Data and Approaches for German Text simplification – towards an Accessibility-enhanced Communication

Thorben Schomacker, Michael Gille,
Jörg von der Hülls and **Marina Tropmann-Frick**
Hamburg University of Applied Sciences
thorben.schomacker@haw-hamburg.de
michael.gille@haw-hamburg.de
joerg.vonderhuells@haw-hamburg.de
marina.tropmann-frick@haw-hamburg.de

Abstract

This paper examines the current state-of-the-art of German text simplification, focusing on parallel and monolingual German corpora. It reviews neural language models for simplifying German texts and assesses their suitability for legal texts and accessibility requirements. Our findings highlight the need for additional training data and more appropriate approaches that consider the specific linguistic characteristics of German, as well as the importance of the needs and preferences of target groups with cognitive or language impairments. The authors launched the interdisciplinary OPEN-LS [1] project in April 2023 to address these research gaps. The project aims to develop a framework for text formats tailored to individuals with low literacy levels, integrate legal texts, and enhance comprehensibility for those with linguistic or cognitive impairments. It will also explore cost-effective ways to enhance the data with audience-specific illustrations using image-generating AI.

1 Introduction

In German-speaking countries, the majority of the population uses everyday language (Alltagssprache) in their daily affairs, with slight regional variations. However, in written texts, a more standardized vocabulary but with similar complexity (Bredel and Maaß, 2016) is typically preferred. In contrast, 12% of the German population faces challenges in comprehending and utilizing standard language due to reduced literacy (Grotlüschen and Buddeberg, 2020). For more accessible and inclusive communication, this group depends on comprehensibility-enhanced language. Currently, specialized human translators convert standard language texts into simplified versions including easy language, with legal texts posing a particular challenge due to their technical nature and normative subject matter. Technical language texts represent one end of the complexity spectrum and easy language texts the other. This is further amplified by the fact that both text forms are linguistic expressions of constructed languages. To categorize training data effectively, we differentiate between "easy language" (Leichte Sprache) and "simple language" (einfache Sprache). "Easy language" refers to a highly comprehensible and rule-based form of German, whereby "simple language" is used to describe a variety of simplified language versions in the gray area between standard language and easy language (Maaß, 2020). Easy language is roughly equivalent with level A2 of the Common European Framework of Reference for Languages (CEFR). Since public entities in Germany are required by law to translate information and communication texts into an accessible language version (BGG, 2022) the costs of this task burden the public budget. Automated approaches based on machine learning techniques promise to solve many of the challenges of text simplification, including the difficulties caused by technical language. A tool to simplify documents from different domains to a degree that facilitates these texts' comprehensibility for people with language or cognitive disabilities does not only improve understanding of these texts. It is also a key to inclusion and social participation (UN, 2008). This holds especially for domain-specific legal texts that are the starting point for the intralingual translation. In the course of our project we aim to build on existing simplification approaches using NLMs and adjust them with respect to the demands of the application domain. To achieve this objective, two specific aspects must be considered: First, the identification and systematic categorization of training data from the legal domain to build a quality-assured dataset to train a large language model for the domain specific simplification tasks in German. Second, the fine-tuning of an NLM under consideration of target-audience related com-

---

[1] For more and up-to-date information, please visit our project homepage https://open-ls.entavis.com

prehensibility requirements. After a brief review of related work on German datasets and approaches (Section 2), this paper delivers a systematic assessment of published German datasets and approaches against the backdrop of the requirements of participatory communication of legal texts (Section 3). Finally, we outline the ongoing research project, "OPEN-LS: Open Data for Easy Language", which adopts a more target-group oriented approaches. We also identify and address several gaps in the existing research (Section 4).

