | 1:0 (deletion) |
|---|---|---|---|---|
| **DEPLAIN-APA** | 12353 | 2712 | 3964 | 5677 |
| **DEPLAIN-WEB** | 5482 | 887 | 1572 | 3050 |

Table 12: Statistics of additional $n : m$ aligned pairs on manual aligned documents, where $n$ and $m$ are $\leq 1$.

Our adaptation of this method is comprised of: i) disabling the first level of aligning the documents as we already had the true document alignments, and ii) modifying the language-dependent tools used within the algorithm to fit the German language (e.g., the stopwords list, the tokenizer model, and the word embeddings model).

**Sentence Transformer** (Reimers and Gurevych, 2020) is a simple straightforward method to find 1:1 sentence alignments by computing cosine similarity between embeddings vectors (produced by a sentence transformer model) of sentences on both sides of the monolingual parallel corpora, and then picking the most similar pairs and labeling them as *aligned*. This method is totally dependent on the used sentence transformer model, and the similarity threshold to set.

| item | Statement |
|---|---|
| **Grammaticality** | The simplified sentence is fluent, and there are no grammatical errors. |
| **Grammaticality (original)** | The original sentence is fluent, there are no grammatical errors. |
| **Simplicity (simple)** | The simplified sentence is easy to understand. |
| **Simplicity (original)** | The original sentence is easy to understand. |
| **Coherence (simple)** | The simplified sentence is understandable without reading the whole paragraph. |
| **Coherence (original)** | The original sentence is understandable without reading the whole paragraph. |
| **Meaning Preservation** | The simplified sentence adequately expresses the meaning of the original sentence, perhaps omitting the least important information. |
| **Overall Simplicity** | The simplified sentence is easier to understand than the original sentence. |
| **Structural Simplicity** | The structure of the simplified sentence is easier to understand than the structure of the original sentence. |
| **Lexical Simplicity** | The words of the simplified sentence are easier to understand than the words of the original sentence. |

Table 13: Statements of the manual evaluation aspects.

Our adaptation of this method is comprised of: i) testing with new and different sentence transformer models that are either multi-lingual, i.e, LaBSE[30] (Feng et al., 2022), or specially designed for German, RoBERTa[31] (Conneau et al., 2020),

[30] https://huggingface.co/sentence-transformers/LaBSE

[31] https://huggingface.co/T-Systems-onsite/cross-en-de-

and ii) testing with different threshold values, i.e, 0.7, 0.75, 0.8, 0.9. We got the best precision score (*precision=0.96*) when we used LaBSE at a threshold value of 0.9.

**Vecalign** (Thompson and Koehn, 2020) is a bilingual sentence alignment method that was designed to align sentences in documents of different languages, however, it was also tested in other works on monolingual parallel corpora (e.g., Spring et al. (2022)). It has two main advantages, it can produce $n$:$m$ alignments, and it can work with more than 200 languages (as it uses the LASER[32] (Artetxe and Schwenk, 2019) sentence representation model in the background; which is multilingual). We used this model only for sentence alignment and not document alignment.

**BertAlign** (Liu and Zhu, 2022) is a attempt to allow sentence-transformer-based methods to produce $n$:$m$ alignments. It was tested on Chinese-English parallel corpora and showed promising results. Our adaptation of this method was only by using a dedicated German sentence transformer model in the algorithm procedure. Following its outperforming results in the sentence transformers experiment, we used the LaBSE sentence transformer model in this experiment.

**MASSAlign** (Paetzold et al., 2017) is a Python package which includes an easy-to-use alignment method on the paragraph- and sentence-level by Paetzold and Specia (2016). The method uses a vicinity-driven approach with a similarity matrix based on a TF-IDF model. It is capable of 1:1, 1:$m$, and $n$:1 alignments. Our adaptation to this model are: i) updating from Python 2 to Python 3 ii) making it more language-independent by flexibly adding a stop word list in the required language. We don't use the updated version of MASSAlign with Doc2Vec by Paun (2021) as we only align on the sentence level. In the first experiment, we found out that paragraph alignment is also required for this algorithm.

**CATS** (Štajner et al., 2018) is an alignment method that can also align paragraphs and sentences. CATS align each original sentence with the closest simple sentence by calculating the similarity of all of them based on n-grams (option: C3G) or word vectors (option: CWASA and WAVG). In our experiments CATS only aligned pairs of type 1:1. Officially the code was published in Java[33], for better integrity with the other alignment methods, we used the existing Python version of it[34]. We did experiments with all three options, for the word vectors we used the German embeddings of fasttext[35] (Athiwaratkun et al., 2018). We just report the best result which was achieved with C3G.

# I Train, Dev, Test Split for Simplification

Table 14 provides the size of train, development and test set of DEplain.

| | document-level | | | sentence-level | | |
|---|---|---|---|---|---|---|
| | WEB | APA | APA+WEB | WEB | APA | APA+WEB |
| train | 481 | 387 | 868 | 1281 | 10660 | 11941<br>(10660+1281) |
| dev | 122 | 48 | 170 | 313 | 1231 | 1544<br>(1231+313) |
| test | 147 | 48 | - | 1846 | 1231 | - |

Table 14: Overview of train/dev/test split.

# J Further Results for Simplification.

We present results of our models trained on DEplain on existing test sets for German text simplification. In subsection J.1, results are shown regarding document simplification and, in subsection J.2, regarding sentence simplification.

