A Corpus for Automatic Readability Assessment and Text Simplification of German

Battisti, Alessia ; Pfütze, Dominik ; Säuberli, Andreas ; Kostrzewa, Marek ; Ebling, Sarah

Abstract: In this paper, we present a corpus for use in automatic readability assessment and automatic text simplification for German. The corpus is compiled from web sources and consists of parallel as well as monolingual-only (simplified German) data amounting to approximately 6,200 documents (nearly 211,000 sentences). As a unique feature, the corpus contains information on text structure (e.g., paragraphs, lines), typography (e.g., font type, font style), and images (content, position, and dimensions). While the importance of considering such information in machine learning tasks involving simplified language, such as readability assessment, has repeatedly been stressed in the literature, we provide empirical evidence for its benefit. We also demonstrate the added value of leveraging monolingual-only data for automatic text simplification via machine translation through applying back-translation, a data augmentation technique.

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3302–3311
Marseille, 11–16 May 2020

A Corpus for Automatic Readability Assessment and Text Simplification of German

Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, Sarah Ebling
Institute of Computational Linguistics, University of Zurich
Andreasstrasse 15, 8050 Zurich, Switzerland
{alessia.battisti, dominik.pfuetze, andreas.saeuberli}@uzh.ch, marek_kostrzewa@yahoo.co.uk, ebling@cl.uzh.ch

Abstract

In this paper, we present a corpus for use in automatic readability assessment and automatic text simplification for German. The corpus is compiled from web sources and consists of parallel as well as monolingual-only (simplified German) data amounting to approximately 6,200 documents (nearly 211,000 sentences). As a unique feature, the corpus contains information on text structure (e.g., paragraphs, lines), typography (e.g., font type, font style), and images (content, position, and dimensions). While the importance of considering such information in machine learning tasks involving simplified language, such as readability assessment, has repeatedly been stressed in the literature, we provide empirical evidence for its benefit. We also demonstrate the added value of leveraging monolingual-only data for automatic text simplification via machine translation through applying back-translation, a data augmentation technique.

Keywords: Simplified German, automatic readability assessment, automatic text simplification, multimodal

1. Introduction

Simplified language is a variety of standard language characterized by reduced lexical and syntactic complexity, the addition of explanations for difficult concepts, and clearly structured layout.[1] Among the target groups of simplified language commonly mentioned are persons with cognitive impairment and learning disabilities, prelingually deaf persons, functionally illiterate persons, and foreign language learners (Bredel and Maaß, 2016).

Two natural language processing tasks deal with the concept of simplified language: automatic readability assessment and automatic text simplification. Readability assessment refers to the process of determining the level of difficulty of a text, e.g., along readability measures, school grades, or levels of the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2009). Readability measures, in their traditional form, take into account only surface features. For example, the Flesch Reading Ease Score (Flesch, 1948) measures the length of words (in syllables) and sentences (in words). While readability has been shown to correlate with such features to some extent (Just and Carpenter, 1980), a consensus has emerged according to which they are not sufficient to account for all of the complexity inherent in a text. As Kauchak et al. (2014, p. 2618) state, "the usability of readability formulas is limited and there is little evidence that the output of these tools directly results in improved understanding by readers". As a result, more sophisticated models employing (deeper) linguistic knowledge such as lexical, semantic, morphological, morphosyntactic, syntactic, pragmatic, discourse, psycholinguistic, and language model features have been proposed (Collins-Thompson, 2014; Heimann Mühlenbock, 2013; Pitler and Nenkova, 2008; Schwarm and Ostendorf, 2005; Tanaka et al., 2013).

Automatic text simplification was initiated in the late 1990s (Carroll et al., 1998; Chandrasekar et al., 1996) and since then has been approached by means of rule-based and statistical methods. As part of a rule-based approach, the operations carried out typically include replacing complex lexical and syntactic units by simpler ones. A statistical approach generally conceptualizes the simplification task as one of converting a standard-language into a simplified-language text using machine translation. Nisioi et al. (2017) introduced neural machine translation to automatic text simplification.

Research on automatic text simplification is comparatively widespread for languages such as English, Swedish, Spanish, and Brazilian Portuguese. To the authors' knowledge, no productive system exists for German. Suter (2015), Suter et al. (2016) presented a prototype of a rule-based system for German.

