EASY – PLAIN – ACCESSIBLE

Regina Stodden

Automatic German Text Simplification: Data, Evaluation, and Models

Frank & Timme

Regina Stodden

Automatic German Text Simplification: Data, Evaluation, and Models

Silvia Hansen-Schirra / Chris Maaß (eds.)
Easy – Plain – Accessible
Vol. 21

Regina Stodden

Automatic German Text Simplification: Data, Evaluation, and Models

Frank & Timme
Verlag für wissenschaftliche Literatur

To read the free, open access version of this book online, visit https://www.frank-timme.de/de/programm/produkt/automatic_german_text_simplification or scan the QR code. At the same place, you can also download all images in scalable, high resolution and download the full appendix. On the last pages of this work, you can find a summary in plain language.

This work was supported by the North Rhine-Westphalian funding scheme called "Forschungskolleg Online-Partizipation" and by the Open Access Fund of the Heinrich Heine University Düsseldorf.

peer reviewed content

Research Organization Registry, https://ror.org/024z2rq82

ISBN 978-3-7329-1216-2
ISBN Open Access 978-3-7329-8700-9
ISSN 2699-1683
DOI 10.57088/978-3-7329-8700-9

Frank & Timme GmbH Verlag für wissenschaftliche Literatur
Berlin 2026.

Herstellung durch Frank & Timme GmbH
Wittelsbacherstraße 27a, 10707 Berlin
info@frank-timme.de
Gedruckt auf säurefreiem, alterungsbeständigem Papier.

www.frank-timme.de

Dissertation Heinrich Heine University Düsseldorf, 2024, D61

"Making the simple complicated is commonplace; making the complicated simple, awesomely simple, that's creativity."

Charles Mingus (1922 – 1979; jazz composer)

Acknowledgments

I have been lucky to get a lot of support from different people while I have been working on this thesis. I would like to thank them all for their help. First, I would like to express my deepest gratitude to my two supervisors, *Prof. Dr. Laura Kallmeyer* and *Prof. Dr Stefan Conrad*. Thank you for your wise advises, remembering myself to concentrate on the primary tasks and stopping me (mostly) before doing superfluous work.

Laura, thanks for offering me the opportunity to do a PhD in your lab. I really appreciate your excellent balance of friendliness and assertiveness. You have always motivated and provided the necessary drive to see this work through to completion. I always received more professional, faithful, and thematic support than I could have ever expected. Thank you for always finding the time to give me feedback and to improve my work, even if I had to shorten our schedule. Your support has allowed me to become a better researcher and writer.

Stefan, thanks for showing great interest in my work. In all our meetings I felt your trust in my research which always calmed me down. Your thoughtful comments has helped me to reflect and improve my work.

I had the pleasure to work with great colleagues in the *Computational Linguistics lab at HHU*. I am sorry for attending our socials only rarely, but when I have attended I felt very welcome and comfortable. Many thanks to *Dr. Behrang QasemiZadeh* for introducing me to the academic world and for encouraging me to do a PhD. I would also like to thank *Dr. Younes Samih* for encouraging me to believe more in me and my own work and to think more big. Without your recommendation I would have probably never thought that my work could also be sufficient for top-tier conferences. I am also grateful to *Prof. Dr. Wiebke Petersen* for inspiring and in-depth discussions on text simplification, and NLP in general. Lastly, I would like to mention my colleagues, *Maya Kruse* and *Omar Momen*, and thank them for their awesome assistance and discussions during the creation of the new resources for German text simplification.

Many thanks goes also to the *text simplification community* (especially the Slack group) for warmly adopting me. The discussions in our reading group and our meetings at conferences inspired me a lot. It was simultaneously fascinating and calming to see that even though we work on different languages, we all face the same or at least similar challenges. It also opened my mind to think in different directions within the scope of text simplification. Special thanks go also to *Miriam Anschütz* and *Thorben Schomacker* for great exchange and musing on German text simplification. Although we got in touch only at the end of my PhD journey, I felt like I would have known and worked with you already for years; I learned a lot from your different views on our research field and really enjoyed working with you.

Additionally, this endeavor would not have been possible without the support from the North Rhine-Westphalian (German) funding scheme *"Forschungskolleg Online Partizipation"*, who financed my research. I am also grateful to my *PhD colleagues of the Forschungskolleg*, I enjoyed our interdisciplinary exchange on and beyond the topic of online participation as well as our entertaining trips to the stakeholders and conferences. It is a pity that the pandemic has meant that we had fewer of these experiences. Thanks should go especially to one (former) PhD colleague, *Dr. Phillip Nguyen*, I do not want to miss our amusive and productive meetings for our study on usage intention. I would like to thank *Dr. Dennis Frieß*, *Lena Schwarz*, and *Prof. Dr. Lars Heilsberger* in their roles as coordinators of the Forschungskolleg for the organization of all our (in-person and online) meetings and for all the support they have given us during the PhD program.

I would like to express my attitude to the data providers of my new corpora, especially the *Austrian Press Agency*, to allow me to use and distribute their data for my research. I also would like to acknowledge, *Katja Gabrovska*, the librarian I trust; thanks for seemingly endless discussions on copyright and data protection.

Finally, I cannot thank my family and friends enough especially *Papa & Mama*, *Irene & Christian*, *Klaus & Luisa*, *Bene* (*& his family*), *Rebecca*, *Dirk & Simone*, *Andreas & Katja*, *Oma*, *Melli*, *Ruben*, *Omar*, *Maik*, *Melitta*, and *Zahra*. Thank you all (and also the ones for which names I did not had enough space) for all your great support, trust, patience, and love. Even though I was often annoyed when you asked me about my progress, I was always happy about your interest in my work. Also I want to apologize for my (partial) absence and ghosting especially in the last time during finalizing the thesis, I will try my best to come to light again.

Dear *Mama* and *Papa*, my book is finally done! Even if my graduation has been that new and unfamiliar to you (and also to me), I always felt your deepest support. I cannot thank you enough for sparking my curiosity to discover the world; I will never forget where my roots are.

Dear *Bene*, I cannot express my gratitude in words for bearing with me, cheering me up, distracting me and most of all getting me and you through this journey in the good and tough times. I am so grateful to have you in my life!

Contents

List of Figures

List of Tables

Part I

Introduction & Foundation

Chapter 1

Introduction

1.1 Motivation & Relevance

16.8 million people (including 6.2 million functional illiterates) in Germany (Grotlüschen et al., 2020) and 1.9 million people in Austria (Kneil et al., 2020) have noticeable problems with reading and writing in German. Non-native German speakers, older people, and persons with low literacy may have a limited vocabulary or know only a few grammatical constructions, so they have problems understanding lexically or syntactically complex sentences (Bredel and Maaß, 2016; Saggion, 2017).

However, the ability to access and understand written texts is essential to reflect the information given and to be able to form and formulate one's own opinion on a topic (Park, 2012). Hence, the comprehension of texts greatly affects self-determination and participation in society (Bock, 2015).

To enhance the readability of texts and, accordingly, enhance the participation of people with reading problems in society, linguists and translators provide guidelines for plain language on how to manually reduce the complexity of texts at the lexical, syntactical, and content level, e.g., in German "Leichte Sprache" (EN: (German) Easy Language; Netzwerk Leichte Sprache 2022; Bredel and Maaß 2016) and "Einfache Sprache" (EN: (German) Plain Language; Deutsches Institut für Normung (DIN) 2024a; Baumert 2018). However, manual simplification is very resource-intensive, so digital, supportive tools were developed, such as proofreading, complexity check tools, or automated text simplification tools.

Automated text simplification harks back to research in the late 1990s (e.g., see Carroll et al. 1998 or Chandrasekar et al. 1996) and is defined as automatically reformulating (lexical simplification) or restructuring (syntactical simplification) texts in a way that they are easier to understand for a specific target group, but that the texts still preserve their original meaning (Siddharthan and Mandya, 2014). Therefore, the overall aim of automated text simplification is to make texts more coherent and easier to follow for, e.g., people with reading problems or foreign language learners. However, as an intermediate objective, machine-generated simplifications could also support professional translators to simplify a text faster by reducing their cognitive load and repetitive work (Hansen-Schirra et al., 2020b).

The automatic text simplification (ATS) process can be expressed in a pipeline or a workflow that is similar to the one of machine translation. Following Garbacea et al. (2021), the text simplification pipeline (including manual, computer-assisted, and automatic simplification) contains three main parts:

1. classifying whether a text should be simplified (complexity prediction) and then identifying complex passages in texts (complexity explanation) (see component 0 in Figure 1.1),
2. generating a simplification (see component G), and
3. validating of the generated simplification (see component H).

I specify this general workflow with a focus on automatic simplification by extending the components regarding the generation and evaluation parts and adding the corpus building process. A workflow diagram for automatic text simplification is provided in Figure 1.1. In general, the workflow can be divided into the training and test phase (upper and lower parts), as well as the data and model parts (left and right parts).

The main components, which hold also for machine translation, are: using a parallel corpus (see components B and E in Figure 1.1) to train (see component G) and evaluate (see component H) a text simplification model (see component G). In order to optimize the model for the given development data, the evaluation can be seen as a loop (from component G to H to G): a model is modified regarding some settings (e.g., hyperparameters or prompts), again trained on the same data, and the new system generations are evaluated to verify whether the model has improved in comparison to the previous run. In consideration of these components, there is significant potential for enhancement with regard to the simplification of German text.

In my PhD thesis, I advance the components of the automatic text simplification workflow for written German Plain German. I focus on improving the workflow by facilitating the building and annotation of new corpora, rethinking the evaluation of TS, and designing new TS models. I incorporate research findings from translation studies and machine translation research to build machine learning-based models for intra-lingual translation, i.e., simplification. These models aim to reformulate a given complex text to the needs of, e.g., people with reading difficulties by applying simplification operations such as rewriting sentences in passive voice to active voice or replacing complex terms with easier synonyms (Shardlow, 2014; Alva-Manchego et al., 2020b).

The language of my interest is German; hence, the main focus of my work is on the simplification of German written texts. I have decided on German because, on the one hand, in recent decades, automatic text simplification has been well studied for English (see, e.g., Shardlow 2014; Stajner 2021; Ryan et al. 2023), but German ATS research is a smaller, growing research field which research has just been initiated within the last ten years (see, e.g., Ryan et al. 2023). On the other hand, German is difficult to understand due to some idiosyncrasies, e.g., umlauts, nested sentences, or inflection of adjectives (Marzari, 2010). Furthermore, much research has been done regarding German manual simplification (see e.g., Jekat et al. 2014, Bredel and Maaß 2016, Bock 2019, Maaß 2020, or Bock and Pappert 2023) and also a huge awareness has been gained in the German society about German simplification due to the amount of available simplified texts.

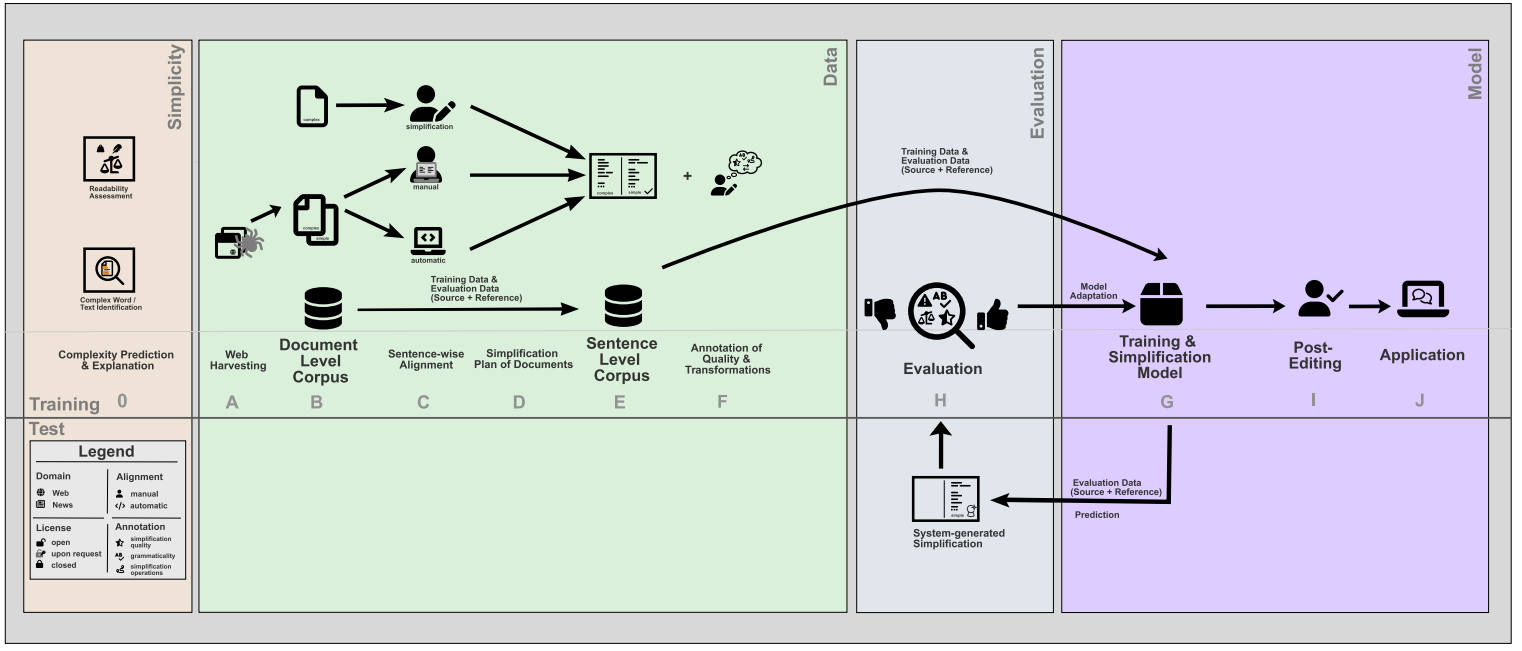


Figure 1.1: Workflow diagram of text simplification including the corpus building process.

I further mainly tackle simplification of texts into German Plain Language (DE: "Einfache Sprache") and only slightly consider other simplification purposes, e.g., expert-laypeople simplification (Trienes et al., 2022), simplification for children (Hewett and Stede, 2021; Aumiller and Gertz, 2022), or simplification into German Easy Language ("Leichte Sprache") (Siegel et al., 2019). The main reason for choosing German Plain Language is that more parallel texts or more complex-simple pairs exist for German Plain Language than for the other simplification purposes. Although many simplified texts exist for children or in German Easy Language, they are often independently written from other texts (see Section 4.7). Hence, no complex-simple parallel version exists for them, and they cannot be directly used for simplification. Although it sounds like a simple reason, the availability of high-quality complex-simple pairs and their size greatly affect the quality of TS systems (Jiang et al., 2020).

Furthermore, German Plain Language is a less controlled language than German Easy Language and still contains a few more complex structures. Hence, German Plain Language is closer to Standard German and contains less strong simplifications. So we currently expect a text simplification system to learn and predict German Plain Language better than German Easy Language.

Overall, in my work, I am combining the research field of linguistic complexity of German, German intra-lingual translation, and machine translation to build and evaluate German text simplification models.

1.2 Research Aims & Contributions of this Thesis

1.2.1 Research Questions

This cumulative thesis addresses the general research question of *how the potential of machine learning methods can be explored for the simplification of German texts, considering data availability and evaluation suitability and comparing their effectiveness for document and sentence simplification approaches* (RQ 1). In order to more clearly outline the research program, I break the thesis down into five thematic segments, which further structure each part of the thesis (i.e., state of the research in Part I, contributions in Part II, and discussion of contributions wrt. research questions in Part III). The thematic segments are as follows:

1. background of complexity and simplification,
2. building text simplification corpora,
3. completed resources for text simplification,
4. evaluation of text simplification, and
5. text simplification models.

Each thematic segment follows subordinated research questions which I introduce and motivate in the remainder of this section. I also describe our contributions[1] regarding each part and explain their value and relevance for research in German text simplification.

1 In the following, the pronoun "we" and the possessive form "our" are including my highly valued co-authors and myself.

1.2.2 Background of Complexity and Simplification

Automatic text simplification is a rising topic in English research, but has not been well studied for other languages such as German. Nevertheless, for German manual simplification, a lot of research has been conducted on what makes a text complex and how to simplify texts for specific target groups manually (see Part I, Chapter 2). Therefore, in this thematic segment of the thesis, I investigate

- *Which simplification operations are commonly used to write simplified German* [Research Question 2-1]?, and
- *How and to what extent can complex passages be automatically identified*? [RQ 2-2]

To answer these questions, Stodden and Kallmeyer (2020) (see Part II, Subsection 7.2.1) present an analysis of linguistic features and simplification operations in different languages (including, e.g., German and English) and domains (including, e.g., news, web, and Wikipedia data). We found that (unsurprisingly) some linguistic operations, such as lexical simplification, are important across all languages. To facilitate lexical simplification, in Stodden and Venugopal (2021) (see Part II, Subsection 7.2.2), we propose a machine learning-based method to detect complex words in English using linguistic features. Although this approach was originally built for English TS, due to previous findings, we argue that it can also be helpful for the identification of complex words in German.

1.2.3 Building Text Simplification Corpora

For automated text simplification, parallel corpora are required to train or to evaluate text simplification models. When building a new (or evaluating an existing) resource, many decisions have to be made (see Part I Chapter 3), e.g., regarding the selection of the right data (which text level, domain, or target group?) or annotation of the data (which annotations, e.g., sentence-wise alignment, simplification operations, simplification quality?). However, this building process can be very complex and time-consuming. Therefore, in this thematic segment, I focus on the following subordinated research questions:

- *What are the challenges faced in the process of creating a representative corpus for German text simplification and how to overcome them?* [RQ 3-1]
- *What are the key characteristics and features of parallel corpora for TS that should be included when building new corpora to improve German text simplification?* [RQ 3-2], and
- *How can the corpus's quality and representativeness be analysed to ensure its suitability for the research of German text simplification?* [RQ 3-3]

We address these questions by working out steps and challenges during the process of building parallel text simplification corpora. In Stodden and Kallmeyer (2022) (see Part II, Subsection 7.3.2) we first name the steps and challenges (e.g., web crawling, sentence-wise alignment, or annotation of simplification pairs) and then solve them by proposing a new annotation tool which supports sentence-wise alignment, manual simplification, annotation of simplification operations, annotation of quality ensuring criteria wrt. simplification. We also propose an extensive annotation schema for German text simplification corpora in Stodden (2022) (see Part

II, Subsection 7.3.1). The schema can be applied to evaluate the quality, representativeness, and suitability of gold simplifications as well as of system-generated simplifications.

