Automatic German Easy Language (Leichte Sprache) Simplification: Data, Requirements and Approaches

Thorben Schomacker[1]

Keywords: Text Simplification; Neural Machine Translation; Evaluation; Dataset; Easy Language; Leichte Sprache; Accessible Language; Simple Language; Comprehensibility-enhanced Language

1 Introduction

With the rise of the internet, it has become convenient and often free to access an abundance of texts. However, not all people, who have access, can really read and understand the texts. Despite the fact that, they speak the language that the text is written in. Most often this problem originates in the too complex nature of the texts. Text Simplification can help to overcome this barrier. In my dissertation, I want to specially focus on *Leichte Sprache* (German Easy Language). Which is a simplified version of German, that is tailored to the needs of people with cognitive disabilities.

In Sect. 2 I will talk about the target group and the motivation behind automatic translation to Leichte Sprache. Further, in Sect. 4 I will briefly discuss recent developments in the field and relevant previous work. Followed by Sect. 5 in which I will outline my project plan as well as my current architectural and implementation ideas. Concluded by my planned evaluation methodology in Sect. 6.

2 Motivation

For their daily affairs, the majority of the population uses everyday language (Alltagssprache) with slight regional variations. However, most written texts are in standard language (Standardsprache), which is similarly complex but has more standardized vocabulary [BM16, p.527-530]. Unfortunately, the complexity of standard language and everyday language is an obstacle for 12% of the population [GB20]. To access information, this group depends on comprehensibility-enhanced language. There are two main forms of simplified German language versions: easy language (Leichte Sprache) and simple language (einfache Sprache). Easy language describes a highly comprehensible and rule-based form of German, and simple language refers to a variety of simplifications in the gray area between standard language and easy language [Ma20].

[1] Hamburg University of Applied Sciences, Hamburg, Germany, thorben.schomacker@haw-hamburg.de

doi:10.18420/ki2023-dc-01

Who needs Leichte Sprache? The foremost reason for the translating complex texts into easy language is to make the information accessible for cognitively and learning impaired people[2]. The second reason, is the fact that translations are also a key to inclusion and social participation [UN08] .

Why automatic Leichte Sprache translation? The goal of the UN Convention on the Rights of Persons with Disabilities [UN08] is to allow all people regardless of their cognitive ability access to as many aspects of the everyday life as possible. Providing information in Leichte Sprache already highly contributes to accessibility and inclusion. Human-made translations require time intensive labor and cost circa 150€ per page [QW23], which limits their providers to a financially strong group. A recent study estimated that only 15% of sites with a ".de"domain provide easy-to-read texts (not necessarily Leichte Sprache) [AHZ23][3]. Since the recipients rarely request translations on their own, it is usually these large organizations that request translations and thus determine the selection of Leichte Sprache texts.

Automatic translation to Leichte Sprache makes Leichte Sprache financially accessible to a larger group of organizations. Additionally, if Leichte Sprache translation systems can be directly used by the recipients, the overall selection of Leichte Sprache text becomes recipient-designed rather than provider-designed. This can further nurture the inclusion of people with cognitive disabilities.

3 Research Question

Machine translation of Leichte Sprache is still under investigated. There are multiple angles to tackle the problem. The following research question outline particular research gaps. Cognitively and learning impaired people are considered as the *target group* (Sect. 2). All the research questions are only applied to Leichte Sprache (German Easy Language) and Germany as the core regions. Theoretically, they can be applied to other forms of Easy Language and other regions.

Q1 How can the target group be involved in creating a Leichte Sprache dataset?
Q2 How can the target group be involved in evaluating the performance of a Leichte Sprache generative AI-system?
Q3 By which rules and regulations[4] is Leichte Sprache defined?

[2] This term is used to be in accordance with DIN SPEC 33429-E [DI23]. Since, the terminology of the German legislation: "people with intellectual or mental disabilities" [BG22] is partly perceived as (too) deficit-oriented and discriminatory.