Machine learning approaches to both readability assessment and text simplification rely on data systematically prepared in the form of corpora. Specifically, for automatic text simplification via machine translation, pairs of standard-language/simplified-language texts aligned at the sentence level (i.e., parallel corpora) are needed. The paper at hand introduces a corpus developed for use in automatic readability assessment and automatic text simplification of German. The focus of this publication is on representing information that is valuable for these tasks but that hitherto has largely been ignored in machine learning approaches centering around simplified language, specifically, information on text structure (e.g., paragraphs, lines), typography (e.g., font type, font style), and images (content, position, and dimensions). The importance of considering such information has repeatedly been asserted theoretically (Arfé et al., 2018; Bock, 2018; Bredel and Maaß, 2016).

The remainder of this paper is structured as follows: Section 2 presents previous corpora used for automatic readability assessment and text simplification. Section 3 describes our corpus, introducing its novel aspects and presenting the primary data (Section 3.1), the metadata (Sec-

[1] The term *plain language* is avoided, as it refers to a specific level of simplification. *Simplified language* subsumes all efforts of reducing the complexity of a piece of text.

tion 3.2), the secondary data (Section 3.3), the corpus profile (Section 3.4), the sentence alignment procedure (Section 3.5), and the results of machine learning experiments carried out on the corpus (Section 3.6).

2. Previous Corpora for Automatic Readability Assessment and Automatic Text Simplification

A number of corpora have been created for use in automatic readability assessment and automatic text simplification. The most well-known example is the Parallel Wikipedia Simplification Corpus (PWKP) compiled from parallel articles of the English Wikipedia and Simple English Wikipedia (Zhu et al., 2010) and consisting of around 108,000 sentence pairs. The corpus profile is shown in Table 1. While the corpus represents the largest dataset involving simplified language to date, its application has been criticized for various reasons (Amancio and Specia, 2014; Xu et al., 2015; Štajner et al., 2018); among these, the fact that Simple English Wikipedia articles are not necessarily direct translations of articles from the English Wikipedia stands out. Hwang et al. (2015) provided an updated version of the corpus that includes a total of 280,000 full and partial matches between the two Wikipedia versions.

Another frequently used data collection for English is the Newsela Corpus (Xu et al., 2015) consisting of 1,130 news articles, each simplified into four school grade levels by professional editors. Table 2 shows the profile of the Newsela Corpus. Together, Tables 1 and 2 show that the difference in vocabulary size between the English and the simplified English side of the PWKP Corpus amounts to only 18%, while the corresponding number for the English side and the level representing the highest amount of simplification in the Newsela Corpus (Simple-4) is 50.8%.[2].

Gasperin et al. (2010) compiled the PorSimples Corpus consisting of Brazilian Portuguese texts (2,116 sentences), each with a natural and a strong simplification, resulting in around 4,500 aligned sentences. Drndarević and Saggion (2012), Bott et al. (2012), Bott and Saggion (2012) produced the Simplext Corpus consisting of 200 Spanish/simplified Spanish document pairs, amounting to a total of 1,149 (Spanish) and 1,808 (simplified Spanish) sentences (approximately 1,000 aligned sentences).

Klaper et al. (2013) created the first parallel corpus for German/simplified German, consisting of 256 parallel texts downloaded from the web (approximately 70,000 tokens).

3. Building a Corpus for Automatic Readability Assessment and Automatic Text Simplification of German

Section 2 demonstrated that the only corpus containing simplified German currently in existence is that of Klaper et al. (2013). Since its creation, a number of legal and political developments have spurred the availability of data in simplified German. Among these developments is the introduction of a set of regulations for accessible information technology (*Barrierefreie-Informationstechnik-Verordnung, BITV 2.0*) in Germany, the approval of rules for accessible information and communication (*Barrierefreie Information und Kommunikation, BIK*) in Austria, and the ratification of the United Nations Convention on the Rights of Persons with Disabilities (CRPD) in Switzerland. The paper at hand introduces a corpus that represents an enhancement of the corpus of Klaper et al. (2013) in the following ways:

- The corpus contains more parallel data.
- The corpus newly contains monolingual-only data (simplified German).
- The corpus newly contains information on text structure, typography, and images.

The simplified German side of the parallel data of the corpus together with the monolingual-only data can be used for automatic readability assessment (cf. Section 3.6). The parallel data in the corpus is useful both for deriving rules for a rule-based text simplification system in a data-driven manner and for training a data-driven machine translation system (cf. Section 3.6.2). A data augmentation technique such as back-translation (Sennrich et al., 2016) can be applied to the monolingual-only data to arrive at additional (synthetic) parallel data.