Although our contributed text simplification annotation tool facilitates the manual sentence-wise alignment of simplification pairs, this process is still very time-consuming. In order to speed up the alignment, in Stodden et al. (2023) (see Part II, Subsection 7.4.2), we propose and compare a few algorithms to automatically align parallel documents also on the sentence level.

1.2.4 Resources for Text Simplification

In addition to building new corpora for the purpose of German text simplification, existing German corpora can also be used to train or evaluate text simplification systems. The same as for English text simplification research, German text simplification research also mainly focuses on corpora derived from news texts, Wikipedia, or web texts (see Part I, Chapter 4). Hence, in this thematic segment, I am trying to widen the field by answering the following questions:

- *Which text domains are currently not considered in automatic or manual German text simplification? Do the texts of these domains require simplification, if so, for which target group should they be simplified?* [RQ 4-1]

Furthermore, for German TS only fewer and smaller corpora exist than for English (see Part I, Chapter 4). New resources for German text simplification might also help to push German text simplification forward. Following this, this thematic segment also addresses the question

- *To what extent can new parallel corpora help to improve German text simplification?* [RQ 4-2]

In Stodden (2021a) (see Part II, Subsection 7.4.1); we evaluate the complexity of user-generated texts as a domain which is not yet considered in either manual or automatic simplification yet to address RQ 4-1. In Part I Chapter 4, existing German text simplification corpora will be introduced and further discussed with respect to the characteristics and issues identified in the part of building new corpora. We come to the conclusion that new high-quality German corpora are required for document and sentence simplification.

Therefore, in Stodden et al. (2023) (see Subsection 7.4.2), we propose five new German corpora for document and sentence simplification. Two of those are news corpora: DEplain-APA-doc and DEplain-APA-sent; and three of those are web corpora: DEplain-web-doc, DEplain-APA+web-doc, DEplain-web-sent. In more detail, DEplain-APA overcomes the problem of missing manually aligned and huge corpora, whereas DEplain-web-sent is partially manually and automatically aligned. In order to analyze the relevance of the new corpora as training data and evaluation data (see RQ 4-2), in Stodden et al. (2023) (see Subsection 7.4.2), we use the corpora as training and evaluation data for new text simplification models.

1.2.5 Evaluation of Text Simplification

The next thematic segment is the evaluation of automatic text simplification in order to explain the relevant aspects of interpreting the quality of text simplification models. Mostly, text simplification is evaluated using automatic metrics, but also some studies include manual evaluation (see Part I, Chapter 5). However, the automatic metrics and the manual evaluation protocol

have been designed for English TS evaluation and, for both, best practices are currently missing. Furthermore, current metrics are highly discussed regarding their suitability for evaluating text simplification (see Part I, Chapter 5). Following this, I derive these subordinated research questions for this thematic segment:

- *Are the tools for manual and automatic evaluation of simplified texts robust and reliable?* [RQ 5-1]
- *Does the effectiveness of text simplification evaluation strategies designed specifically for the English language differ when applied to texts in other languages?* [RQ 5-2]
- *What are the key aspects that need to be considered when using evaluation approaches for manual and automatic evaluation processes in German text simplification? And what are the possibilities of including them in new evaluation approaches?* [RQ 5-3]

To answer these questions, we first confirm that there are different evaluation strategies for manual evaluation (see Stodden (2021c), Part II Subsection 7.5.1) and in automatic evaluation (see Stodden (2024a), Part II Subsection 7.5.3). As solutions, in Stodden and Kallmeyer (2022) (see Part II Subsection 7.3.2), we propose a new manual evaluation protocol and, in Stodden (2024a) (see Part II Subsection 7.5.3) a new framework for a more easy automatic evaluation of German TS. Furthermore, in Arps et al. (2022), we propose a new method to automatically assess the complexity of German sentences.

1.2.6 Text Simplification Models

Last but not least, I also analyze systems and system generations of German text simplification. For English and German text simplification, there already exist some text simplification systems following different machine learning-based approaches (see Part I Chapter 6). However, it is unclear which approach performs best when considering our previous contributions regarding new data and new evaluation methods. Furthermore, the suitability of existing models has not yet been discussed with respect to real-world usage, i.e., whether TS models can be used in applications without major restrictions or caution. In order to tackle these research gaps, I will also focus on the following subordinated research questions:

- *How to connect automatic document and sentence simplification more?* [RQ 6-1]
- *How does the architecture and training data affect the efficacy of simplification models?* [RQ 6-2]
- *How does manual or automatic text simplification affect the usage of real-world applications, such as e-participation processes?* [RQ 6-3]

Our contributions in Stodden et al. (2023) (see Part II Subsection 7.4.2) tackle these research questions by contributing new German models for document and sentence simplification based on fine-tuning an mBART model. We compare the capability of this model and other models wrt. different training data (and different sizes) in Stodden (2024b) (see Part II Subsection 7.6.1), i.e., a reproduction study for German TS. To answer the question about the effect of text simplification on real-world applications, in Stodden and Nguyen (2024), we conduct a near-realistic experimental study in which we analyze the acceptance of participants regarding online participation processes considering the absence of simplified texts versus manually and automatically simplified texts.

1.3 Structure of the Thesis

My thesis is structured into three main parts: an introduction with research questions and a summary of the current state of TS (current part, i.e. Part I), publications (see Part II), and discussion (see Part III). The whole structure of my PhD thesis is visualized in Figure 1.2 (including the TS workflow on the left).

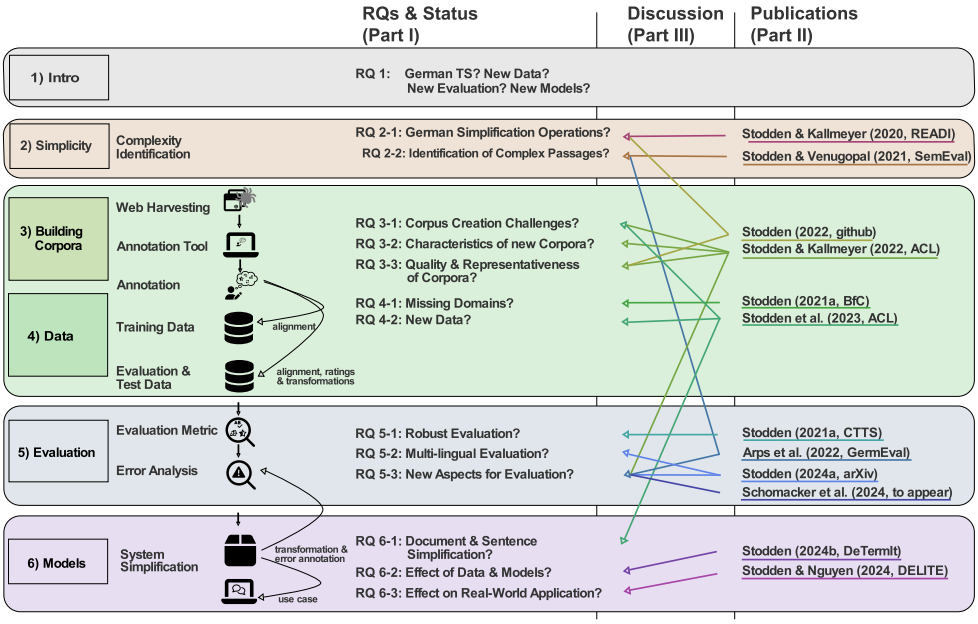


Figure 1.2: Contributions of my thesis including corresponding chapters and publications.

In the introduction part (current part, that is, Part I), I present and discuss the current state of the research, derive research gaps regarding German text simplification, and provide more motivation regarding our research questions. I include an overview regarding five main points, i.e., background regarding simplicity, corpora for text simplification (including the building process and completed corpora), evaluation of text simplification, and automated text simplification systems.

Then, in the part of publications (see Part II), I bring in contributions made by me and my co-authors (in the form of my publications) to narrow or close the research gaps summarized in Part I. The included contributing research papers have been published in the proceedings of international peer-reviewed conferences or workshops. For all publications, links are povided to access their open accessible full text version. For a concise overview of the publications in chronological order see the Appendix. It comprises a comprehensive list of the publications included in my dissertation, together with additional materials (e.g., code, data, poster or videos).

Finally, in the discussion part (see Part III), I combine the research contributions and results presented in Part II. I discuss the contributions of my co-authors and me in a comprehensive manner with respect to the current state of research, the identified research gaps, and research questions presented in Part I.

Each of the three parts follows the same structure derived from the simplification workflow diagram. Each part starts with a short preface (see Part I current chapter, Chapter 1; Part II Chapter 7, and Part III Chapter 8) followed by preliminary notes on complexity and simplification (see Part I Chapter 2; Part II Section 7.2; Part III Section 8.2). Then it contains two chapters focusing on resources for text simplification, first, with a focus on the building process of text simplification corpora (see Part I Chapter 3; Part II Section 7.3; Part III Section 8.3) and then regarding completed German simplification corpora (see Part I Chapter 4; Part II Section 7.4; Part III Section 8.4). Afterwards, I concentrate more on the technical side of text simplification: i.e., evaluation of text simplification (see Part I Chapter 5; Part II Section 7.5; Part III Section 8.5), and models of German text simplification (see Part I Chapter 6; Part II Section 7.6; Part III Section 8.6).

Chapter 2

Complexity and Simplification

In this section, I explain concepts that are necessary to understand the remainder of this work, e.g., comprehensibility, complexity, readability, complexity-reduced varieties of German, simplification operations, and differences between plain, easy, and simplified (see Section 2.1). As completion of this section, I define the automatic text simplification workflow and its components (see Section 2.2).

2.1 Background of Simplification

First, I introduce important terms such as comprehensibility, readability, and complexity before I use them to describe more advanced concepts, e.g., simplification operations, or literacy levels.

2.1.1 Comprehension & Comprehensibility & Readability

I begin with defining the distinction between *comprehension* and *comprehensibility* as well as between *legibility* and *readability*. Following Wolfer (2015, p. 34), comprehension is "the process of understanding a text by building up a mental representation", whereas they define comprehensibility as "the concept of [...] how easy a text can be comprehended"(Wolfer, 2015, p. 34). Hence, the first term refers more to the process of understanding a text, whereas the second term is more closely related to the complexity and readability of a text, but refers more to the user perspective than the text perspective.

Further, in order to comprehend a text, it must be legible and readable to the reader. Legibility refers to the physical skill of being able to identify characters and words properly; therefore, it is not hampered by unclear fonts, contrasts, or font sizes (Wolfer, 2015). In contrast, readability refers to assessing the comprehensibility of a text by measuring its surface characteristics, e.g. the average sentence or word length in texts (Wolfer, 2015; Shardlow, 2014). The comprehension process can be further analyzed by measuring the reading time of a text or tracking the eye movements during reading a text. As a result of successful comprehension, a reader should have understood a text well, which can be verified with multiple-choice or open comprehension questions, as well as a recall of its content. Simplification addresses all introduced terms, because it aims at reducing the cognitive processing costs of comprehension as well as improving the readability for easier comprehensibility.

2.1.2 Linguistic Complexity

Linguistic complexity is also called *system complexity* or *structural complexity*; it can be separated into *absolute complexity* and *relative complexity* Miestamo (2008). Absolute complexity comprises the complexity of a system, e.g., one language, considering its variety and richness in morphology, phonology, syntax, and semantics (Dahl, 2004; Pallotti, 2015; Tolochko and Boomgaarden, 2019).

In comparison, relative complexity is more closely related to *difficulty*: it describes the difficulty or the cost of processing or learning a language feature from an agent's point of view. A feature is complex if it is costly or difficult to process. Based on experiences with languages previously learned, it can be more difficult, complex, or a higher mental effort for a person to learn a language than for another without or other experiences. Therefore, relative complexity is rather subjective; what some people find complex can be simple for other people Miestamo (2008).

2.1.3 Text Simplicity & Text Complexity & Text Difficulty

Following Shardlow (2014) and Vecchiato (2022), for *text simplicity* no clear definition exists; most often text simplicity is defined as the antonym of text complexity Vecchiato (2022).

I define *text complexity* and *text difficulty* according to Mesmer et al. (2012) as follows: While *difficulty* refers to the comprehension performance of the readers in a text, *complexity* refers to measurable factors of a text to describe its difficulty for the readers. The higher the complexity of a text, the higher the required literacy skill. The complexity of a text can tackle the lexical, syntactical, or conceptual level. However, in my thesis, I focus on the lexical and syntactical complexity of a text and neglect the conceptual complexity of the text due to the lack of research available in this field (Eschenbruecher, 2021). In addition, readability measurement or analysis of writing style, soundness, structure, or choice of words are a few ways to assess syntactical and lexical complexity. Using these features, the complexity of texts in the same language can be compared and labeled.

2.1.4 Simplification Operations & Simplification Rules

I will use the term *simplification operations* to denote rewriting strategies that cause a lower complexity of a text compared to the original text. In more detail, simplification operations are rules regarding how to rewrite a complex linguistic feature into a simpler version (Cardon and Bibal, 2023), e.g.,

1. instead of using long one-token compound nouns, visually segmenting them into its components,
2. instead of using complex syntactic structures with nationalizations, rephrasing the text with easier, verbalized structures, or
3. instead of using rare words, replacing them with more frequent synonyms of the basic vocabulary.

Simplification operations are not to be confused with *simplification rules* even if both are closely related. While the second describes guidelines or a set of rules for writing simple, which denote which writing styles are permitted in a simplified language, the first refers to strategies on how to rewrite a complex linguistic feature into a more simple version. Simplification operations, therefore, aim at intralingual translation, whereas guidelines for simplified languages aim at directly writing texts in simplified language (without translation from a complex text). Nevertheless, simplification guidelines sometimes also include instructions or examples on how to rewrite complex texts into simpler texts.

2.1.5 Language & Literacy Levels

Previously, I described that complexity is subjective and that readers can have more or less problems learning a language or reading a text. Differences in comprehensibility of a text by a person can be due to their prior knowledge, language experience, and literacy. The term *content literacy* focuses on the ability for the acquisition of unknown texts by reading and writing, which includes, e.g., literacy skills and prior knowledge of a topic (McKenna and Robinson Richard D., 1990). According to Park (2012), content literacy includes three interdependent levels: a) find & filter relevant content, b) understand & reflect about the content, and c) form & post opinions about this and related content. In the scope of this work, I focus on understanding and reflecting on the content. A well-established framework for measuring the literacy of Germans introduces literacy levels called alpha levels (Grotlüschen et al., 2020).

Compared to content literacy, the term *language skills* describes the abilities required to read, speak, or write a foreign language. The Common European Framework of Reference for Languages (CEFR) is a framework containing foreign language levels (Council of Europe, 2020).

Alpha Levels The alpha levels are six levels that describe the literacy of adults living in Germany based on a large-scale survey called level one (LEO) (Grotlüschen et al., 2020; Grotlüschen et al., 2020).[1] Alpha levels of 1, 2, and 3 indicate low literacy, which means people with level 1 or 2 are able to read individual words but no full sentences. At alpha level 3, people can read full sentences but no full texts. On alpha level 4, people can slowly read full texts, but sometimes read them incorrectly. Following the latest LEO results of 2018, 12% of the German adult population corresponds to alpha level 1, 2, or 3, and at least another 20% to level 4. These numbers underscore the need and importance of both language varieties for German society.

CEFR Levels The CEFR contains a set of skills for listening, reading, speaking, and writing a language (Council of Europe, 2024). Based on the acquired skills, a language learner can be graded or self-assessed into 3 groups with each two levels. The groups can be named as follows: A) basic user (A1: Breakthrough, A2: Waystage), B) independent user (B1: Threshold, B2: Vantage), and C) proficient user (C1: Advanced, C2: Mastery). CEFR levels are also often used as a label on texts to describe which level is at least required to understand the text. As the CEFR levels are more known than the terms of plain language and easy language, the texts

1 The questionnaire of this survey is based on a questionnaire of the Program for the International Assessment of Adult Competencies (PIAAC) (Reder, 2017).

in these varieties are also often labeled with CEFR levels even if they are not designed for this purpose due to the different target groups (broad target group vs. language learner) (Bock and Pappert, 2023).

Text Complexity Labels As previously mentioned, CEFR levels and alpha levels are mainly developed to describe or categorize the reading and writing skills of people, these schemata are nowadays also used as labels for text complexity to make it easier for the reader to find the right text corresponding to their reading skills.

Following first attempts to align CEFR levels and simplified German language varieties, German Easy Language is labeled between A1 and A2 and German Plain Language between B1 and B2 (Oomen-Welke, 2015; Bock and Pappert, 2023; capito, 2024). However, in other works, German Easy Language is described as similar to texts in CEFR level A1, whereas German Plain Language is described as A2 or B1 (Helmle, 2017; Klar & Deutlich – Agentur für Einfache Sprache, 2018).

Following Bock and Pappert (2023), alpha levels would be more suitable for differentiation into complexity levels than CEFR levels, but CEFR levels are now more frequently used for that. Unfortunately, currently, no detailed analysis exists in order to align alpha levels with German simplified language varieties, e.g., German Plain or Easy Language. I assume that people with alpha levels 1 and 2 can benefit from German Easy Language and people with alpha levels 3 and 4 (or higher) can benefit from German Plain Language.