## J.1 Results on Document Simplification.

Table 15 shows results of our document-level TS experiments trained on different parts of DEplain using long-mBART with vocabulary reduced to 35k tokens. *APA* correspond to DEplain-apa and *web* to DEplain-web. For a better comparison, we also add the results of a baseline model (last part) and a comparable model reported in Rios et al. (2021) (first part, numbers are copied from them).

| train data | n | SARI | BLEU | BS-P | FRE |
|---|---|---|---|---|---|
| 20min | 18305 | **33.29** | **6.29** | | |
| DEplain-APA | 387 | 22.805 | 1.706 | 0.03 | 63.9 |
| DEplain-web | 481 | 27.113 | 1.81 | 0.007 | 63.5 |
| DEplain-APA+web | 868 | 24.265 | 1.804 | 0.029 | 64 |
| src2src | | 1.953 | 2.051 | 0.029 | 54.45 |

Table 15: Results on Document Simplification Testing on 20min with long-mBART

roberta-sentence-transformer

[32] https://github.com/facebookresearch/LASER

[33] https://github.com/neosyon/SimpTextAlign

[34] https://github.com/kostrzmar/SimpTextAlignPython

[35] https://fasttext.cc/docs/en/crawl-vectors.html

### J.2 Results on Sentence Simplification

In this section, we present results on existing test sets, i.e., *ZEST* (Mallinson et al., 2020) (see Table 16), *APA-LHA C2-A2* (Spring et al., 2021) (see Table 17), *APA-LHA C2-B1* (Spring et al., 2021) (see Table 18), and *TCDE19* (Naderi et al., 2019) (see Table 19).

| | SARI | BLEU | BS-P | FRE |
|---|---|---|---|---|
| ZEST | 39.09 | 56.68 | - | - |
| U-NMT | 35.22 | 52.02 | - | - |
| U-SIMP | 40.0 | 61.1 | - | - |
| mBART-APA | **45.81** | **56.802** | 0.769 | 67.282 |
| mBART-APA+web | 44.913 | 54.718 | 0.778 | 66.588 |
| src2src | 26.812 | 67.116 | 0.856 | 61.5 |

Table 16: Results on the test set of *ZEST*.

| | SARI | BLEU | BS-P | FRE |
|---|---|---|---|---|
| Sockeye | **42.04** | **15.2** | - | - |
| mBART-APA | 27.987 | 5.294 | 0.232 | 57.865 |
| mBART-APA+web | 28.468 | 5.464 | 0.236 | 56.969 |
| src2src | 4.092 | 3.635 | 0.184 | 44.9 |

Table 17: Results on the test set of *APA-LHA C2-A2*.

| | SARI | BLEU | BS-P | FRE |
|---|---|---|---|---|
| Sockeye | **40.73** | **12.3** | - | - |
| mBART-APA | 29.086 | 6.495 | 0.272 | 57.299 |
| mBART-APA+web | 28.527 | 6.604 | 0.273 | 56.848 |
| src2src | 5.325 | 6.18 | 0.236 | 44.9 |

Table 18: Results on the test set of *APA-LHA C2-B1*.

In each of the tables, the first part includes the results of other models trained on other data than DEPLAIN, the middle part includes the results of our models trained on DEPLAIN, and the last part is the result of the baseline model. The scores of the other models are extracted from the corresponding papers, we do not calculate them ourselves as the model checkpoints or model predictions are not available. Hence, scores of some metrics are missing if they were not reported, e.g., FRE or BERTScore. Furthermore, different implementations of the metrics might be used, therefore the scores should be interpreted with caution.

| | SARI | BLEU | BS-P | FRE |
|---|---|---|---|---|
| ZEST | **41.12** | **21.11** | - | - |
| U-NMT | 35.97 | 11.72 | - | - |
| U-SIMP | 37.4 | 15.03 | - | - |
| mBART-APA | 38.964 | 16.85 | 0.539 | 44.85 |
| mBART-APA+web | 36.937 | 16.321 | 0.542 | 43.65 |
| src2src | 14.999 | 27.348 | 0.546 | 28.1 |

Table 19: Results on the test set of *TCDE19*. The scores of the other models are copied from Mallinson et al. (2020).

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Not numbered, after the conclusion called "limitations"*

☑ A2. Did you discuss any potential risks of your work?
*not numbered, Ethics & Impact Statement*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*Abstract and 1 introduction*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☑ Did you use or create scientific artifacts?

*creation: 3, 4, 5, usage: 6.*

☑ B1. Did you cite the creators of artifacts you used?
*6: test sets and appendix: overview*

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*Not numbered, Ethics & Impact Statement*

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*section 6: test sets for simplification. Not numbered: Ethics & Impact Statement*

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*Details on DEplain-web*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*3, 4, 5 and appendix: Details on DEplain-web*

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Table 1, plus Tables in appendix. Also in 5.*

### C ☑ Did you run computational experiments?

*5, 6*

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Ethics & Impact Statement, appendix*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*6, and Description of Adaptations of Alignment Methods*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*6*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*4, 5, 6. The citations of the alignment methods can be found in the appendix (with references to the implementations).*

**D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?**
*4*

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*If we would publish them, we would hurt our anonymity. We will provide the annotation schema upon acceptance.*

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*4*

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*4 (part of annotation schema)*

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*No ethic board required.*

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*4*