3.1 Primary Data

The corpus contains PDFs and webpages collected from web sources in Germany, Austria, and Switzerland at the end of 2018/beginning of 2019. The web sources mostly represent websites of governments, specialised institutions, and non-profit organisations (92 different domains). The documents cover a range of topics, such as politics (e.g., instructions for voting), health (e.g., what to do in case of pregnancy), and culture (e.g., introduction to art museums). For the webpages, a static dump of all documents was created. Following this, the documents were manually checked to verify content and language. The main content was subsequently extracted, i.e., HTML markup and boilerplate removed using the Beautiful Soup library for Python.[3] Information on text structure (e.g., paragraphs, lines) and typography (e.g., boldface, italics) was retained. Similarly, image information (content, position, and dimensions of an image) was preserved.

For PDFs, the PDFlib Text and Image Extraction Toolkit (TET) was used to extract the plain text and record information on text structure, typography, and images.[4] The toolkit produces output in XML format (TETML).

3.2 Metadata

Metadata was collected automatically from the HTML (webpages) and TETML (PDFs) files, verified, manually complemented, and recorded in the Open Language Archives Community (OLAC) Standard.[5] OLAC is based

[2] Vocabulary size as an indicator of lexical richness is generally taken to correlate positively with complexity (Vajjala and Meurers, 2012)

[3] https://pypi.org/project/beautifulsoup4/ (last accessed: November 21, 2019)

[4] https://www.pdflib.com/ (last accessed: November 21, 2019)

[5] http://www.language-archives.org/OLAC/olacms.html (last accessed: November 21, 2019)

	English	Simple English
Number of sentences	108,016	114,924
Number of tokens	2,645,771	2,175,240
Avg. no. of words per sentence	24.49	18.93
Vocabulary size	95,111	78,009

Table 1: Parallel Wikipedia Simplification Corpus (PWKP) (Zhu et al., 2010): Profile (from Xu et al., 2015)

	Original	Simple-1	Simple-2	Simple-3	Simple-4
Number of sentences	56,037	57,940	63,419	64,035	64,162
Number of tokens	1,301,767	1,126,148	1,052,915	903,417	764,103
Avg. no. of sentences per document	49.59	51.27	56.12	56.67	56.78
Avg. no. of words per document	1,152.01	996.59	931.78	799.48	676.2
Avg. no. of words per sentence	23.23	19.44	16.6	14.11	11.91
Vocabulary size	39,046				19,197

Table 2: Newsela Corpus (Xu et al., 2015): Profile

on a reduced version of the Dublin Core Metadata Element Set (DCMES).[6] Of the 15 elements of this "Simple Dublin Core" set, the following 12 were actively used along with controlled vocabularies of OLAC and Dublin Core:

- `title`: title of the document, with the language of the document specified as the value of an `xml:lang` attribute and alternatives to the original title (e.g. translations) stored as `dcterms:alternative` (cf. Figure 1 for an example)
- `contributor`: all person entities linked to the creation of the document, with an `olac:code` attribute with values from the OLAC role vocabulary used to further specify the role of the contributor, e.g. `author`, `editor`, `publisher`, or `translator`
- `date`: date mentioned in the metadata of the HTML or PDF document or, for news and blog articles, date mentioned in the body of the text, in W3C date and time format
- `description`: value of the description in the metadata of an HTML document or list of sections of a PDF document, using the Dublin Core qualifier `TableOfContents`
- `format`: distinction between the Internet Media Types (MIME types) `text/html` (for webpages) and `application/pdf` (for PDFs)
- `identifier`: URL of the document or International Standard Book Number (ISBN) for books or brochures
- `language`: language of the document as value of the attribute `olac:code` (i.e., `de`, as conforming to ISO 639), with the CEFR level as optional element content[7]
- `publisher`: organization or person that made the document available
- `relation`: used to establish a link between documents in German and simplified German for the parallel part of the corpus, using the Dublin Core qualifiers `hasVersion` (for the German text) and `isVersionOf` (for the simplified German text)
- `rights`: any piece of information about the rights of a document, as far as available in the source
- `source`: source document, i.e., HTML for web documents and TETML for PDFs
- `type`: nature or genre of the content of the document, which, in accordance with the DCMI Type Vocabulary, is `Text` in all cases and additionally `StillImage` in cases where a document also contains images. Additionally, the linguistic type is specified according to the OLAC Linguistic Data Type Vocabulary, as either `primary_text` (applies to most documents) or `lexicon` in cases where a document represents an entry of a simplified language vocabulary

The elements `coverage` (to denote the spatial or temporal scope of the content of a resource), `creator` (to denote the author of a text, see `contributor` above), and `subject` (to denote the topic of the document content) were not used.