2.1.6 Clear, Plain, Simple, Easy, and Simplified

In the scope of text simplification research, the words *clear*, *easy*, *plain*, or *simple* are often used interchangeably to describe the low complexity of a text (Vecchiato, 2022). One reason for this might be its similarity and synonymity as shown in the thesaurus of Merriam-Webster.com (2024). However, in the following, I try to order the terms regarding their simplicity extent.

Clear Compared to the terms mentioned above, *clear* is the adjective that refers to the most complex texts. Clear writing aims to be "brief, simple, comprehensible and concise. Precision, on the other hand, refers to an exactness of expression, to an absence of ambiguity, an attempt to reduce contestability" (Coleman, 1998, p.393). Clarity and precision are, therefore, characteristics of plain language (Coleman, 1998). But, they do not describe a simplified language variety in more detail, as they do not address a specific target group with limited literacy skills.

Plain The term *plain* is used in the context of the plain language movement to describe a simplified language variety. The plain language movement started in the 1970s and has grown since then (Maaß, 2020). Following International Plain Language Federation (2024, p. 1), plain language is defined as follows:

> "A communication is in plain language if its wording, structure, and design are so clear that the intended readers can easily find what they need, understand what they find, and use that information."

In contrast to clear writing, plain writing also considers the lexic, syntax, and design of a text to make it better readable and comprehensible. In addition, plain language is sometimes also called *simple language* (Vecchiato, 2022).

Easy In this comparison, *easy* or *easy-to-read* refer to the lowest complexity and the easiest comprehensibility. In contrast to the previous adjectives, if a text is described as "easy" or "easy-to-read" it also addresses people with intellectual disabilities. Besides most simply written texts, it also considers easy-to-read typographic features such as font, font size, spacing, or contrast between font and background. In addition, easy-to-read texts are often enriched with images to emphasize the content of the text (Inclusion Europe, 2024).

Simplified The previous terms are also used to describe complexity-reduced language varieties, for example, plain language or easy language. In contrast to the previous adjectives, *standard*, in this context, refers to the standard level of language or the everyday language. Following (Spring et al., 2021), I am using the term *simplified language* as an umbrella term comprising complexity-reduced language varieties and texts that are simplified for children, foreign language learners, or lay people.

2.2 Automatic Text Simplification

Based on the previously denoted terms and knowledge, I can now define the task of *text simplification* (TS). I refer to *text simplification* as the process of making a text simpler, i.e., adapting the style of a text and reformulating it in a way that it is better comprehensible by another target group, e.g., children, laypeople, foreign language learners, or people with cognitive impairments. This process includes several rewriting strategies (further called *simplification operations*) or simplification rules on different linguistic levels, e.g., morphological, syntactical, semantic, or lexical (Vecchiato, 2022). To make a text easier to read for a target group, these strategies and rules can be used to rewrite the original text, e.g., by writing more coherently, writing more explicitly, using basic vocabulary, or using basic grammar. The strategies can be applied manually or with machine learning strategies.

Manual simplification, or intra-lingual translation, of texts into German Plain or Easy language is a complex and time-intensive task (Maaß, 2015a). To enhance the manual simplification process and to assist professional and non-professional translators (as public authorities), machine-generated simplifications can be used as a starting point for a simplification (Hansen-Schirra et al., 2020b; Anschütz et al., 2023; Carrer et al., 2024). The task of machine-generated simplification is also called "automatic text simplification" and is a trending research topic in computational linguistics (see Shardlow 2014; Stajner 2021; Ryan et al. 2023).

Automatic text simplification refers to the process of modifying written texts with the help of machine learning to make texts more accessible and better understandable to a wider audience, particularly for people with limited reading abilities, for example, children, people with cognitive impairments, or people learning a new language (Siddharthan, 2014). The primary aim of text simplification is to improve the readability and comprehensibility of a text by changing the wording or structure of a text but still retaining the meaning of the original text (Alva-Manchego

et al., 2020b). Common strategies of simplification are, for example, substitution of complex words with simpler synonyms or explaining their meaning, removing redundant words, splitting long sentences into several shorter ones, or restructuring sentences (Alva-Manchego et al., 2020b).

2.2.1 Text Units of Simplification

In contrast to professional manual translation which mostly performs holistic translation of full documents, ATS research has long focused on building sentence simplification models over document simplification models (Alva-Manchego et al., 2020b), which is among other things due to research in previous decades in which there was less computing power to process whole documents (same as for machine translation). However, in modern text simplification three text units are common, i.e., document simplification (e.g., see Rios et al. 2021), paragraph simplification (e.g., see Devaraj et al. 2021; Trienes et al. 2022), and sentence simplification (e.g., see Mallinson et al. 2020; Ebling et al. 2022).

In addition to the word and sentence level, text simplification can be applied to the discourse level. Even if the task is named *text* simplification, most approaches focus on simplification within a sentence. Although, according to Alva-Manchego et al. (2019b) discourse features, e.g., anaphora resolution, reordering, or joining sentences can help to make a text more coherent, less ambiguous, and easier to understand. But, yet, only a few papers address cross-sentence simplification, e.g., Siddharthan (2003) or Alva-Manchego et al. (2019b).

2.2.2 Text Simplification Workflow

The automatic text simplification process can be expressed in a workflow or pipeline that is similar to the one of machine translation. As previously already stated, the text simplification pipeline (including manual, computer-assisted, and automatic simplification) contains three main parts (Garbacea et al., 2021): i) classifying whether a text should be simplified (complexity prediction) and then identifying complex passages in texts (complexity explanation) (see component 0 in Figure 2.1), ii) generating a simplification (see component G), and iii) validating of the generated simplification (see component H).

However, I adapt this (more linear) pipeline to a more dynamic workflow for our purposes. Therefore, I specify this general workflow with a focus on automatic simplification by extending the components regarding the generation and evaluation parts and adding the corpus building pipeline. Once more, a visualization of the workflow diagram for automatic text simplification is provided in Figure 2.1. In general, the process can be divided into the training and test phase (upper and lower parts), as well as the data and model parts (left and right parts).

The main components, which hold also for machine translation, are: using a parallel corpus (see components B and E in Figure 2.1) to train (see component G) and evaluate (see component H) a text simplification model (see component G). In order to optimize the model for the given development data, the evaluation can be seen as a loop (from component G to H to G): a model is modified regarding some settings (e.g., hyperparameters or prompts), again trained on the same data, and the new system generations are evaluated again to verify whether the model has improved in comparison to the previous run.

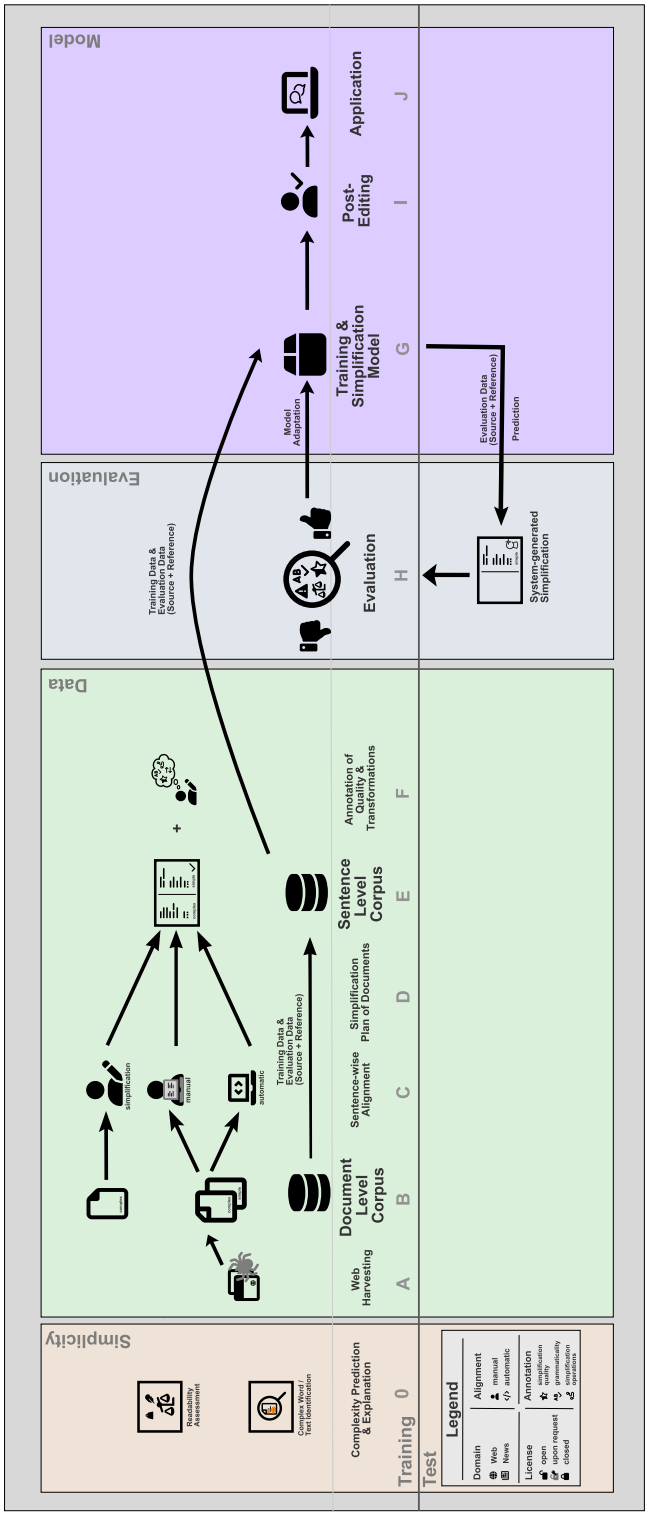


Figure 2.1: Workflow diagram of text simplification with the corpus building process (same as Figure 1.1).

Overall, the TS workflow is characterized by a high degree of flexibility and adaptability at numerous points, e.g.,

i) text simplification can be performed on document or sentence level; hence, for document simplification, component C to E could be skipped and the parallel documents (component B) can be directly used as input for a text simplification model (component G),

ii) a few components are optional or not considered in all TS studies, e.g., generating simplification plans (component D), manual annotation of quality (component F), post-editing (component I), or application (component J),

iii) findings regarding word or sentence complexity (component 0) can be either used to identify complex and simple data for building TS corpora or as additional information for the TS system whether a word or a sentence required simplification,

iv) findings from an initial component can influence later components, e.g., simplification operations found during manual quality estimation (see component F) can influence the choice of TS model architecture (see component G) or how to evaluate the system outputs (component H), or

v) the whole workflow can be repeated several times in order to find the best performing TS model, e.g., the selection of data wrt. size, domain, simplification extent, or simplification operations (components A to F), can influence the quality of a TS model. Also information from post-editing (component I) can be integrated as human feedback into the training process of a TS model (component G, "human-in-the-loop" approach).

2.2.2.1 Data & Corpus Building Workflow

Text simplification can be applied to different text units, e.g., document (see component B in Figure 2.1), paragraph, or sentence level (see component E), depending on the complex-simple pairs of the parallel corpus. However, the process of building a parallel corpus differs depending on the chosen text unit: for document simplification, it involves fewer components (see component A to B) than for sentence simplification (see component A to E). Both have in common the ability to harvest some data from the Web (see component A) or manually simplify it (see component C) and then align the documents (see component B). The aligned documents can also be aligned on the paragraph or sentence level (see component C to E) before using them for training the TS model.

In Chapter 3, I will go into further detail on how to build TS corpora and, in Chapter 4, I will compare the corpora built for German TS.

2.2.2.2 Text Simplification Models

ATS models (component G in Figure 2.1) are often data-driven approaches that are trained with aligned complex-simple text pairs to learn how to apply simplification operations based on the provided examples (Alva-Manchego et al., 2020b). Depending on the training data, the simplification models can be focused on specific tasks, for example, if the sentence pairs contain mostly syntactic changes, the model is most likely to perform only those operations.

However, most corpora contain sentence pairs that apply lexical and syntactical changes at the same time; sometimes the change can be assigned to both levels. For example, in Example 1, the complex word "veröffentlichungspflichtig" (EN: required to be published) is in Example 2 and 3 verbalized into "veröffentlichen müssen" (EN: must publish) which can be seen as a lexical change in case of a word replacement and also as a syntactical change in case of restructuring the sentence by using the word as predicate.[2] Furthermore, due to a lack of parallel corpora with aligned sentence pairs regarding only syntactic or lexical simplification or only simplifications of one target group, ATS models are often very general and not yet directly helpful for the target group (Alva-Manchego et al., 2020b).

In Chapter 6, I will go into more details about text simplification models for German.

2.2.2.3 Evaluation of ATS

Manual or automatic evaluation is required to measure the quality of the generated simplifications (component H in Figure 2.1). A good simplification should be grammatically correct, simpler, and better readable than the original text, and preserve the original meaning of it. For manual evaluation, people are asked to rate the extent of these three aspects for the generated simplification with respect to the original sentence. Manual evaluation is very time-consuming; hence, for a first quality check, automatic metrics are used for the evaluation of sentence simplification models. To the best of my knowledge, there are no metrics for evaluating document simplification yet (except for an adaptation of SARI, i.e., D-SARI).

Readability metrics, such as FKGL (Flesch, 1948), are utilized to measure simpleness and other NLG metrics, such as BLEU (Papineni et al., 2002) or BERTScore (Zhang* et al., 2020), for meaning preservation. To measure the overall simplicity quality, SARI (Xu et al., 2016) is the primary metric. SARI compares a generated simplification sentence with the source sentence and several references to estimate the quality of the lexical simplification. For the evaluation of syntactical simplification, SAMSA (Sulem et al., 2018b) was proposed, which is a reference-less metric based on annotations of semantic structures. LENS (Maddela et al., 2023) is a newer method to evaluate ATS models; it is a metric trained on human assessments. Following (Alva-Manchego et al., 2021), BERTScore can also be used to measure overall simplicity even if it was not implemented for this use case.

In Chapter 5, I will go into more detail about the evaluation of German text simplification.

2.2.2.4 Complexity Prediction & Explanation

Following Garbacea et al. (2021), one part of the TS workflow, which is often neglected, is to identify whether a text is too complex for a target group and whether it should be simplified. If so, the parts of the text that make it complex should also be identified. The task of complexity identification is also called readability assessment, and the task of identifying complex parts on the lexical level is, e.g., complex word identification (see component 0 in Figure 2.1).

Regarding readability assessment of German texts, there are some research works, e.g., Suter et al. (2016), Weiss and Meurers (2022), Klepp (2022a), or Seiffe et al. (2022). With a simi-

2 For typologies of simplification operations, see (Cardon et al., 2022) for English, Brunato et al. (2022) for Italian, (Stodden and Kallmeyer, 2022) for German, and (Cardon and Bibal, 2023) for a multi-lingual overview.

lar research aim in mind, but focusing more on simplifying German sentences, Mohtaj et al. (2022) have organized a shared task related to evaluating the text complexity of German content. Specifically, the participants are required to predict a mean opinion score, which essentially represents the average assessment of a German sentence's complexity by non-native German speakers. This task can also be described as predicting the subjective complexity level of a sentence for a particular target group, in this case, non-native German speakers.

Yet, there is less research on German complex word identification (CWI). A few shared tasks have been performed on CWI or lexical complexity prediction (LCP) of words or sentences, e.g., CWI shared task 2018 (Yimam et al., 2018), LCP shared task 2020 (Shardlow et al., 2021), or MLCP shared task 2024 (Shardlow et al., 2024)[3]. Only the first and the latter contain German data; the other just focuses on English LCP. It is unclear whether the approaches regarding English LCP can also be transferred to German.

2.2.3 Subtasks of Text Simplification

In literature, the following subtasks of text simplification are named: lexical simplification, syntactical simplification, and conceptual simplification.

Lexical Simplification Lexical simplification focuses on the change in vocabulary to make a sentence better readable; it includes the processes of identifying complex words, finding simpler synonyms for them, and replacing complex words with simple alternatives within the sentence (Shardlow, 2014). Hence, automatic approaches of lexical simplification are often split into the following subtasks:

1. complex word identification (or lexical complexity prediction),
2. easier but similar candidates are generated, often with context-aware word embeddings (Glavaš and Štajner, 2015; Gooding and Kochmar, 2019),
3. select the best fitting candidate for the substitution (Paetzold and Specia, 2016b), and
4. integrating the substitute with correct inflection in the original sentence.

Syntactic Simplification In contrast to lexical simplification, syntactic simplification (also called structural simplification) can be processed within one sentence or across several sentences. Syntactic simplification is the task of simplifying the grammar and structure of a text. It contains, for example, the splitting of a long sentence into its component clauses, merging a few sentences into one sentence, or rewriting passive to active (Shardlow, 2014). In contrast to lexical simplification, syntactic simplification systems are often rule-based and not data-driven. Based on a dependency parse tree (Niklaus et al., 2019a) or a constituency parse tree (Paetzold and Specia, 2013) of a sentence, rules are hand-crafted and applied to sentences to identify subclauses and move, delete, or separate them into a new sentence.

3 Unfortunately, the results regarding the LCP shared task 2024 have not been published during writing the thesis.

Conceptual Simplification Conceptual simplification comprises the simplification of a complex concept which will be elaborated within the original sentence or by adding a new sentence (Gooding, 2022). An example of a complex concept is a complex word whose naming is important or for which there is no simple synonym. Another type of conceptual simplification can be the simplification of implicit structures that are elaborated to be more straightforward.

2.2.4 Simplification Purposes

Text simplification can be performed for many different purposes, e.g., from standard texts to texts for foreign language learners, from standard text in a simplified language variety, as well as from expert language to standard language (Shardlow, 2014). In order to understand the diversity of the purposes (see Subsection 2.2.4), I first describe different (simplified) language varieties (see Subsubsection 2.2.4.1).

2.2.4.1 Language Varieties

The term "*language variety*" describes a language that is used by people of a particular group or in a particular situation (Southerland and Katamba, 1997), e.g., standard language, technical language, or simplified language. Examples of simplified language varieties are plain language and easy language.