[3] The study of [AHZ23] only investigated the Oscar Corpus [Ab22], therefore this generalization should be considered with a grain of salt. In my opinion, the number can be roughly generalized and is the best estimation possible, since Oscar is based on the largest collection of web crawl data.

[4] I use the term *rules* to describe everything that describes and formalizes the linguistic facets of Leichte Sprache. *Regulation* means the legal framework for the use, production, and provision of Leichte Sprache. I also plan to discuss these two terms and their use in more detail in my dissertation.

Q4 Which requirements can be derived from current Leichte Sprache rules and regulations?
Q5 What annotation rules for a Leichte Sprache dataset cover the requirements from:
Q5a) the target group
Q5b) Leichte Sprache rules and regulations
Q5c) AI-regulation

4 Related Work

Text simplification can be described as a machine task translation, converting one version of a language to another: standard language to simple language. However, compared to other machine translation tasks, automatic text simplification is relatively new. The first data-based automatic text Simplification System [Sp10] for Portuguese was released in 2010.

Hancke et. al. [HVM12] released the first German corpus, that involves text simplification, in 2012. It consists of unaligned articles from GEO (similar to National Geographic) and GEOlino (GEO's edition for children). They used this new data set to train statistical classifiers to predict the reading level. Similar datasets with one adult-targeting and one children-targeting version of the same text have been published by [AG22; WM18].

In 2013 the the first sentence-aligned German simplification data set was published [KEV13]. It was used for the text simplification system for German, that used statistical machine translation [SEV16]. They argued that the corpus is not sufficiently enough large to such a system that works reasonably well. [SEV20] firstly applied Transformer encoder-decoder models to German Text Simplification on their novel APA corpus, which contains Despite being far larger than previous German dataset, this corpus was not large enough to sufficiently train a neural machine translation [SEV20].

[Ri21] adapted mBART [Li20] with Longformer Attention [BPC20] and applied it to the task of document-level text simplification. This approach was further improved by [Eb22]. This approach was expanded to a larger number of specific datasets [St22] in 2023. More recently, decoder-only Transformers become more prominent [An23; De23], which can be trained with monolingual data (simple language) only.

Survey Studies Automatic German text simplification has becoming increasingly scholarly investigated. This year, 2023, three surveys have been published about the current data sets and approaches in German text simplification [An23; Sc23; SMK23]. Prior to this year, no surveys have published at all.

Tools [SK22] published a web-tool to make the creation and modification of simplification corpora less difficult. However, they only offer very limited options of automatic sentence-alignment.

Automatic Translation Companies There are two companies, which offer the automatic translation to simple German: capito with a hybrid approach and SUMM with a fully automatic approach.

Regulation The automatic text simplification to Leichte Sprache is subject to two regulations: DIN SPEC 33429 [DI23] and AI-Act [Gi23]. DIN SPEC 33429 is a standard by the German Institute for Standardization Registered Association (DIN), that offers Guidance for Leichte Sprache. It aims at combining the interest of all interest groups involved in Leichte Sprache. Furthermore, its target is to combine all existing guidelines, such as [De16], into a unified form. Even though it is currently still a draft, it is foreseeable that it will be the most defining and authoritative set of rules for Leichte Sprache in the coming years.

5 Approach

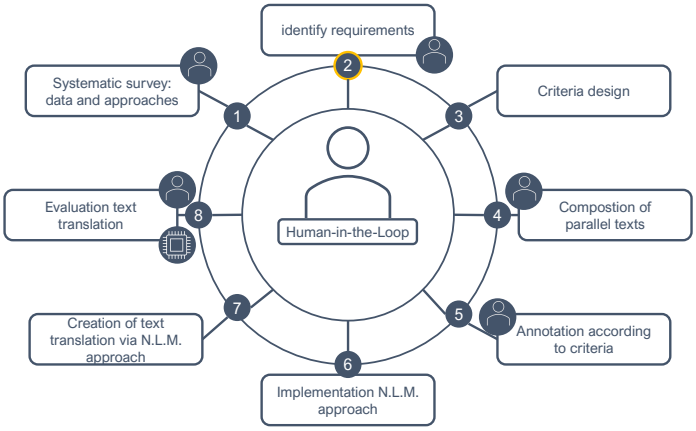

Fig. 1: Project Cylce

I have systemized a project cycle that involves eight steps, which is illustrated in Fig. 1. I am aware of the iterative nature of the project and incorporated into the design of the approach. The following subsections describe the steps:

1. Systematic Survey I want to provide a comprehensive review of the existing literature on automatic text simplification with neural language models. Furthermore, I will narrow my focus on the specifics of Leichte Sprache. This survey should enable me to gain a deep understanding of the practices, techniques, and challenges in German text simplification. I have already identified gaps and areas that require further exploration, but will continue the process. The systematic survey aims at producing an exhaustive tabular overview of parallel German Easy and Simple Language datasets. Tab. 1 shows a work-in-progress version of it.

2. Identify Requirements All requirements will be collected and represent in the evaluation methodology outlined in Sect. 6.

Name	Doc. Pairs	Simplicity Versions	Genre	Domain	Published		URL
20 Minuten	18305	STD, SIM	ART	News	[Ri21]	2021	-
KLEXIKON	2899	CH, AD	ENC	Encyclopedic	[AG22]	2022	[Au23]
APA	2472	A2, B1	ART	News	[SEV20]	2021	-
(apo)	2311	STD, SIM	ART	Medical	[TB23]	2022	[To22]
Geo-Geolino	1627	CH, AD	ART	Science	[HVM12]	2022	-
Lexica	1090	CH, AD	ENC	Encyclopedic	[HS21]	2021	[He22]
capito	752	A1, A2, B1	UNK	Unknown	[Ri21]	2021	-
Tagesschau / Logo	415	CH, AD	SUB	News	[WM18]	2018	-

Tab. 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. A complete version of this table is already published [Sc23].

3. Criteria Design All criteria will be collected and represent in the evaluation methodology outlined in Sect. 6.

4. Composition of parallel texts In an iterative process, I want to collect existing parallel data and create a novel dataset. This dataset will especially focus on relevant data samples and legal texts.

5. Annotation according to criteria After I gathered a set of requirements and criteria and compiled them into an evaluation catalog. And additionally composed a parallel dataset, I will start to semi-manual annotate the dataset.

Some annotation will be done by human annotators. They rate the quality of the text with a questionnaire with question such as "The text is written in a way, so that the coherence of the content and concepts becomes clear. Please rate on a scale from 0-10".

Additionally, automatic evaluation will be added by using tools such as the "LanguageTool Leichte Sprache rules"[5] or spacy[6].

6. Implementation N.L.M approach Previous approaches for German text simplification were always based on parallel data sets (text in the original standard language version and the corresponding simplification). Initially, statistic rules were used [SEV16]. From 2020 on, NLMs text simplification were increasingly used [Eb22; Ri21; SDT23; SEV20; SMK23; SRE21]. These approaches were all based on encoder-decoder transformer models. Recently, a decoder-only transformer approach was presented [An23], which can theoretically be trained only with monolingual data (simple/light language).

[5] https://github.com/languagetool-org/languagetool/blob/master/languagetool-language-modules/de-DE-x-simple-language/src/main/resources/org/languagetool/rules/de-DE-x-simple-language/grammar.xml

[6] https://spacy.io/models/de

Inline with previous work, I want to primarily implement and test encoder-decoder Transformer models. But I will explore implementation and architectural details in the future progression of my project.

7. Creation of text translation via N.L.M. approach I will use the approach discussed in the previous section to translate text to Leichte Sprache.

8. Evaluation text translation I will discuss my planned evaluation in Sect. 6.

6 (Planned) Evaluation

[GS22] and other works conclude, that the evaluation of simplification remains understudied. Evaluating simplified text is a hard task to solve, which roots in the challenge of defining a ”gold standard” simplification output. [GS22] names two reasons behind this root challenge: (1) it is not factual since it relies on transformations managed by ”simplificators” (human or automatic nature) and (2) it is heavily based on own the knowledge and opinion of people and thereby not consensual. Furthermore, unlike for standard language, a *native simplified-language speaker does not exist* [Si14].