[6] `http://dublincore.org/` (last accessed: November 21, 2019)

[7] *capito*, the largest provider of simplification services for German, recognizes three levels of simplified language corresponding to the CEFR levels A1, A2, and B1; `https://www.capito.eu/` (last accessed: November 26, 2019). Note that while the CEFR was designed to measure foreign language skills, with simplified language it is partly applied in the context of first-language acquisition (Bredel and Maaß, 2016).

```
<?xml version='1.0' encoding='utf-8'?>
    <olac:olac xmlns:cld="http://purl.org/cld/terms/"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:oai="http://www.openarchives.org/OAI/2.0/"
    xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
    xmlns:olac="http://www.language-archives.org/OLAC/1.1/"
    xmlns:schemaLocation="http://www.language-archives.org/OLAC/
    1.1/olac.xsd"
    xmlns:tei="http://www.tei-c.org/ns/1.0"
    xmlns:xs="http://www.w3.org/2001/XMLSchema"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns="http://www.openarchives.org/OAI/2.0/static-repository">
        <dc:title xml:lang="de">Maria sagt es weiter...
        Ein Bilder-Lese-Buch über sexuelle Gewalt und Hilfe holen.
        </dc:title>
        <dc:language xsi:type="olac:language" olac:code="de">A2</dc:language>
        <dc:publisher>Frauenbüro der Stadt Linz</dc:publisher>
        <dc:publisher xsi:type="dcterms:URI">www.linz.at/frauen</dc:publisher>
        <dc:contributor xsi:type="olac:role" olac:code="author">
        Verein Hazissa</dc:contributor>
        <dc:contributor xsi:type="olac:role" olac:code="translator">
        capito Oberösterreich</dc:contributor>
        <dc:contributor xsi:type="olac:role" olac:code="illustrator">
        Müller, Silke</dc:contributor>
        <dc:identifier xsi:type="dcterms:URI">
        https://www.linz.at/images/MariaD.pdf</dc:identifier>
        <dc:date xsi:type="dcterms:W3CDTF">2016</dc:date>
        <dc:format xsi:type="dcterms:IMT">application/pdf</dc:format>
        <dc:type xsi:type="dcterms:DCMIType">Text</dc:type>
        <dc:type xsi:type="dcterms:DCMIType">StillImage</dc:type>
        <dc:type xsi:type="olac:linguistic-type" olac:code="primary_text"/>
        <dc:source>mariad.tetml</dc:source>
        <dc:rights/>
        <dcterms:tableOfContents>
        Maria sagt es weiter Seite 7; Informationen zu sexueller Gewalt Seite 12;
            Adressen von Beratungs-Stellen Seite 17; Wörterbuch Seite 32
        </dcterms:tableOfContents>
    </olac:olac>
```

Figure 1: Sample metadata in OLAC for a PDF document from the corpus

Figure 1 shows an example of OLAC metadata. The source document described with this metadata record is a PDF structured into chapters, with text corresponding to the CEFR level A2 and images.

Metadata in OLAC may be converted into the Component MetaData Infrastructure (CMDI) standard in a straightforward manner. CMDI is the supported metadata version of CLARIN, a European research infrastructure for language resources and technology.[8]

Information on the language level of a simplified German text (typically A1, A2, or B1) is particularly valuable, as it allows for conducting automatic readability assessment and graded automatic text simplification experiments on the data. 52 websites and 233 PDFs (amounting to approximately 26,000 sentences) have an explicit language level label.

3.3 Secondary Data

Annotations were added in the Text Corpus Format by WebLicht (TCF) developed as part of CLARIN.[9] TCF supports standoff annotation, which allows for representation of annotations with conflicting hierarchies. The format does not assign a separate file for each annotation layer; instead, the source text and all annotation layers are stored jointly in a single file. A token layer acts as the key element

[8] https://www.clarin.eu/ (last accessed: November 21, 2019)