Texts written in different language varieties have different aims, different purposes, are addressed to different targeted readers, are written by different writer groups, and are written in different text styles. In more detail, technical language is written by experts for experts, and hence is the most complex variety; examples of technical texts are medical texts, legislative texts, or scientific texts. Standard language refers to the language learned in school and texts of everyday life written for a broad audience, e.g., governmental texts, news articles or blog posts (Southerland and Katamba, 1997). In contrast, plain and easy language are both complexity-reduced languages; texts written in these varieties are addressed to people with reading problems, e.g., foreign language learners or people with learning difficulties. The aim of these texts is similar to texts in standard language, which pass on information, but in this case they are written with the special needs of the target group in mind. As discussed previously in Subsection 2.1.6, a text written in an easy language is less complex than a text written in plain language (Maaß, 2020). For an overview of the characteristics and differences of German technical language, standard language, plain language and easy language see Table 2.1.

Focusing more on simplified language varieties: The different needs of people with divergent knowledge and literacy cannot be satisfied with the same simplified sentence, so there is no perfect simplification that suits all people (Siddharthan, 2014). As discussed previously, comprehension and readability are subjective assessments based on the skills and needs of each person. However, people with similar skills can be grouped into target groups, e.g., people with cognitive impairments, native speakers with low reading skills, or foreign language learners. For example, most language learners can understand complex concepts and comprehend the content of complex statements, but could have problems with the grammar in the foreign language, here German (Marzari, 2010). In contrast, people with cognitive impairments, or

	Technical Language	Standard Language	Plain Language	Easy Language
Target Group (Reader)	Experts	People with average literacy and language skills	all those for whom texts in standard language are too difficult	people with learning difficulties and little knowledge of German
Target Group (Writer)	Experts (e.g., people trained on medicine or law)	People with average literacy and language skills	all those who want to write texts that are easy to understand	trained translators
Levels of Simplification	*n/a*	*n/a*	choice of words, sentence structure, text structure, style	choice of words, sentence structure, text structure, text design
Characteristic	precise and full of technical terms	long sentences, complex words, overall rather complex language	Short but concise simplification of texts	strong reduction in content and language, review necessary
Aim	passing on information on a specific topic from expert to expert	passing on information	Simplify (technical) language for laypeople but retain linguistic style and produce correct language	Create strongly simplified texts with a wide reach
Literature			de Oliveira (2016), Baumert (2018), Maaß (2020)	Bredel and Maaß (2016), Bock (2019), Maaß (2020)

Table 2.1: Language varieties.

with learning difficulties, have a problem comprehending complex words and several statements per sentence (e.g., see Bredel and Maaß 2016 or Maaß 2020). Therefore, a simplification should always be written with the target group and their skills in mind, e.g., by following recommendations of writing in a simplified language variety.

In Germany, two main comprehensibility-enhanced language varieties exist, i.e., German Plain Language (DE: "Einfache Sprache") and German Easy Language (or easy-to-read German; DE: "Leichte Sprache") (Maaß, 2020). The target group of German Easy Language includes people with learning difficulties, people with dementia, people who do not speak German very well, and people who cannot read very well (Netzwerk Leichte Sprache, 2022). German Plain Language is instead addressed to people with more language knowledge or more language skills, e.g., people with reading problems and German language learners (Maaß, 2020). Due to the divergent reading and comprehension skills of the people in these target groups, the simplification into German Plain Language contains a lower reduction of complexity than the simplification into German Easy Language. Following this, the texts in German Easy Language are simpler than the texts in German Plain Language.

To describe, define, and to some extent also control the language, many guidelines and recommendations on how to write texts in (German) Plain Language or (German) Easy Language exist: For example, Inclusion Europe (2021) provides an Easy Language standard with recommendations for 18 European languages, including German and English. In addition to the multilingual guidelines of Inclusion Europe, many language-specific information and guidelines exist, for German Easy Language, e.g., see: Bredel and Maaß (2016), Netzwerk Leichte

Sprache (2022), Bock (2019), or Deutsches Institut für Normung (DIN) (2023). Or for German Plain Language: Baumert (2018), Maaß (2020), Deutsches Institut für Normung (DIN) (2024b), Deutsches Institut für Normung (DIN) (2024a). Common simplification operations are, e.g., rewriting passive to active voice, splitting relative clauses into distinct sentences, explaining complex terms, or replacing digits with number words.

However, the same complex German sentence (see item 1) can be rewritten in the different varieties to address the special needs of the different target groups, e.g., in German Plain Language (see item 2) and in German Easy Language (see item 3). Therefore, different linguistic operations can be applied to the complex sentence that are described in more detail in the guidelines of the language varieties (see above). For German Easy Language, the complexity of sentences is reduced to a minimum by applying several simplification operations at once until no further simplification is possible.

(1) *Standard German:*
Veröffentlichungspflichtig sind alle Informationen, die bei den Behörden vorliegen.
'Required to be published are all information, which is available with the authorities.'

(2) *German Plain Language:*
Alle Informationen, die den Behörden vorliegen, müssen veröffentlicht werden.
'All information, which is available to the authorities, must be published.'

(3) *German Easy Language:*
Die Behörden haben viele Informationen. Die Menschen möchten mehr über die Informationen von den Behörden wissen. Deshalb müssen die Behörden alle Informationen veröffentlichen. Veröffentlichen ist ein schweres Wort. Es bedeutet: Etwas offen zeigen.
'The authorities have a lot of information. People want to know more about the information from the authorities. That is why the authorities must publish all information. Publish is a difficult word. It means to show something openly.'

Example 2 is simplified to German Plain Language. To make the sentence more readable for foreign language learners or people with reading difficulties, the adjectivization and nominalization of the complex adjective "veröffentlichungspflichtig" is rephrased into several easier-to-understand words and used as the new predicate of the sentence ("müssen veröffentlicht werden", EN: "must be published". During the simplification process, the structure of the sentence is simultaneously changed, and the clause is moved from the end of the sentence to the middle of the sentence.

Many more linguistic operations have been applied to Example 3 during simplification to enhance comprehensibility, for example, for people with cognitive impairments: content addition by making implicit information more explicit (first two sentences), explaining the complex term "veröffentlichen" (EN: to publish) (last three sentences). The third sentence contains the main content of the original sentence, but the complex adjective "veröffentlichungspflichtig" is rephrased into a verb phrase "müssen …veröffentlichen" (EN: must publish), the passive voice is changed to active voice by naming the grammatical agent ("die Behörden", EN: the authorities), and removing the clause.

Quality Control of Manually Simplified Texts Following the guidelines on how to write in German Easy Language, (professionally) simplified texts in German Easy Language require that the target group proofreads them before publishing (Maaß, 2015a). Hence, texts in German Easy Language have high usefulness and a high quality for the target group, which can also enhance the quality of data-driven text simplification systems.

As a short digression, I want to highlight the high quality of manually simplified texts in German. As previously said, following the German Easy Language ("Leichte Sprache") guidelines, the manually simplified text must be manually evaluated by the target group (Maaß, 2015a; Schiffler, 2022). If the target group still identifies issues with the simplified text, the translator has to rewrite the text. This iterative process continues until the target group can no longer identify any issues in the simplified text. Proofreading also has the benefit that the target group is more involved in the process of the simplification process, so that they can decide for themselves whether a text is easy to read and what they identify as complex (see paternalism issue, e.g., Gooding 2022). There are ongoing discussions whether texts in German Easy Language should be proofread or not. On the one hand, Maaß (2015a, p. 164ff) argue that professionally trained translators are able to write adequate simplified texts for the target group, and therefore proofreading would not be required. On the other hand, the guidelines of German Easy Language enforce proofreading by the target group, although they do not name specific reasons why or define the process (Schiffler, 2022, p. 19ff).

2.2.4.2 Target Groups

Furthermore, the complexity-reduced language varieties can be mainly divided by their addressed reader groups (see the first row in Table 2.1) which define the characteristics of the language based on the needs of the reader group. The target group of German Easy Language includes people with learning difficulties, people with dementia, people who do not speak German very well, and people who cannot read very well (Netzwerk Leichte Sprache, 2022). German Plain Language is instead addressed to people with more language knowledge or more language skills, e.g., people with reading problems, elderly people, German language learners (Baumert, 2018). Therefore, the size of the target groups also differs.

Following Auswärtiges Amt (2020), in 2020, 15.45 million people around the world have learned German as a foreign language. Grotlüschen et al. (2020) found in their literacy study in 2018 that 10.6 million German adults can be classified with alpha level 4, and 4.2 million with alpha level 3. Furthermore, in the year 2022, 18.6 million people have been 65 years old or older (rising tendency) and could also benefit from German Plain Language (Statistisches Bundesamt, 2023). Hence, 14.8 million people with lower literacy, 15.45 million foreign language learners, and 18.6 million elderly Germans could benefit from texts written or simplified into German Plain Language.

Following Maaß (2020), the target group of German Easy Language in Germany comprises 400,000 to 800,000 people with cognitive disabilities, 1.3 million people with dementia, partially 80,000 people with pre-lingual hearing impairments, and 130,000 to 240,000 people with aphasia. However, the numbers cannot be summed to a total number of people in the target group, as some people might refer to more than one of these groups. Furthermore, people who are not

directly addressed in the target group can also benefit from German Easy Language, such as foreign language learners.

In an ideal world, many texts could be directly prepared in plain language so that they can be understood by many people. However, as previously discussed, the complexity of texts is highly subjective because of the foreknowledge and experiences of a reader. Furthermore, many texts have already been published in only complex versions; hence, simplification can help to produce a new variant of the text which is better readable.

2.2.4.3 Simplification Purposes

Text simplification can also be seen as an intra-lingual translation task in which texts are translated from a language variety into another variety of the same language (Hansen-Schirra et al., 2021), e.g., standard German to German Plain Language or German texts with jargon into standard language.

Now, focusing on simplification rather than writing in simplified varieties, each text (except texts written in easy language) can be considered as complex (or to-be-simplified) and can be simplified into any variety with lower complexity. In Table 2.2, I summarize the simplification purposes from technical to standard and to complexity-reduced varieties (see Table 2.2a), from standard language to texts better comprehensible for language learners with different skills (see Table 2.2b), from standard language to texts suitable for children of different ages (see Table 2.2c), and from standard language into complexity-reduced varieties (see Table 2.2d).

The longer the bars are in Table 2.2, the more strong the simplification is, e.g., a simplification from technical language to easy language is expected to be the most strong simplification, whereas simplification within one CEFR step (e.g., from A2 to A1 or B1 to A2) is expected to be the most mild simplification.

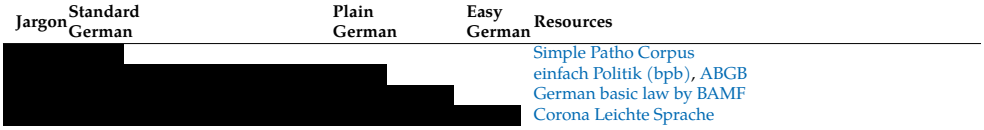
Each simplification purpose further requires special simplification operations: While it might be sufficient in expert-laypeople simplification to explain technical terms, in standard-children simplification many more operations are necessary, e.g., reducing the number of used tenses, replacing rare terms with simpler terms, and making the content more explicit. In order to include these factors in automatic simplification, parallel resources for this purpose are required (see the last column in Table 2.2).

2.2.5 German Simplification

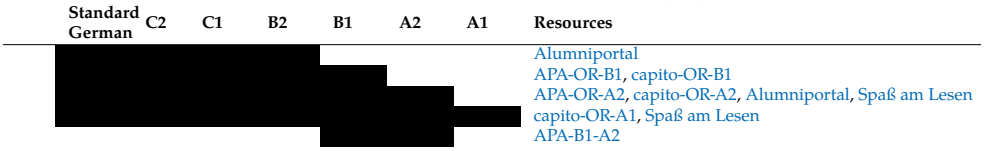
In the previous sections, I have mostly described a general overview regarding complexity and text simplification. In this section, I will focus more on the complexity of the German language and German texts. I discuss if and to what extent the German language and German texts are complex in order to motivate more why one could benefit from German text simplification.

2.2.5.1 Complex German Language – Linguistic Complexity

There is a debate on whether languages are all equally complex or if they can be ranked according to their complexity (Joseph and Newmeyer, 2012). Nevertheless, in the following, I present some research results on the linguistic complexity of the German language.

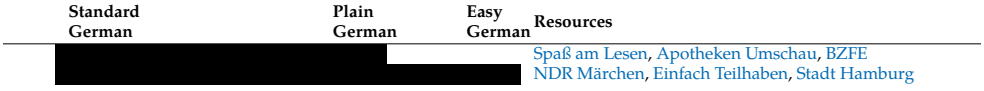


(**a**) Simplification of German technical language.



(**b**) Simplification for foreign language learners.

(**c**) Simplification for youth and children.



(**d**) Simplification into simplified German language varieties.

Table 2.2: Simplification purposes. The length of the bars indicates the degree of simplification. All URLs have lastly been accessed at July 24, 2024.

Marzari (2010) evaluated the seven European languages with the highest amount of speakers in Europe (i.e., Spanish, English, French, Polish, Russian, and German) regarding their complexity for language learners. The analysis includes linguistic features regarding grammar, lexis, and text composition. He found that out of these languages, German has the most complex text composition, most complex lexis (shared with Russian), and a more complex grammar than English, Spanish, Italian, and French. On the one hand, he argues that German is difficult to read due to idiosyncrasies in the alphabet (i.e., umlauts and case sensitivity), in the basic vocabulary (i.e., mostly Germanic), in the morphology (i.e., inflection of adjectives), in the sentence construction (i.e., participle constructions and nested sentences), and in the textual structure (i.e., subject-object-verb order in clauses and less linearity) (Marzari, 2010). On the other hand, following Marzari (2010), German is easy to read due to the writing system (i.e., in total 26 Latin characters + umlauts) and medium complexity of the conjugation (less forms than in French, Italian, Spanish, or Polish) and usage of tenses and mood (i.e., only conjunctive being complex). However, he also argues that this complexity is subjective due to foreknowledge in other languages, e.g., for native English speakers, the basic vocabulary would be easier than for native speakers of a Romance language.[4]

4 In the analysis, always the highest complexity value is considered even if it may be easy for some language learners.

Dammel and Kürschner (2008) analysed morphological complexity of German and compared it to other Germanic languages. Following them, the German morphology is more complex than the morphology of the other languages due to stem involvement, number of allomorphs, direction of determination, but less complex in terms of the fusion of number and case (Dammel and Kürschner, 2008).

Following these results, simplification of German texts seems to be useful in making texts easier to read for foreign language learners.

2.2.5.2 Complex German Texts – Text Complexity

Even though there are languages that are more complex (but also less complex) than German, still many people have trouble reading or writing German (Maaß, 2020; Grotlüschen et al., 2020; Baumert, 2018). By nature, technical language texts are complex in any language, as they are written by experts for experts (Baumert, 2018). For example, law texts, scientific texts, political texts are in general more complex than other texts because they are written for an audience with specific expert knowledge. In contrast, newspapers are generally more simple than the texts in the other domains named before because news texts are written for a broad audience without expecting foreknowledge (Bamberger and Vanecek, 1984).

Due to different literacy and language skills, as well as different foreknowledge and experiences, readers comprehend texts differently; therefore, text comprehension is very subjective (Bock and Pappert, 2023). However, in order to reduce the cognitive processing cost when reading texts, texts can be rewritten in simpler language, i.e., via intra-lingual translation or simplification (Hansen-Schirra et al., 2020a). Therefore, linguistic phenomena have been identified that are often described as complex, e.g., long compound nouns, nested sentences, or passive voice. In order to avoid these phenomena, guidelines have been written to write more simply; for example, they include: splitting nested sentences into several sentences, visually segmenting one-token compounds, or writing in active instead of passive voice (e.g., see Netzwerk Leichte Sprache 2022, Baumert 2018, or Maaß 2020).

Unfortunately, research is missing whether and to what extent texts in German Easy or Plain Language actually comply with these rules and guidelines. Furthermore, it is unclear how the applied simplification operations differ between languages, e.g., between English and German.

2.3 Summary & Outlook

In summary, in this section I have provided background knowledge regarding complexity, simplicity, and text simplification. I have motivated why and how German texts could be simplified with respect to different simplification purposes and simplification operations. Currently, there is a research gap between manual and automatic simplification and, hence, it is unclear which simplification rules (in the form of simplification operations) have been applied during the manual and automatic simplification of German texts. More research is required on the definition and differences of operations in the scope of German simplification (see RQ 2-1).

Furthermore, I have also briefly introduced the automatic text simplification workflow and have identified research gaps regarding the first part of the simplification process, i.e., complex-

ity identification and prediction. In text simplification research, it is often overlooked whether a text requires simplification or if it is already simple enough for a specific target group. To advance the integration of complexity prediction in TS research, more research is necessary regarding how to automatically identify complex text passages (see RQ 2-2).

In the following, I describe the state of research regarding the other parts of the simplification workflow. Therefore, I provide more details on the construction of TS corpora (see Chapter 3), existing TS resources (Chapter 4), evaluation methods (Chapter 5), and models (Chapter 6). In each of these sections, I will present the current state and discuss current challenges.

Chapter 3

Building Text Simplification Corpora

Such as in all machine learning tasks, in the field of automatic text simplification, also a data collection (also called corpus) is required to evaluate and/or train automated text simplification systems. A corpus for text simplification is a monolingual data collection that always consists of a collection of pairs, where one side of the pair consists of a *complex text* in one language and the other side consists of at least one simplified version of these texts (called *simplified texts*) in the same language. Ideally, both sides of a simplification pair contain the same content but differ in their complexity level.

Building complex-simple pairs (or parallel TS corpora) is as important for TS as for other natural-language generation tasks. Based on the complex-simple texts, a supervised machine learning model can learn how to generate a text close to the provided target simplification. Furthermore, (un)supervised models can be evaluated against TS corpora to estimate their quality.