2 Related Work

Text simplification can be described as a machine translation task, converting one version of a language to another (Standard → Simple). However, compared to other machine translation tasks, automatic text simplification is a relatively new task. It started with a rule based statistical approach in 2010 (Specia, 2010) on a small parallel Portuguese corpus (roughly 4,500 parallel sentences). The first German simplification corpus was introduced in 2012 (Hancke et al., 2012) and consisted of articles from GEO (similar to National Geographic) and GEOlino (GEO's edition for children). In the initial paper, the corpus was only used for the training of statistical classifiers to predict the reading level of German texts. Their corpus was later improved and enlarged (Weiß and Meurers, 2018). In 2016 the first rule-based automatic text simplification system for German was released (Suter et al., 2016). In 2020 the first parallel corpus for data-driven automatic text simplification for German was published (Säuberli et al., 2020) and a first investigation of the use of a neural machine translation system for this problem in German was conducted. They concluded that the Austrian Press Agency corpus was not large enough to sufficiently train a neural machine translation system that produces both adequate and fluent text simplifications. In a later study, the same neural machine translation architecture was use and further evaluated concerning the levels of simplification which were generated by these models (Spring et al., 2021). In 2021 (Rios et al., 2021) adapted mBART (Liu et al., 2020) with Longformer Attention (Beltagy et al., 2020) and applied it to the task of document-level text simplification. It has been further explored on different domains, recently (Schomacker et al., 2023; Stodden et al., 2023). Furthermore, the first Decoder-only approach for German text simplification has been released (Anschütz et al., 2023).

Automatic simplification of legal documents has only recently, in 2022, emerged (Collantes et al., 2015; Cemri et al., 2022; Manor and Li, 2019; Gallegos and George, 2022; Gille et al., 2023; Kopp et al., 2023). All of these works had to rely on monolingual datasets and state, that the task is still underinvestigated. To this day, there is no dataset with parallel legal documents (standard → simple language). In section 5.1 we will further discuss features and constraints of legal texts.

3 Systematic and accessibility-oriented assessment of dataset landscape

3.1 Parallel Datasets

To find all aligned German text simplification datasets, we focused our Google Scholar search on papers which prioritize German by including the word "German" in the title. Further, we wanted to find textual datasets, so used its synonyms: "dataset", "corpus", "data" or "texts". The task of text simplification can be covered by datasets with "simple" language or which investigate "readability" or text "complexity". So, we concluded on this query: `allintitle: German corpus OR dataset OR data OR texts "Simple" OR "simplification" OR "readability" OR "complexity"`. This resulted in an identification of 14 parallel German datasets or sub-datasets as listed in Table 1. By reading the dataset descriptions in their corresponding publication and checking the underlying data sources, we identified inductively text genres and domains in the dataset. We categorized them in three exclusive genres: 1) Encyclopedic (ENC) texts are summaries of knowledge either general or special to a particular field; 2) Articles (ART), are published nonfiction texts; and 3) Unknown (UNK), are texts, of which its author did not provide sufficient information to be clearly categorized. In addition to the genre, we tagged the datasets with seven domains: 1) Medical, which covers all aspects of human health; 2) Disability, which covers all aspects of the life and interests of people with disabilities; 3) News, are texts about current events without defining a field of interest; 4) Politics, discussing topics about politically viewpoints or activities such as electoral programs of political parties; 5) Government, any information, that is published by public authorities and/or containing administrative and non-partisan legal information; 6) Encyclopedic, collection of texts that

could form a reference work without any specific field of interest; 7) Unknown, are texts, of which its author did not provide sufficient information to be clearly categorized. Aligned datasets thematically focused on legal aspects were not identified. We provide an overview of the datasets in Figure 1 by the number of documents. With a percentage of 73% of the documents, News is the largest domain. The more practical and life-oriented categories Government, Disability and Medical are forming together less than 10% of the available data. A significant proportion of 20% of the available simple data is targeted to children. Training machine learning models with children-oriented simple language could lead to a bias. So, this type of data should be used with caution.

3.2 Monolingual Datasets and Sources

To gain a more complete picture of the datasets, we further investigated collections of German easy language, that have no standard language equivalent. Many newspapers or lexicons target children, e.g., "Dein Spiegel" from "Der Spiegel". We decided to only include resources that use simple or easy language and did not research any children-targeted content because children-targeted content does not necessarily mean that it is accessible for the target we defined for simple and easy language. Furthermore, we focused on resources that cover different genres to show the variety of genre currently used in easy language. Many text genres have no published parallel dataset despite the fact, that there are monolingual resources (e.g., narrative texts, legal texts). Similar to the parallel datasets, the majority of texts are news and encyclopedic articles. A comparatively large number of monolingual datasets address the interests of people with disabilities, not least because public authorities in Germany are obliged to communicate in simple and understandable language.

3.3 (Non-)Consideration of Accessibility and Participation in Existing Datasets

For the reasons outlined in Section 1 we focus on two particularly critical dimensions when considering accessibility and participation aspects for the evaluation of existing datasets and approaches: Legal texts and the concrete needs of the addressees.