[9] https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format (last accessed: November 21, 2019)

```
<TextCorpus>
    <text>...</text>
    <tokens>
        <token ID="t_0" font="F0">Vorwort</token>
        <token ID="t_1" font="F0">Liebe</token>
        <token ID="t_2" font="F0">Leserinnen</token>
        <token ID="t_3" font="F0">!</token>
        ...
    </tokens>
    <sentences>
        <sentence ID="s_0" tokenIDs="t_0"/>
        <sentence ID="s_1" tokenIDs="t_1 t_2 t_3"/>
        ...
    </sentences>
    <textstructure>
        <textspan start="t_0" type="paragraph" end="t_0"/>
        <textspan start="t_0" type="line" end="t_0"/>
        <textspan type="paragraph" start="t_1" end="t_3"/>
        <textspan type="line" start="t_1" end="t_3"/>
        <textspan type="paragraph" start="t_4" end="t_27"/>
        ...
    </textstructure>
    <lemmas>
        <lemma ID="l_0" tokenIDs="t_0">Vorwort</lemma>
        <lemma ID="l_1" tokenIDs="t_1">lieb</lemma>
        ...
    </lemmas>
    <POStags tagset="stts">
        <tag tokenIDs="t_0">NN</tag>
        <tag tokenIDs="t_1">ADJA</tag>
        ...
    </POStags>
    <morphology>
        <analysis tokenIDs="t_0">
            <tag>
                <fs>
                    <f name="category">Common name</f>
                    <f name="gender">Neuter</f>
                    <f>_</f>
                    <f name="number">Singular</f>
                </fs>
            </tag>
        </analysis>
        ...
    </morphology>
    <depparsing>
        ...
    </depparsing>
    <images>
        <image ID="I0" page="1" x="-1.07" y="112.47" width="597.12" height
           ="730.56"/>
        ...
    </images>
    <fonts>
        <font id="F0" name="TradeGothic-BoldTwo" fullname="UDSPGZ+TradeGothic-
           BoldTwo" type="Type_1_CFF" embedded="true" ascender="977" capheight
           ="722" italicangle="0" descender="-229" weight="700" xheight="520"/>
        ...
    </fonts>
    </TextCorpus>
```

Figure 2: Sample corpus annotation

to which all other annotation layers are linked.
The following types of annotations were added: text structure, tokens, parts of speech, morphological units, lemmas, sentences, dependency parses, images, and fonts. TCF does not readily accommodate the incorporation of all of these types of information. We therefore extended the format in the following ways:

- Information on the font type and font style (e.g., italics, bold print) of a token and its position on the physical page (for PDFs only) was specified as attributes to the `token` elements of the `tokens` layer (cf. Figure 2 for an example).
- Information on physical page segmentation (for PDFs only), paragraph segmentation, and line segmentation was added as part of a `textspan` element in the `textstructure` layer (cf. Figure 2 for an example).
- A separate `images` layer was introduced to hold `image` elements that take as attributes the x and y coordinates of the images, their dimensions (width and height), and the number of the page on which they occur (cf. Figure 2 for an example).
- A separate `fonts` layer was introduced to preserve detailed information on the font configurations referenced in the `tokens` layer (cf. Figure 2 for an example).

Linguistic annotation was added automatically using the ParZu dependency parser for German (Sennrich et al., 2009) (for tokens and dependency parses), the NLTK toolkit (Bird et al., 2009) (for sentences), the TreeTagger using the pretrained model for German (Schmid, 1995) (for part-of-speech tags and lemmas), and Zmorge (Sennrich and Kunz, 2014) (for morphological units).
Figure 2 shows a sample corpus annotation. Together, the metadata shown in Figure 1 and the annotations presented in Figure 2 constitute a complete TCF file.

3.4 Corpus Profile

Table 3 shows the profile of the resulting corpus. The corpus contains 6,217 documents (5,461 monolingual documents plus 378 documents for each side of the parallel data). The monolingual-only documents on average contain fewer sentences than the simplified German side of the parallel data (average document length in sentences 31.64 vs. 55.75). The average sentence length is almost equal (approx. 11 tokens). Hence, the monolingual-only texts are shorter than the simplified German texts in the parallel data. Compared to their German counterparts, the simplified German texts in the parallel data have clearly undergone a process of lexical simplification: The vocabulary is smaller by 51% (33,384 vs. 16,352 types), which is comparable to the rate of reduction reported in Section 2 for the Newsela Corpus (50.8%).