In the remainder of this chapter, I am explaining relevant terms and concepts to build TS corpora, e.g., parallel corpora vs. comparable corpora (see Section 3.2), or one vs. many target simplifications (see Section 3.1). Furthermore, I am introducing the workflow for building text simplification corpora (see Section 3.3) including i) finding relevant data (see Section 3.4), ii) sentence-wise alignment (see Section 3.6 and Section 3.7), and iii) additional annotation regarding simplification operations and quality assessment (see Section 3.9). Before I conclude the chapter (see Section 3.11), I give a short overview on existing annotation interfaces which can support the building process (see Section 3.10).

3.1 One vs. Many Target Simplifications

As TS research is based on machine translation research, often the same terminology is used. Hence, the complex text can also be named as *source text* and the simplified text as *target text*.

For a source text, there exists not one perfect simplification, but several equally good target simplifications that are tailored to the individual needs of people (Siddharthan, 2014). Therefore, in TS research, target texts are also referred to as *reference texts* (Alva-Manchego et al., 2020b). As there is a lack of high-quality TS corpora in languages other than English, many non-English TS corpora contain one-to-one pairs with only one gold simplification for a complex sentence (Martin et al., 2023).

But, many English TS corpora contain pairs with one complex text and many references. If more than one reference is available, a generated simplification can also be evaluated against more possible reference simplifications and not only against one idealized target simplification.

3.2 Comparable vs. Parallel Corpora

As previously mentioned, the collection of complex-simple pairs is called a TS corpus. However, a TS corpus can be further described by its degree of meaning overlap between both sides of the pair, i.e., *parallel* or *comparable*.

Following Teubert (1996), a corpus is called *parallel corpus* if a complex text at hand is directly translated or simplified into the corresponding version in another language or language variety.[1]. In contrast to parallel corpora, comparable corpora do not contain direct translations (or simplifications) of the source texts; they comprise texts in another (or simplified) language with similar but not necessarily exactly the same content (Teubert, 1996).

For a better illustration, if two individuals independently write a text on the same topic X, it is called a comparable pair: Person A posts a complex-to-read text with topic X on Wikipedia. Person B publishes a text in simpler words on the same topic X on a Wikipedia-based page for children without knowing or considering the text of person A. Then this pair of a complex document by person A and the simple document by person B would be considered following the named definition as *comparable* as both texts might contain different insights even if both address the same topic X.

On the other hand, if person B would a) read the text by person A, b) would try to keep most of the content when rewriting it, and c) at the same time would try to simplify the wording and the structure of the original text, then, we would call this a parallel document pair, following my definition above.

With regard to text simplification, most resources are comparable resources as the complex and simple texts are written independently. Another indicator for this is that for a huge proportion of the document pairs it is not possible to identify or assign a simplified sentence to an origin in the complex document.

Comparable resources for German TS are, for example, the German version of Wikipedia and a Wikipedia version for children (e.g., Vikidia[2], Klexikon[3], or Miniklexikon[4]), or news articles written in simplified German without a corresponding news article in standard German (e.g., Deutschlandfunk[5] or NDR[6]).

In comparison, the Austrian News Agency[7] offers their news articles in three different complexity levels, all containing nearly the same content. In addition, the simplified texts of the

1 Different definitions for parallel and comparable corpora exist (see McEnery and Xiao 2007), however, I follow the definition of Teubert (1996) as it is easily transferable to text simplification.

2 https://de.vikidia.org/wiki/Hauptseite [last access: July 24, 2024]

3 https://klexikon.zum.de/ [last access: July 24, 2024]

4 https://miniklexikon.zum.de/ [last access: July 24, 2024]

5 https://www.nachrichtenleicht.de/ [last access: July 24, 2024]

6 https://www.ndr.de/fernsehen/barrierefreie_angebote/leichte_sprache/Nachrichten-in-Leichter-Sprache,nachrichtenleichtesprache100.html [last access: July 24, 2024]

7 https://science.apa.at/nachrichten-leicht-verstandlich/ [last access: July 24, 2024]

Hamburg Public Authority[8] are a direct translation of the text into standard German with low information loss. Therefore, the last two are examples of *parallel* resources for German TS.

3.3 Building Process

The creation of text simplification corpora most often contains the following steps or components:

A) Building a web harvester to download and harvest documents in standard German and simplified German,

B) Alignment of complex and simple documents to document pairs – resulting in a document simplification dataset,

C) Sentence-wise alignment (manually or automatically using some alignment algorithms),

E) Resulting in a sentence simplification dataset,

F) Adding some human annotations on the aligned sentence pairs for insights into the simplifications.

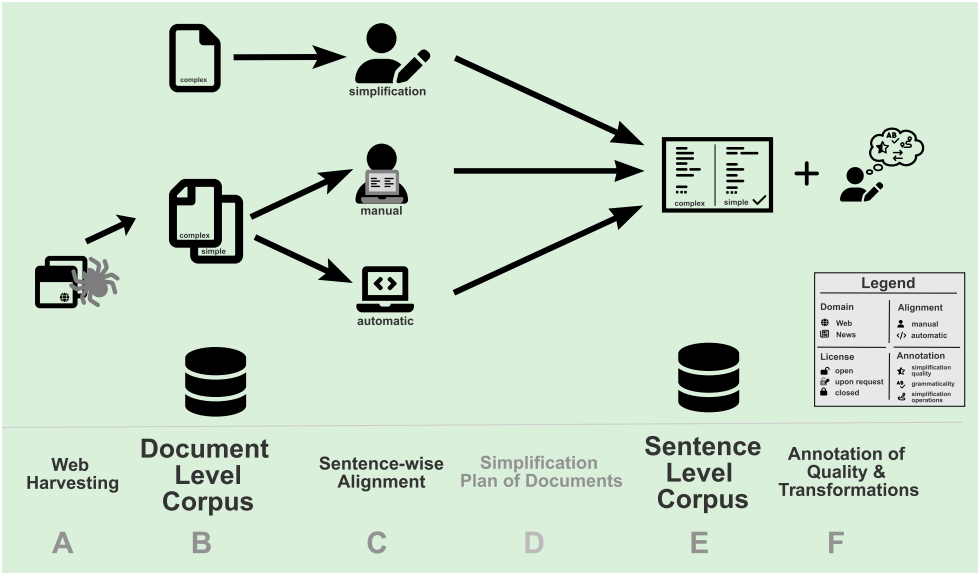


Figure 3.1: Corpus building process (Cutout of the whole TS workflow of Figure 1.1).

For a better overview, the process and the connections between the components are visualized in Figure 3.1. When creating parallel corpora, a distinction can be made between the text units of interest. For a document simplification corpus, components C and E would be skipped. For a paragraph simplification corpus, the same procedure can be applied as for a sentence simplification corpus (components A to F).

8 https://www.hamburg.de/barrierefrei/leichte-sprache/ [last access: July 24, 2024]

In the following, I will describe each step or component in more detail (see Section 3.4 - Section 3.9) and explain how to facilitate the process with digital interfaces (see Section 3.10).

3.4 Finding Suitable Data (Component A & B)

In order to build a TS corpus, comparable or better parallel resources are required in which one part is more complex than the other. A good origin for parallel corpora could be i) professional translators, who could share their simplifications as well as the original texts, ii) data providers, which offer their material directly, or iii) crawling the simplified and complex texts from web pages. A prominent example for type ii) is Newsela[9], a news agency that offers professionally simplified English news articles for scientific purposes. This resource has been used for the first time for automatic text simplification by Xu et al. (2015).

Following guideline 3.1 of the Web Content Accessibility Guidelines (WCAG) 2.0 (World Wide Web Consortium, 2008), the international standard for web content, texts on webpages should be understandable by people with at least nine years of school education. If the content is more complex, they recommend providing a simpler summary, providing visual illustrations, providing a spoken version, providing a version in sign language, or simplifying the text (at W3C, 2024). Based on these recommendations, many webpages contain texts in easily understandable language and for some webpages also parallel versions in standard language and simplified language are provided. Hence, in recent years, an increasing number of websites have offered texts that have been simplified both manually and professionally.

A good start to find relevant simplified German web pages are references listed on the web pages of professional translators or looking at public authority web pages because public authorities in Germany are obliged to offer easily readable texts on their web pages following the barrier-free information technology ordinance ("German Barrierefreie-Informationstechnik-Verordnung" (BITV, 2011). Nevertheless, I want to highlight the importance of the copyright of the complex and simplified data. Before downloading the data, I recommend carefully checking the copyright if it is permitted to download, use, and distribute the data.[10]

Furthermore, if simplified or even parallel texts are available on websites, but the website owners do not provide a bundled download, building and/or using a web scraper could be helpful to build a new TS corpus. Examples of simplified German text web scrapers are proposed in Battisti et al. (2020), Anschütz et al. (2023), or Toborek et al. (2023). [11]

Depending on the aimed text unit, if interested in sentence simplification but only parallel (or comparable) documents are available, manually or automatically identifying parallel pairs on the sentence level could be another option (see Section 3.6 and Section 3.7). Common English corpora following this strategy are, e.g., WikiLarge (Zhang and Lapata, 2017) or WikiAuto (Jiang et al., 2020).

9 https://newsela.com/data/ [last access: July 24, 2024]

10 For more information regarding copyright of data and re-using it, I refer the interested reader to Moorkens and Lewis (2020). It is beyond the scope of this thesis to provide a more detailed analysis of this topic.

11 More details on web scrapers used for building German TS corpora can be found in Section 4.1.

3.5 Manual Simplification (Component C)

If no parallel documents or paragraphs are available or are not specific enough for the simplification purpose, professional translators could be asked to simplify complex texts wrt. a specific target group or simplification guideline. In SimplePatho (Trienes et al., 2022), for example, German medical students have simplified German medical reports to ensure the correctness of the technical language.

If simplifying general texts for a more general audience, also non-professional translators, such as crowd-workers, could be asked to simplify original texts based on given simplification guidelines. This strategy has been applied, for example, to the English TS corpora, for example, TurkCorpus (Xu et al., 2016) or ASSET (Alva-Manchego et al., 2020a).

3.6 Alignment (Component C)

3.6.1 Alignment

If parallel texts are available, the next step is to find the corresponding simple text for a complex text if not yet grouped or linked. The process of identifying the corresponding parallel or comparable parts in a bunch of texts is called *alignment* (McEnery and Xiao, 2007). Following McEnery and Xiao (2007) or Paetzold et al. (2017), aligning can take place at various levels, e.g.

- Document level: Assignment of documents with similar content,
- Paragraph level: Assignment of paragraphs with similar content within similar documents,
- Sentence level: Assignment of sentences with similar content within similar documents, or
- Word level: Assignment of words with similar content within similar sentences.

In text simplification, simplification pairs are most often aligned on the sentence level, but in recent years also some approaches regarding document and paragraph simplification with aligned document pairs and aligned paragraph pairs have arisen (e.g., Sun et al. 2021 or Cripwell et al. 2023a).

Looking more closely at sentence simplification, English TS corpora mostly contain 1:1, n:1 and 1:m sentence pairs, but for German n:m and 0:1 pairs are also frequent. In addition to rephrasing (mainly in 1:1 pairs), splitting (1:m pairs) and merging (n:1 pairs), in German simplification, new sentences are often added to explain a term (instead of substitution; 0:1). Furthermore, there exists an option in which a few sentences are combined, rephrased, and reordered at the same time, resulting in new sentences (n:m pairs).

In the following, I use *complex sentence* and *complex text* interchangeably. With both terms, I refer to the complex side of a complex-simple pair which can either contain one or more sentences. The same holds for *simple sentence*, *simple text*, and *simple side*.

3.6.2 Alignment Types

To the best of my knowledge, no guidelines and clear definitions of alignment types can be found in the literature. Therefore, I am introducing my own definition per alignment type in the following.

1:1 alignments Probably the most common alignment option in text simplification is 1:1 alignment. 1:1 alignments (e.g., see Petersen and Ostendorf 2007; Zhang and Lapata 2017; Alva-Manchego et al. 2020b) are annotated when the meaning of a source sentence is transferred into exactly one simplified sentence. Examples include reordering the sentence, omitting words, adding words, or substituting words in the simplification (e.g., replacing a long or infrequent word with a short or common word).

1:*m* alignment 1:*m* alignments are considered as pairs in which one source sentence has been split into several sentences (see Petersen and Ostendorf 2007; Zhang and Lapata 2017; Alva-Manchego et al. 2020b). This is the case, for example, where subordinate clauses are rewritten as independent sentences. If a fraction of the content of the original sentence also appears in other simplified sentences, all of these simplified sentences should be annotated in the pair. Sentences on one side of the pair can (but do not have to) be adjacent to each other. The order and distance between the simplified sentences do not matter; hence, simplifications corresponding to one complex sentence can be scattered throughout the simplified text. Spring et al. (2023) call this phenomenon *crossing alignment* as the inlying sentences can also be aligned in another alignment pair. Hence, the alignments of both pairs would cross each other. An example of a crossing alignment is provided in Figure 3.2 between the complex sentences 13 to 15 and the complex sentences 19 and 20.

However, if another simplified sentence just refers to one word of the previous simplified sentence, e.g., as a term explanation or exemplification, and this explanation is not part of the complex document, I do not consider this sentence as part of the alignment pair. Instead, this simplified sentence should be aligned as 0:1 and not assigned to any source sentence.

***n*:1 alignment** If fractions of content from more than one source sentence have been conveyed into one simplified sentence, it is an *n*:1 alignment (e.g., see Petersen and Ostendorf 2007; Zhang and Lapata 2017; Alva-Manchego et al. 2020b). In a common *n*:1 alignment, a few pieces of information from the source sentences are omitted and no longer included in the simplified sentence to make the sentence shorter and more concise. However, if the deletion does not change the original meaning, a merge of sentences can still be a valid simplification.

***n*:*m* alignment** If there is a mixture of a summary and information deletion of multiple clauses into multiple sentences, the sentences are annotated with an *n*:*m* alignment. Thus, several cross-sentence simplification transformations occur simultaneously. In an *n*:*m* alignment, many different simplification operations are applied to several source sentences at the same time. For example, a subordinate clause from a first complex sentence might be split into a new simplified sentence, while at the same time the main information of a second complex sentence is

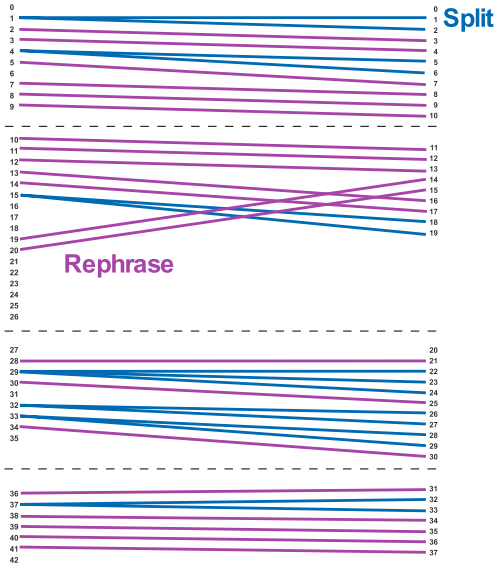


Figure 3.2: Crossing alignment in a news document pair of the Austrian Press Agency (Title: "News from June 21, 20191"; ID = 403).

integrated into the main clause of the first complex sentence. This kind of alignment can be annotated whenever no clear 1:1, 1:*m* or *n*:1 sentence assignment can be made, but the meaning of a complex passage is still preserved in the simplified document.

1:0 or 0:1 alignments As content deletion and word explanations are also relevant simplification operations, 1:0 (deletions; see (Petersen and Ostendorf, 2007)) and 0:1 alignments (additions) can also be considered as alignment types. If, after the annotation of a document pair, some source sentences are not aligned with any simplified sentence, they are considered as deletions and vice versa for additions. These alignment pairs can be helpful for getting insight into the simplification of a document, e.g., how similar the documents are to each other, how much new content has been added, or how much of the original document was removed. Usually, text simplification systems are not trained directly on 1:0 and 0:1 sentence pairs. However, the additions could be used as augmented data to fine-tune a language model on simplified sentences before fine-tuning on the complex-simple pairs (e.g., see Subsection 4.7.3).

3.7 Automatic Alignment (Component C)

Manual sentence-wise alignment of parallel documents is a very time-consuming task as it requires skimming two documents (the original and the simplified document) at the same time and trying to identify sentences with the same or very similar meaning, keeping all possible alignment types in mind. As previously stated, a faster way of alignment – compared to manual alignment following the above definitions – is automatic alignment which could be trained or evaluated on manual alignments. The two most commonly used automatic alignment methods are CATS (Štajner et al., 2018) and MASSalign (Paetzold and Specia, 2016c; Paetzold et al.,

2017), both developed for English (and CATS also for Spanish). However, in recent years, more alignment methods have been developed; for an overview, see Table 3.1. Only some of the methods are capable of aligning *n*:*m* and crossing alignment pairs.

Name	Reference	Crossing Alignments	Nature of Alignment
MASSAlign	Paetzold et al. (2017)	no	*n*:*m*
linear feature-based classifier	Cardon and Grabar (2018)	*n/a*	n:1
CATS	Štajner et al. (2018)	yes	*n*:1
LHA	Nikolov and Hahnloser (2019)	yes	*n*:*m*
VecAlign	Thompson and Koehn (2020)	no	*n*:*m*
SentenceTransformer	Reimers and Gurevych (2020)	no	1:1
Neural CRF	Jiang et al. (2020)	yes	*n*:*m*
BERTAlign	Liu and Zhu (2022)	yes	*n*:*m*

Table 3.1: Automatic alignment methods. Extended Table of Table 1 in Spring et al. (2023).

MASSAlign by Paetzold et al. (2017) is a Python package that includes an easy-to-use alignment method on the paragraph and sentence levels by Paetzold and Specia (2016c). The method uses a vicinity-driven approach with a similarity matrix based on n-grams in a TF-IDF model. It is capable of 1:1, 1:*m*, and *n*:1 alignments.