Legal Texts: Legal language comprises many different types of text such as laws and regulations, court judgements, witness statements, complaints,

legal opinions etc. In addition, a large (and increasing) number of legal sub-domains, e.g., constitutional law, criminal law, AI law, exist. All of these different types of texts in the different (sub-)domains share similar linguistic traits, such as the use of legal jargon ('legalese'), formalization, long and complex sentences, a very high degree of intertextuality, mixed authorship (at least to some degree), a wide range of addressees and a unique tension between accuracy and vagueness (Baumann, 2020). In addition, many legal texts are designed to be legally binding and establish rights and obligations. In establishing and organizing legal relationships these texts are fundamentally different from statements of fact that are subject of most intralingual and monolingual corpora. Thus, texts with legal content differ in many respects from texts in standard language. Furthermore, Legal texts fulfill certain text functions (DIN-Normenausschuss Ergonomie, 2023). This text function, e.g. a legal binding, can deviate in the translation into plain language. These deviations should be consciously handled. For these reasons alone, the training of neural language programs for the legal domain must be based on suitable German-language training materials.

Specialized format: We pursue a participatory approach and collaborate with a large service provider and stakeholder of easy language recipients. The largest proportion of people with low literacy are disabled in some form. Most of them have difficulties to read texts, that exceed a half DIN A4 page, even if the text is written in easy language. Translating legal texts to a version that both maintains its meaning and is comprehensible to people with cognitive or language impairments, we need to define a specialized format. We propose the following four-level complexity hierarchy:

1. A summary in easy language of the underlying standard language /legal document, which has a pre-defined maximum length. This text version should be easy to read and understand for people who need low barrier text forms. It also helps the reader to appreciate the central meaning of the underlying document.

2. A longer version in easy language with jump markers that refer the reader to a glossary. This version is especially meant to be digital, so that the reader can access the glossary by a one-click action, that does not disturb the

Name	Doc. Pairs	Simplicity Versions	Genre	Domain	Published		URL
20 Minuten	18305	STD, SIM	ART	News	(Rios et al., 2021)	2021	-
KLEXIKON	2899	CH, AD	ENC	Encyclopedic	(Aumiller and Gertz, 2022)	2022	(Aumiller, 2023)
APA	2472	A2, B1	ART	News	(Säuberli et al., 2020)	2021	-
(apo)	2311	STD, SIM	ART	Medical	(Toborek and Busch, 2023)	2022	(Toborek et al., 2022)
Geo-Geolino	1627	CH, AD	ART	Science	(Hancke et al., 2012)	2022	-
Lexica	1090	CH, AD	ENC	Encyclopedic	(Hewett and Stede, 2021)	2021	(Hewett, 2022)
capito	752	A1, A2, B1	UNK	Unknown	(Rios et al., 2021)	2021	-
Tagesschau / Logo	415	CH, AD	SUB	News	(Weiß and Meurers, 2018)	2018	-
	378	STD, SIM	ART	Unknown	(Battisti et al., 2020)	2020	-
(bra), (mdr), (taz)	377	STD, SIM	ART	News	(Toborek et al., 2022)	2022	(Toborek et al., 2022)
	256	CH, AD	ART	Disability	(Klaper et al., 2013)	2013	-
(koe)	82	STD, SIM	ART	Government	(Toborek et al., 2022)	2022	(Toborek et al., 2022)
(beb), (lmt)	66	STD, SIM	ART	Disability	(Toborek et al., 2022)	2022	(Toborek et al., 2022)
TextComplexityDE	23	STD, SIM	ENC	Encyclopedic	(Seiffe et al., 2022)	2019	(Naderi, 2023)
(soz)	15	STD, SIM	ART	Politics	(Toborek et al., 2022)	2022	(Toborek et al., 2022)

Table 1: All available German parallel text simplification datasets and sub- datasets according to the Google Scholar results by using the query in section 1. For more details about the categorization, please refer to section 3.1. Simplicity Version are Standard Language (STD), any form of simple language (SIM), children-targeted (CH), adult-targeted-language (AD), and A1, A2, B1 are language level from the CEFR.

reading flow.

3. A complete version in easy language that should only reduce the linguistic complexity and not the complexity of content. It aims at conveying most of the (legal) statements of the original document. We assume that this version may be longer than the original text on which it is based.