3.5 Sentence Alignment

Automatic text simplification approaches that are based on machine translation (cf. Section 1) currently rely on data in the form of sentence alignments. Two freely available tools exist specifically for generating sentence alignments of standard-language/simplified-language document pairs: Customized Alignment for Text Simplification (CATS) (Štajner et al., 2018) and MASSAlign (Paetzold et al., 2017). Preliminary investigation for our data showed the performance of CATS to be superior to that of MASSAlign. Therefore, we used CATS to align the sentences of the parallel part of the corpus. CATS aligns units of types paragraph and sentence. The tool requires a number of parameters to be specified:

- **Similarity strategy:** CATS offers a lexical (character-n-gram-based, CNG) and two semantic similarity strategies. With regard to CNG, Štajner et al. (2018) found trigrams (C3G) to perform best for aligning English and Spanish data. The two semantic similarity strategies, WAVG (Word Average) and CWASA (Continuous Word Alignment-based Similarity Analysis) (Franco-Salvador et al., 2016), both require pretrained word embeddings. WAVG averages the word vectors of a paragraph or sentence to obtain the final vector for the respective text unit. CWASA is based on the alignment of continuous words using directed edges.
- **Alignment level:** CATS is capable of aligning paragraphs only, paragraphs followed by sentences (two-step alignment), or sentences only.
- **Alignment strategy:** CATS allows for adhering to a monotonicity restriction, i.e., requiring the order of information to be identical on the standard-language and simplified-language side, or abandoning it.

Table 4 shows the performance of CATS on the parallel part of our corpus for all three available similarity strategies (C3G, CWASA, WAVG). For CWASA and WAVG, we experimented with three different embedding algorithms: word2vec (Mikolov et al., 2013), gloVe (Pennington et al., 2014), and fastText (Bojanowski et al., 2016). For gloVe, we trained word vectors on data from public sources such as the Europarl corpus, the UN corpus and the News Commentary Corpus.[10] For word2vec, we use readily available word vectors trained on German Wikipedia and German news articles (May 15, 2015) (Mueller, 2015).[11] FastText has its own database of pretrained word vectors in 157 languages, derived from Wikipedia and Common Crawl (Grave et al., 2018).[12]
We tested two alignment levels: aligning sentences directly (column "sentence" in Table 4) and pararaph alignment followed by sentence alignment (column "paragraph+sentence" in Table 4).

[10] `http://data.statmt.org/wmt17/translation-task/preprocessed/de-en/` (last accessed: November 25, 2019)

[11] `https://devmount.github.io/GermanWordEmbeddings/` (last accessed: November 25, 2019)

[12] `https://fasttext.cc/docs/en/crawl-vectors.html` (last accessed: November 25, 2019)

	German	Simplified German
Monolingual		
Number of documents		5,461
Number of sentences		172,773
Number of tokens		1,916,045
Avg. no. of sentences per document		31.64
Avg. no. of tokens per sentence		11.09
Parallel		
Number of documents	378	378
Number of sentences	17,121	21,072
Number of tokens	347,941	246,405
Avg. no. of sentences per document	45.29	55.75
Avg. no. of tokens per sentence	20.32	11.69
Vocabulary size	33,384	16,352
Parallel (total)		
Number of documents		756
Number of sentences		38,193
Number of tokens		594,346
Avg. no. of sentences per document		50.52
Avg. no. of tokens per sentence		15.56
Monolingual and parallel (total)		
Number of documents		6,217
Number of sentences		210,966
Number of tokens		2,510,391
Avg. no. of sentences per document		33.93
Avg. no. of tokens per sentence		11.90

Table 3: Corpus profile

Embedding type	**Size (in GB)**	**Similarity strategy**	**Cosine similarity**			
			sentence		paragraph+sentence	
			mean	SD	mean	SD
-	-	C3G	0.371	0.284	0.319	0.296
gloVe	0.48	CWASA	0.568	0.199	0.529	0.201
		WAVG	0.843	0.149	0.829	0.142
word2vec	1.7	CWASA	0.589	0.259	0.510	0.251
		WAVG	0.714	0.244	0.677	0.238
fastText	4.5	CWASA	0.646	0.177	0.607	0.179
		WAVG	0.853	0.142	0.827	0.137

Table 4: CATS sentence alignment: Scores (SD: standard deviation)

As our alignment strategy, we dismissed the monotonicity restriction due to our observation that the order of information in a simplified-language text is not always preserved compared to that of the corresponding standard-language text.

Table 4 displays the mean cosine similarity scores produced by CATS across all sentence and paragraph+sentence pairs, respectively. Human evaluation is needed to determine the best-performing configuration.