Linear Feature-based Classifier by Cardon and Grabar (2018) is, as the name suggests, a feature-based linear classifier for sentence-wise alignment. It has been tested for multiple domains as well as multiple languages. Its features include, e.g., stop-word overlap, word overlap, sentence length, and word-based similarity.

CATS by Štajner et al. (2018) is an alignment method that can also align paragraphs and sentences. CATS aligns each original sentence with the closest simple sentence by calculating the similarity of all of them based on n-grams (option: C3G) or word vectors (option: CWASA and WAVG). C3G is a character tri-gram model with similarity metrics, whereas CWASA and WAVG both calculate the cosine similarity between word embeddings of a complex or simplified entity. To make CATS suitable for languages other than English and Spanish, the selection of the word embeddings can be changed to embeddings in the language of interest. Officially the code was published in Java[12], for better integrity with the other alignment methods, but there is also a Python implementation of it[13].

LHA by Nikolov and Hahnloser (2019) is an unsupervised method that finds 1:1 sentence alignments in monolingual parallel corpora where documents do not need to be aligned beforehand. It works with a hierarchical strategy by aligning documents on the first level and then aligning sentences within these documents.

VecAlign by Thompson and Koehn (2020) is a bilingual sentence alignment method that was designed to align sentences in documents of different languages. However, it was also tested in other works on monolingual parallel corpora (e.g., Spring et al. (2022)). It has two main advantages: it can produce *n*:*m* alignments, and it can work with more than 200 languages (as it uses the LASER[14] (Artetxe and Schwenk, 2019) sentence representation model in the background, which is multilingual).

12 https://github.com/neosyon/SimpTextAlign [last update: December 4, 2020; last access: July 24, 2024]
13 https://github.com/kostrzmar/SimpTextAlignPython [last update: March 22, 2021; last access: July 24, 2024]
14 https://github.com/facebookresearch/LASER [last update: May 2, 2024; last access: July 24, 2024]

Sentence Transformer by Reimers and Gurevych (2020) can be used as a straightforward method to find 1:1 sentence alignments by computing cosine similarity between embedding vectors (produced by a sentence transformer model) of sentences on both sides of the parallel corpora, and then picking the most similar pairs and labeling them as *aligned*. This method is totally dependent on the transformer model (e.g., LaBSE Feng et al., 2022 or RoBERTa Conneau et al., 2020), and the similarity threshold to set (e.g., values between 0.7 and 0.9).

Neural CRF by Jiang et al. (2020) is a neural conditional random field (CRF) classifier which is trained on manual alignment pairs. This alignment method classifies for each sentence of a complex document and each sentence of a simplified document whether these sentences should be (partially) aligned or not. Hence, it learns on a bulk of data (i*j where i = length of the complex document and j equals the length of the simple document) with unbalanced data because the category of non-aligned data is overrepresented. Due to this strategy, the method is capable to align crossing *n*:*m* alignment pairs. However, because *Neural CRF* is a supervised metric, it requires training data in the language and domain of interest in order to be applicable to other languages and domains than the ones it was trained on.

BertAlign by Liu and Zhu (2022) is an attempt to allow sentence transformer-based methods to produce *n*:*m* alignments. It was tested on Chinese-English parallel corpora and showed promising results.

In summary, there are a few methods for automatic sentence-wise alignment of parallel documents, but only a few of them are capable of *n*:*m* alignments and crossing alignments. Furthermore, most of the metrics were originally developed for aligning English document pairs, but most of them (except the neural CRF) can be applied to other languages with small adaptations.

In previous work on building German corpora, it has already been shown that a sentence-wise alignment is not suitable for some corpora (mostly comparable and non-parallel corpora) (Aumiller and Gertz, 2022), or the alignment results in low accuracy even when aligning mild simplifications automatically (see Spring et al. 2023). In Chapter 4, I will present in more detail the results of automatic alignment on different German TS corpora.

3.8 Simplification Plans (Component D)

Based on the alignment pairs per document pair, a simplification plan can be built. In this plan, each complex sentence of a complex document can be labeled with a suitable simplification operation corresponding to the alignment type, e.g., *merging*, *splitting*, or *rephrasing*. This plan can either be helpful to train an automatic sentence-wise alignment method (e.g., Jiang et al. 2020) or to train a sentence simplification system with additional simplification operation instructions (e.g., Cripwell et al. 2023b).

3.9 Annotation of Simplification Operations and Quality Assessment (Component F)

In addition to the alignment of a corpus, the annotation or marking of the sentence pairs with the change(s) made during the simplification and an assessment regarding the quality of the

simplifications can also be part of building a text simplification corpus. In the following, I will first highlight the relevance of the annotation on simplification operations and give an overview on existing typologies for the operations (see Subsection 3.9.1). Then I summarize the relevance and strategies for the assessment of the quality of a TS corpus (see Subsection 3.9.2).

3.9.1 Simplification Operations

During manual simplification of a text, many different operations or transformations can be applied to the complex sentence. Following Cardon and Bibal (2023), sentence level simplification operations can be split into syntactical operations (e.g., reordering sentence elements), lexical operations (e.g., paraphrasing or synonym substitution), or morphological operations (e.g., changing the verb's tense or mood). On the document level, further operations can be applied, for example, coreference resolution, sentence deletion, or reordering of sentence positions (Cardon and Bibal, 2023).

Information on simplification operations can be helpful for analyzing the characteristics of a corpus. For example, the operations provide information on how the sentence pairs can be grouped with respect to lexical or syntactic simplifications or whether they are suitable data for evaluation of syntactical TS models.

For text simplification, for example, the following typologies of simplification operations exist for the following languages (Cardon and Bibal, 2023):

1. Bott and Saggion (2014) for Spanish,
2. Brunato et al. (2015), Brunato et al. (2022) for Italian,
3. Gonzalez-Dios et al. (2018) for Basque,
4. Koptient et al. (2019) for French,
5. Caseli et al. (2009) for Brazilian Portuguese, and
6. Amancio and Specia (2014) for English.

Furthermore, simplification operations can be used as part of a guideline for translators on how to simplify texts. However, in this case, often only one simplification is performed at once. Alva-Manchego et al. (2020a) present the first corpus (so-called ASSET) that specifically contains simplifications via multiple operations. Barancikova and Bojar (2020) also describes a controlled text simplification approach. Following their simplification guidelines, several operations have been proposed for each complex sentence, which have been consecutively applied to a complex sentence.

In addition, an existing corpus can be extended by annotation of simplification operations, for example, to derive hand-crafted rules for a TS system (Koptient et al., 2019), to get more insights into the behaviors of metrics for automatic TS evaluation (e.g., see ASSET$_{ann}$ by Cardon et al. 2022), to evaluate system outputs based on applied operations (e.g., see Yamaguchi et al. 2023), or train evaluation metrics including performed operations (e.g., see Heineman et al. 2023). The list of simplification operations can also be extended with error categories, e.g., hallucination or information deletion, to get more insights and comparisons in what have been correctly and wrongly simplified (e.g., see Yamaguchi et al. 2023 or Heineman et al., 2023).

3.9.2 Simplification Quality Assessment

When building an TS corpus, I follow the proxy that a sentence in the target document is simpler than its corresponding sentence(s) in the source document. To check whether this assumption holds, the simplicity of both documents or sentence pairs should be compared to guarantee a high quality of the corpus. Most often, the simplification quality of new corpora is evaluated with readability metrics (e.g., see Vajjala and Lučić 2018), linguistic criteria such as sentence length or word length (e.g., see Scarton et al. 2018 or Ryan et al. 2023), or manual inspection of random samples (e.g., see Joseph et al. 2023).

As suggested in Alva-Manchego et al. (2021), manually simplified sentences (or manual references) should be annotated with their simplicity level when building a new TS corpus. The simplicity level can help to evaluate the model performance based on the maximum simplicity level they could achieve wrt. the given references. Unfortunately, the simplicity or other criteria (e.g., the grammaticality or the meaning preservation) in the gold simplifications are currently very rarely evaluated during the creation of a new TS corpus.

A few English TS corpora have been extended with annotations regarding simplification quality. Examples for this evaluation strategy are the English TS corpora: QATS (Štajner et al., 2016b), HSplit (Sulem et al., 2018a), PWKP (Sulem et al., 2018b), ASSET (Alva-Manchego et al., 2020a), Simplicity-DA (Alva-Manchego et al., 2021), and Fusion (Schwarzer et al., 2021). Furthermore, Taylor et al. (2022) have evaluated the general acceptability of all their professionally simplified documents in qualitative interviews prior to sentence-wise alignment. In addition, Joseph et al. (2023) verified the validity or meaning preservation of their aligned sentence pairs in a random sample of their multi-lingual medical TS corpus.

In some studies (e.g., see Mallinson et al. 2020 or Ryan et al. 2023), the quality of the target texts is assessed at the same time as when evaluating the automatic simplifications. This kind of annotation is more motivated by getting comparable scores of the annotation of system outputs and references and focuses less on the overall quality of the corpus. This procedure has the advantage that the ground truth ratings are originated from the same group of participants, which also evaluates the system outputs. In manual evaluation, the sentence pairs are most often evaluated wrt. simplicity, grammaticality, and meaning preservation (see Chapter 5).

Previously to our work, no German TS corpus has been annotated regarding simplification quality during the building of the corpus, except TextComplexityDE (Naderi et al., 2019). Naderi et al. asked German language learners to rate the simplicity of 1000 sentences from German Wikipedia; the most complex 250 sentences have then been manually simplified. However, no additional evaluation has been made on the simplified sentences. Further, I am aware of one German TS research paper in which the gold reference has been evaluated wrt. grammaticality, meaning preservation, and simplicity during the evaluation of the output of the TS system, that is, Mallinson et al. (2020).

3.10 Annotation Interfaces

As previously presented, most of the creation of a TS corpus corresponds to manual annotation, which is very costly. Some annotation interfaces have been built to facilitate the creation

and annotation of TS corpora. The interfaces focus either on the alignment process (see Subsection 3.10.1), the annotation of the simplification operations (see Subsection 3.10.2), or the annotation of the simplification quality.

3.10.1 Interfaces for Sentence-wise Alignment

Manual sentence-wise alignment of parallel documents is a very time-consuming task, as it requires skimming two documents (the original and the simplified document) at the same time and trying to identify sentences with the same or very similar meaning, keeping all possible alignment types in mind.

To assist annotators in the alignment task, there are a few alignment interfaces (also called bi-text editors or alignment tools), e.g., ISA by Tiedemann (2006), MASSalign's interface Paetzold et al. (2017), or the interface by Jiang et al. (2020). However, they all have some limitations with respect to text simplification corpora construction: The MASSalign interface focuses on the analysis of an alignment pair after automatic alignment and does not have an option for manual alignment or correction. Further, ISA (Tiedemann, 2006) supports only 1:1 alignment, and hence, many sentence pairs for TS would not be recognized or wrongly aligned. The annotation interface of Jiang et al. (2020) is built directly for the purpose of aligning sentence pairs for TS. Their tool is a simple HTML script plus JavaScript in which an uploaded document pair can be aligned. The tool allows a user to mark unlimited spans of texts in both documents (which are *n*:*m* alignments); however, it does not support crossing *n*:*m* alignments as the span in their annotation is not allowed to be interrupted by a sentence which is not part of the alignment. Unfortunately, there are details missing on how to use it, e.g., it is unclear in which format the documents have to be uploaded in the tool.

3.10.2 Interfaces for Annotation of Simplification Operations

The identification and annotation of rewriting transformations can be described as a sequence labeling challenge. Consequently, widely used sequence labeling tools such as BRAT (Stenetorp et al., 2012) can be used for this purpose. Gonzalez-Dios et al. (2018) modified BRAT to annotate rewriting in the context of text simplification by indicating the affected sequence in the original sentence and assigning the corresponding transformation label. In comparison, Koptient et al. (2019) labeled transformations at the word level in parallel text using a modified version of YAWAT (Germann, 2008). Heineman et al. (2023) built an annotation interface called SALSA which is especially designed for the annotation of simplification operations and the quality of system outputs.[15]

3.10.3 Interfaces for Annotation of Simplification Quality

To the best of my knowledge, previously to our work, no annotation tool has focused on supporting the annotation of simplification quality of manual references. However, interfaces that support the manual assessment of system outputs could also be applied for the evaluation of the

15 This interface has been proposed after the release of our text simplification annotation tool TS-ANNO (Stodden and Kallmeyer, 2022).

gold data. For this, crowd-workers have been asked to rate the simplicity through crowdsourcing platforms such as Amazon Mechanical Turk[16] or Figure Eight[17] (e.g., see Alva-Manchego et al. 2020a) or a small group of annotators or experts have evaluated the system outputs (e.g., see Cripwell et al. 2022).

3.11 Summary & Outlook

In summary, I have presented the typical process of building a text simplification corpus including finding suitable data, aligning the data on the sentence level, annotating the simplification quality and operations, and how to support this process with digital interfaces. For the most part, I have presented strategies referring to English TS research due to the imbalanced research between English and non-English TS.

3.11.1 Challenges & Research Gaps

However, most of the strategies can also be applied to German TS to some extent. For example, the resource-finding strategies or the alignment types are universal across languages. But some of the automatic alignment strategies are language-dependent and need some adaptation to also support German. However, the challenges can also be transferred to German TS, e.g., lack of accessible resources for parallel corpora (further called Building Challenge A), time-consuming manual alignment (further called Building Challenge B), or missing reliability of automatic alignment methods (further called Building Challenge C).

Furthermore, prior to my work, there was no typology of simplification operations for German TS except guidelines for manual simplification into German Plain or Easy Language. A comparison between both resources could benefit the research directions of manual as well as automatic simplification. I have further shown that these simplification typologies and annotations have more use cases than just being instructions for simplifying, e.g., they could be utilized as criteria for quality control of corpora (further called Building Challenge D).

On the one hand, the manual annotation of simplification operations or quality assessment aspects involves a high human effort, but, on the other hand, it can be supported by annotation interfaces. However, existing annotation interfaces have also not been utilized to build German TS corpora (Building Challenge B). In addition, before our work, there has been a lack of an annotation tool that comprises all parts of the process of building a TS corpus. Following this, more alignment tools are required that support manual alignment of all types of alignment, including crossing alignment.

The named challenges are inline with my previously introduced research questions which I will address in the following of my thesis, i.e., how to overcome challenges during corpus creation (see RQ 3-1), determining relevant characteristics of new corpora (see RQ 3-2) and identifying the quality and representativeness of corpora (see RQ 3-3).

16 https://www.mturk.com/ [last access: July 24, 2024]

17 https://www.figure-eight.com/; recently called Appen (https://www.appen.com/) [last access: July 24, 2024].

3.11.2 Outlook

In the next section (see Chapter 4), we will have a closer look at how the corpus building strategies have been applied on resources to build German TS corpora and which challenges are arising in this sub-field.

Chapter 4

German Simplification Corpora

In the era of pretrained and large language models, high quality datasets are still of high relevance. Indeed, there is a current trend to promote (new) datasets for pre-training models, e.g., by curating previous corpora or adding new resources (Li et al., 2024), or building new corpora to critically evaluate large language models (Röttger et al., 2024). In the case of TS as a downstream task of natural language generation, most often already pre-trained models are fine-tuned for automatic simplification. Therefore, smaller but parallel text pairs are required to fine-tune as well as to judge the quality of TS systems on unseen data.

Thus, in this chapter, I am introducing corpora for training and evaluating German text simplification models. Such as for other machine learning tasks, a corpus for text simplification also consists of a training set, development set, and test set. The training set, as the name suggests, is used to train or fine-tune a (pre-trained) model, whereas a development set is used for first evaluations of the model and to tune parameters of the model based on the first results. The test set is then used to finally evaluate the model by verifying the predictions of the model on a second unseen evaluation set to check that the model is overall capable of solving the task and does not only perform well on the development set (Jurafsky and Martin, 2009).

In order to build a TS corpus, the corpus building strategies introduced in the previous sections have already been applied to data in many languages, including many for English, but also some for German, Spanish, or French. Besides the building strategy and language of the corpora, the TS corpora also differ regarding the simplification purpose, target group, and text domain. In German Easy Language, most texts correspond to the domains of news, health, and politics (including public authorities) (Maaß, 2020, p. 176).

In the remaining of this chapter, I am introducing existing German sentence and document simplification corpora and shortly linking them to non-German corpora. I group the TS corpora according to the domain or type of their texts. Nonetheless, all digital data could be made available or accessible via the Web, e.g., shared with a cloud service, sent via email, made available on web pages, made available as a downloadable database. Consequently, to an extent, all of these resources could be named "web data". In order to distinguish between the corpora, I have decided to group them by their type or domain of the texts, even if they are still somehow all related to the Web. Therefore, I think of corpora from the web domain as a collection of text data from the Web originated from different webpages and containing texts in different

domains. Hence, it is a mixed group of several texts having in common that they are available on HTML pages and are too few to be gathered in separate corpora. TS corpora assigned to this group will be introduced in Section 4.1. Referring back to the corpus building process of the previous section, web corpora are the group of corpora which always include component A "web harvesting" (see Figure 3.1).

On the other hand, corpora of other domains are more often derived by data collection besides web crawling, e.g., direct simplification or data preparation and provision by the data providers or curators. Corpora are assigned to a domain other than web, if several texts can be grouped to one particular webpage or one particular text domain. Corresponding corpora per these domains will be introduced in dedicated sections: In more detail, for TS corpora which are originated from Wikipedia see Section 4.2. For TS corpora consisting of only news texts see Section 4.3, for corpora with medical corpora see Section 4.4, for corpora with political texts see Section 4.5, and for corpora with narrative texts see Section 4.6.

Additionally, I will present corpora with only simplified texts (see Section 4.7), and data augmentation strategies on how to extend a training dataset (see Section 4.8).

For an extensive overview of non-German TS corpora, I refer to, e.g., Trienes et al. (2024); Ryan et al. (2023); Martin et al. (2023); Madina et al. (2023); Štajner et al. (2022); Alva-Manchego et al. (2020b); Brunato et al. (2022).