DIN SPEC 33429		**Evaluation**			
Level	Rule	Rule	Language Tool	manual	participative
Word	Metaphor		METAPHERN	-	-
Sentence	Negation		VERNEINUNG	-	-
	Satzzeichen	Check forbidden characters	-	-	-
Text	Coherence				YES
Situation	Situation of use	*Always assume a specific situation of use and annotate it.*			

Tab. 2: The original criteria will be written in German. Since the DIN-SPEC is also in German. This table provides an excerpt, which was translated to English.

Evaluation measures are intended to serve as an assessment of the quality of simplification research, they can be classified in 1) human, 2) automatic and 3) semi-automatic (or hybrid) evaluation approaches. Since simplification can be considered as monolingual translation of documents from original to simplified languages, traditionally translation-related metrics are also applied to simplification.

Leichte Sprache poses a special challenge compared to other text simplification tasks. It is specified by rules and regulations and, in the future, also by a DIN standard [DI23]. In this respect, all texts would have to be checked for precisely these sets of rules in order to determine their level of Leichte Sprache. Currently, there are no scientific publications on this topic. I would like to make this form of evaluation possible through a new semi-automated approach. For this purpose, I will work through the DIN standard in order to derive requirements. For each of these requirements, I will define a form of evaluation and then

calculate an overall result from the individual evaluations. Tab. 2 outline a work-progress version of this rule catalog.

References

[Ab22] Abadji, J.; Ortiz Suarez, P.; Romary, L.; Sagot, B.: Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp. 4344–4355, June 2022, URL: https://aclanthology.org/2022.lrec-1.463, visited on: 06/23/2023.

[AG22] Aumiller, D.; Gertz, M.: Klexikon: A German Dataset for Joint Summarization and Simplification. In: Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022). Pp. 2693–2701, June 2022.

[AHZ23] Asghari, H.; Hewett, F.; Züger, T.: On the Prevalence of Leichte Sprache on the German Web. In: Proceedings of the 15th ACM Web Science Conference 2023. WebSci '23, Association for Computing Machinery, New York, NY, USA, pp. 147–152, Apr. 2023, ISBN: 9798400700897, URL: https://dl.acm.org/doi/10.1145/3578503.3583599, visited on: 06/23/2023.

[An23] Anschütz, M.; Oehms, J.; Wimmer, T.; Jezierski, B.; Groh, G.: Language Models for German Text Simplification: Overcoming Parallel Data Scarcity through Style-specific Pre-training, arXiv:2305.12908 [cs], May 2023, URL: http://arxiv.org/abs/2305.12908, visited on: 06/09/2023.

[Au23] Aumiller, D.: Klexikon: A German Dataset for Joint Summarization and Simplification, original-date: 2022-01-05T09:09:42Z, Feb. 2023, URL: https://github.com/dennlinger/klexikon, visited on: 05/06/2023.

[BG22] BGG: § 11 Disability Equality Act BGG, May 2022, URL: https://www.gesetze-im-internet.de/bgg/__11.html, visited on: 05/18/2023.

[BM16] Bredel, U.; Maaß, C.: Leichte Sprache theoretische Grundlagen, Orientierung für die Praxis. Dudenverlag, 2016, ISBN: 978-3-411-75616-2.

[BPC20] Beltagy, I.; Peters, M. E.; Cohan, A.: Longformer: The Long-Document Transformer, arXiv: 2004.05150, Dec. 2020, URL: http://arxiv.org/abs/2004.05150, visited on: 05/15/2021.

[De16] Deutsche Gesellschaft für Leichte Sprache: Regelwerk Leichte Sprache. 2016.