3.6 Empirical Validation of the Corpus

3.6.1 Cluster Analysis

Battisti (2019) and Battisti et al. (2019) applied unsupervised machine learning techniques to the simplified German texts of the corpus presented in this paper (i.e., the simplified German side of the parallel data along with the monolingual-only data) with the aim of investigating evidence of multiple complexity levels. The authors found features based on text structure (e.g., number of paragraphs, number of lines, adherence to a one-sentence-per-line rule),

typography (number of words of a specific font type), and images (number of images) to be predictive of the level of difficulty of a simplified German text. To our knowledge, this is the first study to deliver empirical proof of the relevance of such features.

3.6.2 Machine Translation

To empirically assess the benefit of the monolingual-only data in our corpus for automatic text simplification via machine translation, we trained a baseline neural sequence-to-sequence model, i.e., recurrent neural networks (RNNs) with attention (Bahdanau et al., 2015), on the aligned sentences of our corpus (cf. Section 3.5, using WAVG as similarity strategy with fastText embeddings and sentence-only as alignment level). We then used back-translation (Sennrich et al., 2016) to generate additional (synthetic) parallel data from the monolingual sentences in simplified German. In our first experiments, even with very low baseline translation quality, adding a sample of these synthetic parallel pairs to the existing parallel data in a 1:1 ratio resulted in an improvement of 2.6 BLEU.[13] The detailed results will be reported in a separate publication.

We attribute the low overall performance of the machine translation experiments to the nature of the data. Many of the simplified German documents in the parallel data of our corpus contain extensive text elaboration, sentence splitting, and reordering, which renders alignment at the sentence level difficult. Therefore, one of our next steps will be to look at ways to tackle text simplification at the document level. Nevertheless, the results presented in this section confirm that there is a substantial advantage to leveraging additional monolingual-only data for the translation task.

4. Conclusion and Outlook

We have introduced a corpus compiled for use in automatic readability assessment and automatic text simplification of German. While these tasks have been addressed for other languages, research on German is still scarce. The features exploited as part of machine learning approaches to readability assessment so far typically include surface and/or deeper linguistic features. The corpus presented in this paper additionally contains information on text structure, typography, and images. These features have been shown to be indicative of simple vs. complex texts both theoretically and, using the corpus described in this paper, empirically.

In this paper, we have also demonstrated the added value of considering monolingual-only data for the translation task. As a next step, we will look into whether using embeddings based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) in CATS will yield an improvement for the sentence alignment task, as determined via human evaluation. As part of a different line of research, we will investigate document-level machine translation using the corpus at hand. In addition, we will look into exploiting the image information present in the corpus for the translation task.

For researchers interested in the corpus, we will provide the Python scripts that download the data from the web, process it and produce an HTML or a TETML file as well as TCF and plain text files. Additionally, we can provide sentence alignments derived with CATS.

[13] Note that SARI (Xu et al., 2016), an evaluation metric developed specifically for assessing the output of automatic text simplification systems, requires more than one reference translation; hence, we were unable to apply it.

5. Bibliographical References

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

Arfé, B., Mason, L., and Fajardo, I. (2018). Simplifying informational text structure for struggling readers. *Reading and Writing*, 31(9):2191–2210.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, May 7-9, 2015, Conference Track Proceedings*, San Diego, CA, USA.

Battisti, A., Ebling, S., and Volk, M. (2019). An Empirical Analysis of Linguistic, Typographic, and Structural Features in Simplified German Texts. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*.

Battisti, A. (2019). Automatic Cluster Analysis of Texts in Simplified German. Master's thesis, University of Zurich.

Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Bock, B. M. (2018). "Leichte Sprache" – Kein Regelwerk. Sprachwissenschaftliche Ergebnisse und Praxisempfehlungen aus dem LeiSA-Projekt. Technical report, Universität Leipzig.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Bott, S. and Saggion, H. (2012). Automatic Simplification of Spanish Text for e-Accessibility. In *Proceedings of the 13th International Conference on Computers Helping People with Special Needs (ICCHP)*, pages 527–534, Linz, Austria.

Bott, S., Saggion, H., and Figueroa, D. (2012). A Hybrid System for Spanish Text Simplification. In *Proceedings of the Third Workshop on Speech and Language Processing for Assistive Technologies*, pages 75–84, Montréal, Canada.

Bredel, U. and Maaß, C. (2016). *Leichte Sprache: Theoretische Grundlagen. Orientierung für die Praxis*. Duden, Berlin.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., and Tait, J. (1998). Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of the AAAI'98 Workshop on Integrating AI and Assistive Technology*, pages 7–10.