4.1 Corpora with Web Texts

As previously discussed, many data can be retrieved from webpages (even when respecting copyright restrictions). The Web also contains many data written in simplified language which is sometimes also linked to a more complex version. With the help of a web crawler, these data can be harvested and stored in simplification corpora. Referring back to the corpus building process of the previous section, web corpora are the group of corpora which always include component A "web harvesting" (see corpus building workflow in Figure 3.1). However, using text data from webpages raises some special challenges compared to other resources, i.e.,

1. they are not static, if a website is crawled at a different time the content might have been changed or deleted.
2. the texts of the websites have often restricted licenses.
3. the domain of web texts is rather broad, as in the Web any texts of any domain can be published.
4. texts of different target groups are often mixed.
5. simplified texts published on the Web often do not pass quality control.

To overcome the first two challenges, at first the webpages can be archived and addressed using archived links, and second, the data is not made available directly. In order to not hurt copyright restrictions, only a web crawler and the name of the web pages can be made available to facilitate others to download the corpus themselves.

Regarding points 3 and 4, mixing different resources and subdomains often also causes a mix of different target groups of the simplifications. However, a combined corpus or a TS model

trained on this corpus cannot meet the needs of all target groups, as each target group has different needs on how a text should be simplified (Siddharthan, 2014; Gooding, 2022).

Focusing on the last point, in contrast to news texts, texts posted on the Web (and also on Wikipedia-like resources) are often not checked whether they are well-written, simple, or meet Plain or Easy Language guidelines. Taking into account these points, the TS corpora of the web domain should be used with caution.

4.1.1 German Resources

In Germany, many texts on webpages are available in simplified German. A recent study by Asghari et al. (2023) showed that 15.5 % of German websites are written in simplified language. They also show that webpages with scientific or governmental content have the least proportion of simplified language (less than 8 %) while news pages and webpages with the topic games have the highest proportion (roughly 30 %). Texts addressing games might be written in general simpler than texts addressing scientific content, and hence no simplification might be required.

One reason for the high amount of simplified texts might be due to an enforcement to provide simplified content on web pages of (at least) public authority websites. Some recent research states that webpages of public authorities are difficult to understand (e.g., see Asghari et al. 2023; Heuer et al. 2024). To counteract, Germany has passed the barrier-free information technology ordinance ("German Barrierefreie-Informationstechnik-Verordnung" (BITV)) for the webpages of German public authorities to improve accessibility on the Web. Regarding the use of language, they strengthen the requirements in the first version of BITV by providing web content only in "the clearest and simplest language that is appropriate" (BITV, 2011). In the second version of BITV (BITV, 2011), the language use is made more clear: public authorities have to provide the most important part of their information additionally in German Easy Language. In more detail, the ordinance states that they should provide at least on introductory and navigating pages in German Easy Language. Hence, it might be just a small proportion in relation to the total number of webpages on a website.

However, to make use of the simplified web texts, several web crawlers or web harvesters have been built in recent years, which extract the content of these parallel web pages. An overview of existing web crawlers is provided in Table 4.1 (for non-parallel web crawlers, see Table 4.15).[1]

Reference	Corpus Name	URL
Klaper et al. (2013)	Simple German Web '13	upon request
Battisti et al. (2020)	Simple German Web '20	not reproducible
Klepp (2022b) †	-	github.com/krupper/transformer-text-readability-classification
Anschütz et al. (2023) †	-	github.com/brjezierski/scrapers
Toborek et al. (2023)	Simple German Web '23	github.com/buschmo/Simple-German-Corpus
Klöser et al. (2024)	-	github.com/MSLars/German-Text-Simplification
Stodden et al. (2023)	DEplain-web	github.com/rstodden/data_collection_german_simplification

Table 4.1: Web crawler of websites with texts in simplified German. The crawler marked with † only extract simplified texts and no parallel text pairs. Last part shows own contributions. All URLs have lastly been accessed at July 24, 2024.

1 I have excluded the web scarper of Hewett and Stede (2021) and Aumiller and Gertz (2022) because they focus on Wikipedia-based web pages which will be explained in Section 4.2.

An overview of the included subdomains and webpages per corpus or web crawler is provided in Table 4.2. Unfortunately, for some of the corpora, the name of the included webpages is not made public, thus, I could not include them in the overview.

4.1.2 Simple German Web Corpus '13

The first who built a parallel corpus based on parallel German web documents are Klaper et al. (2013). Their corpus is very small with overall 1,888 manually aligned sentence pairs of 256 short parallel texts (see component B to E in the corpus building process in Figure 3.1). The simplifications of their corpus are mostly written in German Easy Language. An overview of the included webpages can be found in Table 4.2. Due to copyright limitations of the web documents, they could not make their corpus publicly available, but it is available upon request.

4.1.3 Simple German Web Corpus '20

A few years later, Battisti et al. (2020) extended the corpus by adding more web pages from public authorities, specialized institutions, and non-profit organizations. In total, the corpus contains 378 parallel documents with 21,072 complex sentences and 17,121 simple sentences that address topics such as politics, health, or culture. Additionally, the web crawler of Battisti et al. (2020) harvests a large amount of web pages with only simplified monolingual data, i.e., 5,461 documents with 172,773 sentences and 1,916,045 tokens. Following Ebling et al. (2022) and Spring et al. (2023) of the same research group, the corpus was manually aligned, resulting in 1,080 sentence pairs of 36 documents (with 1,454 complex sentences and 1,440 simple sentences). These alignments have been used as gold data for the evaluation of automatic alignment methods. They compared the alignment methods named CATS (Štajner et al., 2018) and MASSalign (Paetzold et al., 2017) and proposed to use CATS on their datasets as it performed better. Spring et al. (2023) has used the same data to verify more alignment methods and found that LHA (F1: 0.355) can clearly perform better than CATS (F1: 0.046) and MASSalign on the data (F1: 0.108).

Unfortunately, neither the dataset paper by (Battisti et al., 2020) nor one of the closely related papers of the same research group (e.g., Ebling et al. 2022 or Spring et al. 2023) provide more information regarding sentence or word length of the data. Furthermore, neither the document pairs (see component B in the corpus building process in Figure 3.1) nor the sentence pairs (see component E) are made publicly available (outside their research group) as their used web data is not openly licensed. Upon request, they do share the code for the web crawler to download the data oneself (see component A), but they do not share the required URLs of the websites to be crawled, which makes it impossible to reproduce the data. Hence, it is unclear which webpages are included in the corpora and the web crawler. Due to the missing details, I cannot add an overview table with statistics here. However, based on the description in the paper, I can assume that the corpus contains simplified texts in German Plain and Easy Language.

Subcorpus	Website Simple	Website Complex	Simple	Complex	Domain	Description	SGC '13	SGC '23
Alumniportal	alumniportal-deutschland.org †	alumniportal-deutschland.org	PL	PL	language learner	Texts related to Germany and German traditions written for language learners.		
Apotheken Umschau	apotheken-umschau.de/einfache-sprache/ ‡	apotheken-umschau.de	PL	SG	health	Health magazine in which diseases are explained in PL		x
BZFE	bzfe.de/einfache-sprache/ †	bzfe.de	PL	SG	health	Information of the German Federal Agency for Food on good nutrition		
Passanten Verlag	passanten-verlag.de/	projekt-gutenberg.org/	PL	SG/OG	fiction	Books in PL		
Spaß Am Lesen Verlag	einfachebuecher.de/	projekt-gutenberg.org/	PL	SG/OG	fiction	Books in PL		
Behinderten-beauftragter	behindertenbeauftragter.de/DE/LS	behindertenbeauftragter.de	EL	SG	accessibility	Official office for disabled people		x
Bibel	offene-bibel.de/ †‡	offene-bibel.de/	EL	SG	bible	Bible texts in EL		
brandeins	brandeins.de/themen/rubriken/leichte-sprache	–	EL	SG		Translating excerpts from various topics		x
Einfach Teilhaben	einfach-teilhaben.de/DE/LS/	einfach-teilhaben.de	EL	SG	accessibility		x	
Gemeinnützige Werkstätten und Wohnstätten Sindelfingen	gww-netz.de/de-LS/	gww-netz.de	EL		accessibility	Non-profit association in social sector	x	
Heilpädagogische Hilfe Osnabrück	*n/a*	os-hho.de	EL		accessibility	orthopaedagogical support	x	
Lebenshilfe	lebenshilfe-main-taunus.de	lebenshilfe-main-taunus.de	EL	SG	accessibility	Non-profit association for disabled people	x	x
MDR Nachrichten	mdr.de/nachrichten-leicht/	-	EL	SG	news	State-funded public broadcasting service		x
Oberschwäbische Werkstätten	*n/a*	owb.de	EL		accessibility	Non-profit association in social sector	x	
Sozialpolitik	sozialpolitik.com/es	sozialpolitik.com/	EL	SG		Explains social policy in Germany		x
Stadt Hamburg	hamburg.de/barrierefrei/leichte-sprache	hamburg.de	EL	SG	public authority	Information of and regarding the German city Hamburg		
Stadt Köln	stadt-koeln.de/leben-in-koeln/soziales/informationen-leichter-sprache	stadt-koeln.de	EL	SG	public authority	Information of and regarding the German city Cologne		x
TAZ	taz.de/Politik/Deutschland/Leichte-Sprache/!p5097/	taz.de/	EL	SG	news	German Newspaper (discontinued)		x

Table 4.2: Resources for German web TS corpora without own contributions. The line separates German Plain (PL) and Easy Language (EL). OG = Old German, SG = Standard German. All URLs have lastly been accessed at July 24, 2024.

4.1.4 Simple German Web Corpus '23

Recently, Toborek et al. (2023) have published a larger web corpus with accessible archived URLs. Hence, the parallel documents are retrievable and available. Their corpus is a collection of texts in various genres, e.g., public authorities, news, politics, accessibility, and health data. In addition, their corpus contains simplifications in different target levels, i.e., mostly in German Easy Language and some in German Plain Language (see Table 4.2). Due to copyright issues, the data can only be downloaded with the help of the provided web crawler and the archived web texts on web-archive and cannot be made available in a pre-processed version (see component A in corpus building process in Figure 3.1). This corpus tackles the problem of non-static web content by providing links to archived web texts on web-archive. Hence, on any date the same data can be retrieved. On the one hand, even if this is an inconvenient way of accessing the data, it solves the problem of non-static data and incomparability of the data (if it would be retrieved on different dates). On the other hand, the dynamic addition of new documents, one main benefit of web corpora, is prevented by this approach.

In total, the corpus comprises ~700 parallel documents from eight webpages (see component B). The SGW corpus '23 overlaps in only one webpage with the SGW corpus '13 (see Table 4.2), i.e., Lebenshilfe Main Taunus[2]. They aligned 39 of their documents manually, in more detail webpages of MDR (16), Stadt-Köln (5), Apotheken Umschau (5), brandeins (4), lebenshilfe (3), TAZ (2), sozialpolitik (2), behindertenbeauftragter (2). Their manual sentence-wise alignment (see component E) results in 391 sentence pairs with 1:m alignment (152 1:1 alignment, 109 1:2 alignment, and 130 1:m (where $m > 2$). Their alignment is limited to a n:1 relation, hence, only rephrasing and merging, but no splits are recognized. The resulting aligned sentence pairs have been used as gold data for experiments with variants of the CATS alignment algorithm (Štajner et al., 2018) (see component C).

They have combined CATS with different similarity measurements, e.g., cosine similarity or SBERT, and have found that SBERT and "maximum" similarity performed best on their test set (F1: 0.32), but they recommend using maximum similarity as it has higher precision. If comparing two similar strategies (CATS with CWASA) on similar data (web data) in different alignment studies, Toborek et al. (2023) report a 10-times higher F1-Score on their web dataset (0.21) than Spring et al. (2023) on their web dataset (0.024, see Subsection 4.1.3). The reasons for these significant different results remain an open question. It might be due to idiosyncrasies of the test set, the pre-processing, the evaluation method, or other factors. More analysis is required to find the reasons by evaluating on the same test set, with exactly the same algorithm settings, the same n:m relation, and the same evaluation procedure.

Furthermore, Toborek et al. (2023) report that the quality of the alignment method is varying with respect to the source documents: they achieve very high scores (F1 of 0.66) for texts on the websites of the commissioner for the disabled[3] but very low for the German newspaper TAZ[4] (F1 0.06) using the same setting. Nonetheless, they used their best-performing alignment algorithm to automatically align the remaining parallel web documents, resulting in 5,942 sen-

2 https://www.lebenshilfe-main-taunus.de [last access: July 24, 2024]

3 behindertenbeauftragter.de [last access: July 24, 2024]

4 taz.de [last access: July 24, 2024]

tence pairs where ~2,300 are derived from Apotheken Umschau (in German Plain Language), ~1,500 from MDR (in German Easy Language), and ~1,100 from the City of Cologne (in German Easy Language), and the remaining pairs from smaller websites.

4.1.5 BiSECT

BiSECT by Kim et al. (2021) is a multi-lingual, synthetic simplification dataset focusing on syntactical simplification, i.e., splitting and rephrasing one complex sentence into two simpler sentences (1:2 alignment) or merging two complex sentences into one simple sentence (2:1 alignment). Its basis is a bilingual corpus (i.e., OPUS Tiedemann and Nygaard, 2004) from which they extracted German-English pairs with 1:2 or 2:1 sentence alignments. The English parts were then automatically translated into German and filtered regarding lexical and semantic overlap. After removing sentence pairs with too little semantic overlap, the assumption is that the translated sentences are good simplifications of the original English sentences as typical simplification strategies were applied, i.e., splitting or merging. Kim et al. (2021) apply the same approach also to German, French, and Spanish. Additionally, an original, non-translated version of BiSECT is also available in English.

The German version of BiSECT consists mainly of texts in the web domain (89 %) and a few in the political domain. After cleaning, the corpus consists of 186,237 sentence pairs. Unfortunately, the test and development sets have a problem with encoding: umlauts are misspelled, i.e., diacritical markers are missing ("ü" is encoded as "u" and so on), and 'ß'-letters are completely omitted. This evokes problems in training and evaluation of a simplification model: During training, a simplification tool would be trained on grammatically wrong texts. During evaluation, a correctly simplified word with umlauts would rather be considered as wrong as it is neither included in the gold reference nor in the original text. Automatic grammar error correction approaches could help to repair the texts. I experimented with an out-of-the-box German grammar error correction system from Huggingface called "mbart-german-grammar-corrector"[5] to repair the data automatically, but a few examples have shown that the model could not resolve all errors (see Example 1 to 4). More research is needed on how to repair the data and make it usable for German TS.

(1) Simple Sentence of BiSECT: "Im Herzen von Warschau , 15 Minuten zu Fu vom Hauptbahnhof entfernt , begrut Sie Nathan's Villa . Die erschwinglichen Unterkunfte bieten kostenfreien Internetzugang und einen kostenlosen Snack am Morgen ."

(2) Output using mBART-German-Grammar-Corrector: "Im Herzen von Warschau , 15 Minuten zu Fus vom Hauptbahnhof entfernt , begrüssen Sie Nathan's Villa . Die erschwinglichen Unterkunften bieten kostenfreien Internetzugang und einen kostenlosen Snack am Morgen ."

(3) Manual Correction: "Im Herzen von Warschau , 15 Minuten zu Fuß vom Hauptbahnhof entfernt , begrüßt Sie Nathan's Villa . Die erschwinglichen Unterkünfte bieten kostenfreien Internetzugang und einen kostenlosen Snack am Morgen ."

5 https://huggingface.co/MRNH/mbart-german-grammar-corrector [last update: August 19, 2023; last access: July 24, 2024]

(4) Correct Translation: "In the heart of Warsaw , a 15-minute walk from the main train station , Nathan's Villa welcomes you . The affordable accommodations offer free internet access and a free snack in the morning ."

4.1.6 Capito Corpus

Spring et al. (2021)[6] propose a corpus for German TS based on professionally translated texts by the capito translation agency[7]. This corpus is based on texts in standard German of different genres, e.g. brochures, information texts, websites, and legal texts, which have been professionally simplified into three simplification levels, i.e., OR-B1, OR-A2, and OR-A1. For the simplification, simplification guidelines have been used, which are unfortunately not publicly available. A small part of the corpus was manually aligned on the sentence level, which serves as evaluation data for automatic alignment algorithms.

Following Spring et al. (2023), the gold data consists of in total 42 documents (OR-A1: 22, OR-A2: 8, and OR-B1: 12), and 1,254 aligned sentence pairs (OR-A1: 416, OR-A2: 412, and OR-B1: 426). The authors have again experimented with different alignment methods, i.e., CATS (Štajner et al., 2018), LHA (Nikolov and Hahnloser, 2019), MASSalign (Paetzold et al., 2017), SentenceBERT (Reimers and Gurevych, 2020) and Vecalign (Thompson and Koehn, 2020). They found Vecalign to perform best on the subcorpora capito OR-A2 (F1-Score: 0.392) and OR-A1 (F1: 0.215), but LHA is best in OR-B1 (F1 of 0.513) (Spring et al., 2023). Spring et al. (2021) found out that ensembling three alignment methods, i.e., LHA, SentenceBERT and Vecalign, performs better on the subcorpora capito OR-A1 (F1: 0.379) and capito OR-A2 (F1: 0.621) but worse on capito OR-B1 (F1: 0.359) compared to one alignment method only. It seems that the more mild the simplifications, the better the performance of the alignment methods.

Based on the results of the evaluation of the alignment methods, they automatically aligned the remaining 3,440 documents (OR-B1: 1,055, OR-A2: 1,546, OR-A1: 839) of the capito corpus with LHA. It results in 54,224 sentence pairs for capito OR-B1, 136,582 sentence pairs for capito OR-A2, and 10,952 sentence pairs for capito OR-A1.