4. The original text in standard language.

4   Research gaps and planned contributions

We identified and categorized existing resources for simplifying German texts with the aim of a preparatory assessment for the development of an NLM-based approach that supports accessible communication through participation-relevant texts. Our assessment of these intralingual-aligned and monolingual datasets as well as the existing approaches revealed the research gaps. Moreover, we observed that all monolingual datasets use illustrations to improve readability and intelligibility, while none of the parallel datasets do so. All identified datasets have a linear structure without any interactive elements, that could improve the readability. Based on our investigation and analyses in relation to the target group, we identify future areas of research:

1. Identification and investigation of existing texts, which are tailored to the needs of the target group and improve the readability of texts both in monolingual and parallel datasets.

2. Extension of parallel datasets by adding topics, domains and sub-domains, that are rele-

vant for the everyday life of the target group.

3. Addition of any form of illustration to the parallel datasets. By including visual elements, such as images, diagrams, or charts, the dataset becomes more inclusive and accessible to a wider range of users.

4. The transferability of the model to domains and sub-domains (e.g., legal sub-domains) for which it has not been trained.

5. The methodological development of evaluation methods that allow for an assessment that is in line with the objectives and purpose of accessibility, inclusion and participation by incorporating appropriate quantitative and qualitative methods. These evaluation methods may consider factors like readability scores, user feedback, comprehension tests, and other relevant metrics to measure the effectiveness of the model in promoting accessibility, inclusion, and participation.

We want to tackle all five research gaps in the future, so that researchers and developers can enhance the quality and applicability of language models for the target group, making information more accessible and engaging for a broader audience. Our current focus is to make legal texts more accessible in German easy language. Documents from this domain are often pivotal to a self-empowered life. Based on texts in this domain, we aim at designing specialized accessibility-enhanced formats.

Limitations

In this work, we examined the current state-of-the-art of German text simplification. It reviews neural language models for simplifying German texts and assesses their suitability for legal texts and accessibility requirements. The general methodology of this paper is applicable for any domain or language, but only works for the task of text simplification. Furthermore, the review only focuses on German, so no definitive conclusions about the situation for other languages can be made based on this work alone. Additionally, this paper relied on the current draft version of the DIN standard (DIN-Normenausschuss Ergonomie, 2023), the final version and its implications could deviate. Moreover, the DIN standard (DIN-Normenausschuss Ergonomie, 2023) is based on assumptions about its addressees, which we have not questioned further but simply adopted. These assumptions, e.g. include a homogeneity bias. Another limitation would be the limited use for pure information texts or transfers into information texts, i.e. that the target text function (in the sense of DIN) is always an informative one.

Ethics Statement

This paper complies with the ACL Ethics Policy[2]. The research field of this paper can help people to gain access to information by translating and transforming in an accessibility-enhanced way. Our presentation aims at motivating further scientific research and debate.

Acknowledgements

References

Miriam Anschütz, Joshua Oehms, Thomas Wimmer, Bartłomiej Jezierski, and Georg Groh. 2023. Language Models for German Text Simplification: Overcoming Parallel Data Scarcity through Style-specific Pre-training. ArXiv:2305.12908 [cs].

Dennis Aumiller. 2023. Klexikon: A German Dataset for Joint Summarization and Simplification. Original-date: 2022-01-05T09:09:42Z.

Dennis Aumiller and Michael Gertz. 2022. Klexikon: A German Dataset for Joint Summarization and Simplification. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 2693–2701.

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 12th Language Resources and Evaluation Conference*, pages 3302–3311, Marseille, France. European Language Resources Association.

Antje Baumann. 2020. Rechtstexte als Barrieren – Einige Merkmale der Textsorte "Gesetz" und die Verständlichkeit. In Christiane Maaß and Isabel Rink, editors, *Handbuch Barrierefreie Kommunikation*, 1 edition, volume 3 of *Kommunikation – Partizipation – Inklusion*, pages 679–702. Frank & Timme, Berlin.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. ArXiv: 2004.05150.

BGG. 2022. § 11 Disability Equality Act BGG.

Ursula Bredel and Christiane Maaß. 2016. *Leichte Sprache theoretische Grundlagen, Orientierung für die Praxis*. Sprache im Blick. Dudenverlag.

Mert Cemri, Tolga Çukur, and Aykut Koç. 2022. Unsupervised Simplification of Legal Texts. ArXiv:2209.00557 [cs].