Chandrasekar, R., Doran, C., and Srinivas, B. (1996). Motivations and Methods for Text Simplification. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 1041–1044, Copenhagen, Denmark.

Collins-Thompson, K. (2014). Computational Assessment of Text Readability. A Survey of Current and Future Research. *ITL International Journal of Applied Linguistics*, 165(2):97–135.

Council of Europe. (2009). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Drndarević, B. and Saggion, H. (2012). Towards Automatic Lexical Simplification in Spanish: An Empirical Study. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 8–16, Montréal, Canada.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.

Franco-Salvador, M., Gupta, P., Rosso, P., and Banchs, R. E. (2016). Cross-language plagiarism detection over continuous-space-and knowledge graph-based representations of language. *Knowledge-based systems*, 111:87–99.

Gasperin, C., Maziero, E., and Aluisio, S. M. (2010). Challenging Choices for Text Simplification. In *Computational Processing of the Portuguese Language. Proceedings of the 9th International Conference, PROPOR 2010*, pages 40–50, Porto Alegre, Brazil.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Heimann Mühlenbock, K. (2013). *I See What You Mean: Assessing Readability for Specific Target Groups*. Ph.D. thesis, University of Gothenburg.

Hwang, W., Hajishirzi, H., Ostendorf, M., and Wu, W. (2015). Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of NAACL-HLT*, pages 211–217.

Just, M. and Carpenter, P. (1980). A theory of reading. from eye fixations to comprehension. *Psychological review*, 87(4):329–354.

Kauchak, D., Mouradi, O., Pentoney, C., and Leroy, G. (2014). Text Simplification Tools: Using Machine Learning to Discover Features that Identify Difficult Text. In *Proceedings of the 47th Hawaii International Conference on System Sciences*, pages 2616–2625.

Klaper, D., Ebling, S., and Volk, M. (2013). Building a German/Simple German Parallel Corpus for Automatic Text Simplification. In *ACL Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mueller, A. (2015). Analyse von Wort-Vektoren deutscher Textkorpora (Bachelor thesis). *http://devmount.github.io/GermanWordEmbeddings/*.

Nisioi, S., Štajner, S., Ponzetto, S. P., and Dinu, L. P. (2017). Exploring Neural Text Simplification Models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 85–91, Vancouver, Canada, July.

Paetzold, G., Alva-Manchego, F., and Specia, L. (2017). Massalign: Alignment and annotation of comparable documents. In Seong-Bae Park et al., editors, *Proceedings of the IJCNLP 2017, Tapei, Taiwan, November 27 - December 1, 2017, System Demonstrations*, pages 1–4. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pitler, E. and Nenkova, A. (2008). Revisiting Readability: A Unified Framework for Predicting Text Quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*, pages 186–195, Honolulu, Hawaii.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the EACL'95 SIGDAT Workshop*, pages 47–50, Dublin, Ireland.

Schwarm, S. and Ostendorf, M. (2005). Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *Proceedings of the 43rd Annual meeting of the Association for Computational Linguistics*, pages 523–530.

Sennrich, R. and Kunz, B. (2014). Zmorge: A German Morphological Lexicon Extracted from Wiktionary. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*.

Sennrich, R., Schneider, G., Volk, M., and Warin, M. (2009). A New Hybrid Dependency Parser for German. *Proceedings of the German Society for Computational Linguistics and Language Technology Conference*, pages 115–124.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, aug. Association for Computational Linguistics.

Štajner, S., Franco-Salvador, M., Rosso, P., and Ponzetto, S. (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)*, pages 3895–3903, Miyazaki, Japan.

Suter, J., Ebling, S., and Volk, M. (2016). Rule-based Automatic Text Simplification for German. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 279–287, Bochum, Germany.

Suter, J. (2015). Rule-based Text Simplification for Ger-

man. Bachelor's thesis, University of Zurich.

Tanaka, S., Jatowt, A., Kato, M., and Tanaka, K. (2013). Estimating content concreteness for finding comprehensible documents. In *Proceedings of the 6th ACM international conference on Web search and data mining*, pages 475–484.

Vajjala, S. and Meurers, D. (2012). On Improving the Accuracy of Readability Classification Using Insights from Second Language Acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP (NAACL HLT '12)*, pages 163–173, Montreal, Canada.

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

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4(401–415).

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of the International Conference on Computational Linguistics*, pages 1353–1361, Beijing, China.