[De23] Deilen, S.; Garrido, S. H.; Lapshinova-Koltunski, E.; Maaß, C.: Using ChatGPT as a CAT tool in Easy Language translation, arXiv:2308.11563 [cs], Aug. 2023, URL: http://arxiv.org/abs/2308.11563, visited on: 08/25/2023.

[DI23] DIN-Normenausschuss Ergonomie: Empfehlungen für Deutsche Leichte Sprache (DIN SPEC 33429), DIN SPEC, Berlin, Apr. 2023.

[Eb22] Ebling, S.; Battisti, A.; Kostrzewa, M.; Pfütze, D.; Rios, A.; Säuberli, A.; Spring, N.: Automatic Text Simplification for German. eng, Frontiers in Communication 7/, Publisher: Frontiers Research Foundation, p. 706718, Feb. 2022, ISSN: 2297-900X, URL: https://www.zora.uzh.ch/id/eprint/218829/, visited on: 07/04/2022.

[GB20] Grotlüschen, A.; Buddeberg, K., eds.: LEO 2018: Leben mit geringer Literalität. wbv, Bielefeld, 2020, ISBN: 978-3-7639-6072-9 978-3-7639-6071-2.

[Gi23] Gille, M.; Schomacker, T.; von der Hülls, J.; Schomacker, T.: Der Einsatz von Neural Language Models für eine barrierefreie Verwaltungskommunikation: Anforderungen an die automatisierte Vereinfachung rechtlicher Informationstexte. In. 2023.

[GS22] Grabar, N.; Saggion, H.: Evaluation of Automatic Text Simplification: Where are we now, where should we go from here. In: Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale. ATALA, Avignon, France, pp. 453–463, June 2022, URL: https://aclanthology.org/2022.jeptalnrecital-taln.47, visited on: 08/03/2022.

[He22] Hewett, F.: lexica-corpus, original-date: 2021-08-13T09:12:24Z, Aug. 2022, URL: https://github.com/fhewett/lexica-corpus, visited on: 05/07/2023.

[HS21] Hewett, F.; Stede, M.: Automatically evaluating the conceptual complexity of German texts. In: Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021). KONVENS 2021 Organizers, Düsseldorf, Germany, pp. 228–234, 2021, URL: https://aclanthology.org/2021.konvens-1.23, visited on: 04/12/2023.

[HVM12] Hancke, J.; Vajjala, S.; Meurers, D.: Readability Classification for German using Lexical, Syntactic, and Morphological Features. In: Proceedings of COLING 2012. The COLING 2012 Organizing Committee, Mumbai, India, pp. 1063–1080, Dec. 2012, URL: https://aclanthology.org/C12-1065, visited on: 07/01/2022.

[KEV13] Klaper, D.; Ebling, S.; Volk, M.: 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. University of Zurich, Sofia, Bulgaria, pp. 11–19, Aug. 2013, URL: https://www.zora.uzh.ch/id/eprint/78610/, visited on: 07/01/2022.

[Li20] Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; Zettlemoyer, L.: Multilingual Denoising Pre-training for Neural Machine Translation. Transactions of the Association for Computational Linguistics 8/, Place: Cambridge, MA Publisher: MIT Press, pp. 726–742, 2020, URL: https://aclanthology.org/2020.tacl-1.47, visited on: 07/21/2022.

[Ma20] Maaß, C.: Easy Language – Plain Language – Easy Language Plus: Balancing Comprehensibility and Acceptability. Accepted: 2020-09-28T09:51:54Z, Frank & Timme, Berlin, 2020, ISBN: 978-3-7329-9268-3.

[QW23] Quabeck, A.; Wolfram, A.: Schlüssel Leichte Sprache: Unsere Leistungen, de-DE, 2023, URL: https://schluessel-leichte-sprache.de/leistungen, visited on: 06/23/2023.