Miguel Collantes, Maureen Hipe, and Juan Lorenzo Sorilla. 2015. Simpatico: A Text Simplification System for Senate and House Bills. In *Proceedings of the 11th National Natural Language Processing Research Symposium,*, volume 11, pages 26–32, Manila.

DIN-Normenausschuss Ergonomie. 2023. Empfehlungen für Deutsche Leichte Sprache (DIN SPEC 33429).

Isabel Gallegos and Kaylee George. 2022. The Right to Remain Plain: Summarization and Simplification of Legal Documents.

Michael Gille, Thorben Schomacker, Jörg von der Hülls, and Marina Tropmann-Frick. 2023. Der Einsatz von Neural Language Models für eine barriere-freie Verwaltungskommunikation.

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

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability Classification for German using Lexical, Syntactic, and Morphological Features. In *Proceedings of COLING 2012*, pages 1063–1080, Mumbai, India. The COLING 2012 Organizing Committee.

---

[2] https://www.aclweb.org/portal/content/acl-code-ethics

Freya Hewett. 2022. lexica-corpus. Original-date: 2021-08-13T09:12:24Z.

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.

David Klaper, S. Ebling, and Martin Volk. 2013. Building a German/Simple German Parallel Corpus for Automatic Text Simplification. In *Klaper, David; Ebling, S; Volk, Martin (2013). Building a German/Simple German Parallel Corpus for Automatic Text Simplification. In: The Second Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2013), Sofia, Bulgaria, 8 August 2013.*, pages 11–19, Sofia, Bulgaria. University of Zurich.

Tobias Kopp, Amelie Rempel, Andreas Schmidt, and Miriam Spieß. 2023. Towards machine translation into Easy Language in public administrations: Algorithmic alignment suggestions for building a translation memory. In Silvana Deilen, Silvia Hansen-Schirra, Sergio Hernández Garrido, Christiane Maaß, and Anke Tardel, editors, *Emerging Fields in Easy Language and Accessible Communication Research*, 1 edition, volume 14 of *Easy – Plain – Accessible*, pages 371–406. Frank & Timme.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742. Place: Cambridge, MA Publisher: MIT Press.

Christiane Maaß. 2020. *Easy Language – Plain Language – Easy Language Plus: Balancing Comprehensibility and Acceptability*, 1 edition, volume 3 of *Easy–Plain–Accessible*. Frank & Timme, Berlin. Accepted: 2020-09-28T09:51:54Z.

Laura Manor and Junyi Jessy Li. 2019. Plain English Summarization of Contracts. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 1–11, Minneapolis, Minnesota. Association for Computational Linguistics.

Babak Naderi. 2023. Text Complexity DE. Original-date: 2020-09-30T09:43:40Z.

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. Tex.ids= riosNewDatasetEfficient2021a.

Thorben Schomacker, Tillmann Dönicke, and Marina Tropmann-Frick. 2023. Exploring Automatic Text Simplification of German Narrative Documents.

Laura Seiffe, Fares Kallel, Sebastian Möller, Babak Naderi, and Roland Roller. 2022. Subjective Text Complexity Assessment for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 707–714, Marseille, France. European Language Resources Association.

Lucia Specia. 2010. Translating from Complex to Simplified Sentences. In *Computational Processing of the Portuguese Language*, Lecture Notes in Computer Science, pages 30–39, Berlin, Heidelberg. Springer.

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.

Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. DEPLAIN: A German Parallel Corpus with Intralingual Translations into Plain Language for Sentence and Document Simplification. ArXiv:2305.18939 [cs].

Julia Suter, Sarah Ebling, and Martin Volk. 2016. Rule-based Automatic Text Simplification for German. In *Proceedings of the 13th Conference on Natural Language Processing*, pages 279–287.

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.

Vanessa Toborek and Moritz Busch. 2023. A New Aligned Simple German Corpus. Original-date: 2022-08-22T10:58:53Z.

Vanessa Toborek, Moritz Busch, Malte Boßert, Pascal Welke, and Christian Bauckhage. 2022. A New Aligned Simple German Corpus.

UN. 2008. UN Convention on the Rights of Persons with Disabilities (CRPD).

Zarah Weiß and Detmar Meurers. 2018. Modeling the Readability of German Targeting Adults and Children: An empirically broad analysis and its cross-corpus validation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 303–317, Santa Fe, New Mexico, USA. Association for Computational Linguistics.