[Ri21] Rios, A.; Spring, N.; Kew, T.; Kostrzewa, M.; Säuberli, A.; Müller, M.; Ebling, S.: A New Dataset and Efficient Baselines for Document-level Text Simplification in German. In: Proceedings of the Third Workshop on New Frontiers in Summarization. tex.ids= riosNewDatasetEfficient2021a, Association for Computational Linguistics, Online and in Dominican Republic, pp. 152–161, Nov. 2021, URL: https://aclanthology.org/2021.newsum-1.16, visited on: 07/04/2022.

[Sc23] Schomacker, T.; Gille, M.; Tropmann-Frick, M.; von der Hülls, J.: Data and Approaches for German Text Simplification – Next Steps toward an Accessibility-enhanced Communication, 2023.

[SDT23] Schomacker, T.; Dönicke, T.; Tropmann-Frick, M.: Exploring Automatic Text Simplification of German Narrative Documents, 2023.

[SEV16] Suter, J.; Ebling, S.; Volk, M.: Rule-based Automatic Text Simplification for German. In: Proceedings of the 13th Conference on Natural Language Processing. Pp. 279–287, 2016.

[SEV20] Säuberli, A.; Ebling, S.; Volk, M.: 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). European Language Resources Association, Marseille, France, pp. 41–48, May 2020, ISBN: 979-10-95546-45-0, URL: https://aclanthology.org/2020.readi-1.7, visited on: 07/04/2022.

[Si14] Siddharthan, A.: A survey of research on text simplification. en, ITL - International Journal of Applied Linguistics 165/2, Publisher: John Benjamins, pp. 259–298, Jan. 2014, ISSN: 0019-0829, 1783-1490, URL: https://www.jbe-platform.com/content/journals/10.1075/itl.165.2.06sid, visited on: 11/07/2022.

[SK22] Stodden, R.; Kallmeyer, L.: 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. Association for Computational Linguistics, Dublin, Ireland, pp. 145–155, May 2022, URL: https://aclanthology.org/2022.acl-demo.14, visited on: 07/04/2022.

[SMK23] Stodden, R.; Momen, O.; Kallmeyer, L.: DEPLAIN: A German Parallel Corpus with Intralingual Translations into Plain Language for Sentence and Document Simplification, arXiv:2305.18939 [cs], May 2023, URL: http://arxiv.org/abs/2305.18939, visited on: 06/09/2023.

[Sp10] Specia, L.: Translating from Complex to Simplified Sentences. In (Pardo, T. A. S.; Branco, A.; Klautau, A.; Vieira, R.; de Lima, V. L. S., eds.): Computational Processing of the Portuguese Language. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, pp. 30–39, 2010, ISBN: 978-3-642-12320-7.

[SRE21] Spring, N.; Rios, A.; Ebling, S.: Exploring German Multi-Level Text Simplification. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). INCOMA Ltd., Held Online, pp. 1339–1349, Sept. 2021, URL: https://aclanthology.org/2021.ranlp-1.150, visited on: 07/04/2022.

[St22] Stodden, R.: Erstellung eines parallelen Vereinfachungskorpus für die deutsche Sprache – Unter Verwendung des HHU Annotationstools TS-anno, original-date: 2020-09-29T16:14:03Z, Düsseldorf, Jan. 2022, URL: https://github.com/rstodden/TS_annotation_tool, visited on: 05/31/2023.

[TB23] Toborek, V.; Busch, M.: A New Aligned Simple German Corpus, original-date: 2022-08-22T10:58:53Z, May 2023, URL: https://github.com/mlai-bonn/Simple-German-Corpus, visited on: 05/06/2023.

[To22] Toborek, V.; Busch, M.; Boßert, M.; Welke, P.; Bauckhage, C.: A New Aligned Simple German Corpus, 2022, URL: https://arxiv.org/abs/2209.01106.

[UN08] UN: UN Convention on the Rights of Persons with Disabilities (CRPD), en, May 2008, URL: https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-persons-disabilities, visited on: 05/18/2023.

[WM18] Weiß, Z.; Meurers, D.: 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. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 303–317, Aug. 2018, URL: https://aclanthology.org/C18-1026, visited on: 07/01/2022.