Spring et al. (2023) also claim that the simplified texts are written in a different order than the original texts as the sentences are not simplified one by one but moved within the new document. To be more clear, in each subcorpus more than 87% of the sentence pairs contain crossing sentence pairs. However, they do not find an effect that corpora with more crossing sentence pairs are more difficult to automatically align than corpora with fewer crossing sentence pairs. Unfortunately, due to copyright issues, the data are not available and, hence, no further details can be reported, nor can the data be used to train TS models.

4.1.7 HDA-Leichte-Sprache-Corpus & GEASY & DE-Lite

The HDA-Leichte-Sprache-Corpus by Siegel et al. (2019) and the GEASY corpus by Hansen-Schirra et al. (2021) are both relatively small corpora with German Easy Language as the target language. In comparison, the GEASY corpus is smaller (93 document pairs) and is focused on texts in German Easy Language, whereas the leichte-sprache-corpus contains also German Plain

6 More versions of the corpus are also introduced in Ebling et al. (2022) and Spring et al. (2023).
7 https://www.capito.eu/ [last access: July 24, 2024]

Language and therefore more parallel documents (351 parallel documents). For both corpora, it is not clear which websites are included in the corpus. However, for the leichte-sprache-corpus, the names of the subcorpora give more information on the genres of the texts, e.g., bible, news, fairy tales, election manifesto, and literature. A long list of websites with comments on their availability has been collected for the GEASY corpus[8], even so it is not clear which documents are finally part of the corpus.

However, in contrast to the leichte-sprache-corpus, the GEASY corpus has been (manually) aligned on the sentence level using memsource[9], resulting in 1,816 alignment pairs. In addition, the authors of GEASY give more insight into the alignment types (see Table 4.3). Their manually aligned corpus contains more *n*:*m* pairs (54.46%, mostly 1:*m* pairs (overall 47.8%)) than 1:1 pairs (36.56%) and also a small proportion of additions (0:1, 8.04%) and only a surprisingly low portion of deletions (1:0, 1%). Following this, GEASY (and more generally German Easy Language) seems to contain many cross-sentence simplification operations (e.g., as sentence splitting or term explanations) than German Plain Language (compared to APA-RST in Table 4.10).

Level	1:1 (total)	*n*:1	1:*m*	*n*:*m*	*n*:*m* (all)	1:0	0:1
all	664	21	868	100	989	17	146

Table 4.3: Characteristics of the document simplification corpus GEASY. The Table is based on Table 3 in Hansen-Schirra et al. (2020b).

Unfortunately, the GEASY corpus is not publicly available. But the leichte-sprache-corpus is available online[10] and can be used as a test set for document simplification.

Jablotschkin et al. (2024) recently proposed another web-based corpus, named DE-Lite. This corpus contains parallel complex-simple pairs and additional monolingual data of the LeiKo corpus (Jablotschkin and Zinsmeister, 2020) (see Subsection 4.7.3). The parallel subcorpus partially overlaps with previously named corpora, i.e., Simple German Web Corpus '20 (see Subsection 4.1.3), Simple German Web Corpus '23 (see Subsection 4.1.4), GEASY, and our proposed corpus DEplain (see Part II Subsection 7.4.2). More details regarding the overlap are not available, e.g., which webpages are overlapping or to what extent. In its current version, the corpus contains roughly 8,000 parallel documents and is planned to be extended soon. At the moment, the data is not available yet, but will hopefully be available online soon[11], which would facilitate experiments with document simplification on German Easy Language.

4.1.8 Semi-synthetic Simple German Web Corpus

Recently, Klöser et al. (2024) also proposed a parallel German corpus for document simplification based on simplified documents crawled from the Web. In order to gather simplified documents, they extended the web crawler proposed by Anschütz et al. (2023). However, it remains open whether they dropped webpages, added new webpages, or both.

8 https://traco.uni-mainz.de/geasy-korpus/ [last update: September 28, 2024; last access: July 24, 2024]
9 https://www.memsource.com/ now called phrase https://phrase.com/ [last update: December 4, 2020; last access: July 24, 2024]
10 https://github.com/hdaSprachtechnologie/easy-to-understand_language [last update: March 16, 2022; last access: July 24, 2024]
11 https://github.com/HeikeZinsmeister/DE-Lite [last access: July 24, 2024]

In total, their corpus contains 8,130 parallel, but semi-synthetic documents. In more detail, they align the manually simplified document with complex texts that are automatically generated by GPT-4 (OpenAI, 2024). For their complex texts, they prompted GPT-4 with 15 different instructions on how to rephrase the simplified texts. Unfortunately, the formulation of the prompts is not available.

Furthermore, their corpus relies on the assumption that the language model is trained on difficult-to-read texts and, hence, that it produces more complex versions of the simplified texts. More analysis is required to justify this proxy. Unfortunately, the data is not yet available. However, it is anticipated that it will be accessible in the near future[12]. Due to the missing data and details regarding the web crawler, I cannot add more statistics regarding the readability or simplification extent of the corpus nor compare the corpus to other corpora in the same domain.

4.1.9 Miscellaneous

Besides the corpora previously named in this chapter, there are some studies which analyze parallel data in standard German and simplified German, e.g., Jekat et al. (2017), or Lange (2018). These datasets are not publicly available, and many details are missing regarding the corpus description, e.g., it is not clear whether they are sentence-wise aligned. Hence, they are not further described in this work because too much information is missing.

4.2 Corpora with Wikipedia Texts & Knowledge Acquisition Texts

Following Ferschke et al. (2013), texts from "Wikipedia – The Free Encyclopedia" are very often used to build NLP corpora because it contains huge data (e.g., more than 6.8 million articles in English Wikipedia contributors, 2024b), is available in more than 300 languages (Wikipedia contributors, 2024b), and most of the content (except images) is openly licensed (Wikipedia contributors, 2024a). Following this, Wikipedia is also a prominent resource for building text simplification corpora.

4.2.1 Non-German Corpora

For English text simplification, Wikipedia-based corpora are a predominant resource because there are two English versions, i.e., the standard English Wikipedia[13] and the Simple English Wikipedia[14]. The Simple English Wikipedia is addressed to non-native English speakers, but can also be helpful for people with learning difficulties (Alva-Manchego et al., 2020b). It seems probable that the majority of the articles in both Wikipedia versions were written independently, a phenomenon also observable in different language versions of Wikipedia articles. Consequently, this resource merely comprises comparable, rather than parallel, articles.

Nevertheless, researchers have utilized this openly licensed resource by linking the documents of both versions and automatically aligning the content sentence-wise. The resulting corpora are, for example, PWKP (108,000 pairs) (Zhu et al., 2010), EW-SEW (392,000

12 https://github.com/MSLars/German-Text-Simplification [last access: July 24, 2024]
13 https://www.wikipedia.org/ [last access: July 24, 2024]
14 https://simple.wikipedia.org/wiki/Main_Page [last access: July 24, 2024]

pairs) (Hwang et al., 2015), or WikiLarge (286,000 pairs) (Zhang and Lapata, 2017). Although these corpora are suitable to train TS models due to their large size, these corpora have been also criticized wrt. i) low quality due to many misalignments, and ii) monotonous simplification operations (Alva-Manchego et al., 2020b).

To address the problem of misalignments (see component C in corpus building process in Figure 3.1), Jiang et al. (2020) proposed a new alignment method (i.e., neural CRF alignment), a new manual aligned (i.e., Wiki-manual), and an automatically aligned corpus (i.e., Wiki-auto). To overcome the challenge of monotonous simplification operations, three Wikipedia-based corpora with higher quality have been built, i.e., TurkCorpus (Xu et al., 2016), HSplit (Sulem et al., 2018a), and ASSET (Alva-Manchego et al., 2020a). All three corpora contain several simplification options for the same original sentences, but focusing on different simplification types, i.e., TurkCorpus on lexical simplification, HSplit on structural simplification, and ASSET on multiple rewriting strategies. Following this, nowadays ASSET is the most common evaluation data set for evaluation of English sentence simplification models.

Also in other languages than English, Wikipedia has been used as a resource to build text simplification corpora, even if no simple version of the original Wikipedia exists in these languages. For example, for Italian, Tonelli et al. (2016) proposed SIMPITIKI, a simplification corpus based on Wikipedia's revision history. In French, Tack et al. (2016), Grabar and Cardon (2018), and Ormaechea and Tsourakis (2023) build a comparable simplification corpus based on the French Wikipedia (written for French speaking people) and a Wikipedia version written for French speaking children called Vikidia[15]. Although Vikidia also exists in other languages (e.g., Arabic, Basque, Greek, German, or Spanish) (Vikidia contributors, 2024), to the best of my knowledge, this resource is currently only used to build French simplification corpora.

Another strategy to use Wikipedia for building simplification corpora is machine translation, e.g., WikiLarge has been automatically translated to Russian (resulting in RuWikiLarge) (Sakhovskiy et al., 2021) and French (resulting in WikiLarge FR) (Cardon and Grabar, 2020).

4.2.2 German Resources

There is no parallel or comparable simple German version of the documents on the original Wikipedia website. But there are a few similar resources, for example, three webpages which include simplified texts in Wikipedia style written for children, i.e., Vikidia, Klexikon[16] and Miniklexikon[17]. Klexikon addresses children between 8 and 13 (Klexikon contributors, 2023), whereas Miniklexikon addresses younger children and also people with low German reading skills (MiniKlexikon contributors, 2024). Klexikon and Miniklexikon are at least comparable documents, as for each document of one source a document exists also in the other source (MiniKlexikon contributors, 2024). Besides Wikipedia-like platforms for children, there is also a platform with Wikipedia-like texts written in German Easy Language called Hurraki[18].

15 https://fr.vikidia.org/wiki/Vikidia:Accueil [last access: July 24, 2024]

16 https://klexikon.zum.de/ [last access: July 24, 2024]

17 https://miniklexikon.zum.de/ [last access: July 24, 2024]

18 https://hurraki.de/ [last access: July 24, 2024]

Another German resource for simple knowledge acquisition texts (except Miniklexikon and Klexikon) is GEOlino[19]. GEOlino is a German science magazine for children between the ages of 8 and 14 years (GEO.de, 2024).

To build German TS corpora upon Wikipedia or similar texts, a few strategies have been applied to use Wikipedia resources to build German TS corpora, i.e.,

1. translation of English Simple Wikipedia pages (see Subsection 4.2.3),
2. translation of English Wikipedia-based TS corpora (see Subsection 4.2.4), and
3. manual simplification of Wikipedia pages (see Subsection 4.2.6), and
4. simplifying simple German texts written in Wikipedia style (see Subsection 4.2.5 and Subsection 4.2.7).

However, in contrast to the English Wikipedia test sets, all German sets (except translated ASSET) contain only one reference, which corresponds to only one gold simplification.

4.2.3 Translated Wikipedia Corpus

A German TS corpus based on standard English and Simple English Wikipedia has been proposed by Ebling et al. (2022): it is a synthetic corpus called Wikipedia-Corpus. In this corpus, Wikipedia articles in standard German were automatically aligned to their English and Simple English counterparts based on inter-language linking of Wikipedia itself. Subsequently, the Simple English Wikipedia articles were automatically translated into German using the DeepL translator[20] and then aligned to the original German Wikipedia articles.

Overall, the size of the corpus is comparably huge, i.e., ~106,000 document pairs with ~7 million tokens on the source side and ~1.1 million tokens on the target side. Hence, the corpus would be large enough for training document simplification systems.

In order to be able to use this corpus also for sentence simplification, the corpus has been sentence-wise aligned. Therefore, approximately 200 of the document pairs have been manually aligned at the sentence level (resulting in 1,382 sentence pairs) to enable the evaluation of automatic alignment methods. Spring et al. (2023) have experienced with many different alignment methods for automatic alignment (see component C in corpus building process in Figure 3.1). In more detail, they used the methods called MASSalign Paetzold et al. 2017, CATS Štajner et al. 2018, and LHA Nikolov and Hahnloser 2019. Based on their report, it seems that the methods do not perform well on this task, which might be due to many misalignments (lower precision than recall). Overall, their best method, i.e., large-scale hierarchical alignment (LHA) (Nikolov and Hahnloser, 2019) achieves an F1-score (F1: 0.170) on the manually aligned sentence pairs. Furthermore, their experiments are limited to their chosen alignment methods as they are not able to identify $n : m$ sentence pairs.

Unfortunately, neither the document level corpus, the sentence level corpus, nor the code to rebuild the corpus is available. Hence, I cannot report on more insight into the corpus. I can only argue that due to its size, the corpus might be very suitable for training TS models, but the alignments and simplification quality should be checked before using it.

19 https://www.geo.de/geolino [last access: July 24, 2024]
20 https://www.deepl.com/translator [last access: July 24, 2024]

It is another open question how this corpus relates to similar English TS corpora such as WikiLarge or PWKP. The articles in Simple English Wikipedia are written in English Plain Language; however, it is unclear whether the articles are of the same complexity after the translation. If the complexity has remained the same during translation, the texts would address language learners on CEFR level A2 or B1.

4.2.4 Translated ASSET

Schlippe and Eichinger (2023) experiment on German TS by training and evaluating on synthetic data also generated with machine translation, in more detail with Google Translate[21]. Their corpus contains automatic translations of 1,000 sentences from the English simplification corpus named TurkCorpus (Xu et al., 2016) and additional 500 sentence pairs of ASSET (Alva-Manchego et al., 2020a). All pairs have been translated into 40 languages including German. Unfortunately, neither their training data nor their test data is publicly available. Therefore, I cannot provide any further insights into this corpus. But similarly to the translated Wikipedia Corpus, I would recommend checking whether the simplifications and the simplification variety of ASSET and TurkCorpus have been lost or kept during the translation.

4.2.5 Lexica Corpus and Klexikon

As previously mentioned, some Wikipedia articles have been simplified for children of different ages, e.g., Klexikon for children between 8 and 13 (Klexikon contributors, 2023) and Miniklexikon for even younger children (MiniKlexikon contributors, 2024). In the following corpora, these resources have been utilized: i.e., a simplification and summarization corpus called Klexikon (Aumiller and Gertz, 2022), a document simplification corpus Lexica-corpus-klexikon (Hewett and Stede, 2021) and another document simplification corpus Lexica-corpus-miniklexikon (Hewett and Stede, 2021).

In all three corpora, the document alignments are based on title matching of the articles. For the Klexikon corpus, the authors automatically aligned the documents of Klexikon and Wikipedia based on their titles, but considered disambiguation of the titles, e.g., different articles for homonyms without disambiguation in the title, e.g., "Bank" vs. "Bank (credit institution)". The title matching process for the lexica corpora is not described in more detail. Overall, Klexikon results in total in 3,000 comparable documents, whereas Lexica contains 300 comparable documents for each subcorpus (see Table 4.4 and Table 4.5). As some articles were too long or were not available on both webpages (Hewett and Stede, 2021), the number of comparable documents in the Lexica corpus is less than in the Klexikon corpus. However, the Lexica corpus is dynamic as with the provided code the corpus can grow over time. An extended version of March 2022 already includes 1000 documents per subcorpus (Hewett and Stede, 2022).

Furthermore, the Lexica corpus is also smaller than the Klexikon corpus in terms of length per document, more precisely, the number of sentences per document. The main reason for that is that in the Lexica corpus only the first original paragraph is aligned with the full simplified texts, whereas in Klexikon the full original and simplified articles are aligned with each other.

21 https://translate.google.com/about/ [last access: July 24, 2024]

Therefore, Klexikon is supposed to be a gold dataset for the combined tasks of text simplification and text summarization, while the Lexica corpus aims for only text simplification.

Comparing standard Wikipedia with Klexikon and Miniklexikon (see Table 4.4 and Table 4.5), as expected, standard Wikipedia contains on average the longest sentences (between 18.41 and 22.7 words) whereas sentences in Klexikon are up to 10 words shorter, and Miniklexikon is even 14 words shorter.

Both corpora are openly available due to the open licenses of their resources (i.e., Wikipedia, Klexikon, and Miniklexikon).

It can be assumed that simplification between Wikipedia and Klexikon or Miniklexikon results in strong simplifications, whereas the simplification between Klexikon and Miniklexikon results in a mild simplification as the target groups are closer than the ones of Klexikon and original Wikipedia.

Both corpora also have in common that they are aligned on the document level but not on the sentence level. Hence, the corpus building process for these corpora ends at component B (see Figure 3.1). Aumiller and Gertz (2022) state that automatic sentence alignment does not apply to these resources as the documents are independently written and contain several *n*:*m* alignments (i.e., sentence splits and merges of sentences).

			Complex			Simple		
	# Doc. Pairs	**n:m**	**FRE↓**	**Sent. Len.↑**	**Word Len.↑**	**FRE↑**	**Sent. Len.↓**	**Word Len.↓**
Klexikon	3,000	100%	40.1	22.7	8.7	66.7	13.5	6.9

Table 4.4: Characteristics of the document simplification corpus Klexikon. Scores are based on Table 4 in Aumiller and Gertz (2022). Word length in characters.

			Complex			Simple		
	# Doc. Pairs	**n:m**	**FRE↓**	**Sent. Len.↑**	**Word Len.↑**	**FRE↑**	**Sent. Len.↓**	**Word Len.↓**
lexica-Klexikon	295	100%	-	18.41	-	-	13.29	-
lexica-Miniklexikon	295	100%	-	18.41	-		9.57	-

Table 4.5: Characteristics of the document simplification corpus Lexica-Corpus. Scores are based on Table 1 in Hewett and Stede (2021).

4.2.6 TextComplexityDE

Naderi et al. (2019) (updated version proposed in Mohtaj et al. 2019) have built a sentence simplification corpus based on manual simplifications of 23 Wikipedia articles originally written in standard German. In the first step, they asked crowd workers with a self-indicated CEFR level of A or B to rate 11 aspects in 1,000 sentences of these Wikipedia articles. Their rating aspects include, for example, understandability, complexity, or lexical complexity. In a second step, the authors picked 265 sentences from these 1,000 sentences, which were ranked as complex and not easily understandable. Then they again assigned crowd-workers to simplify these sentences manually.[22] As a result, their corpus, called TextComplexityDE, contains 250 sentence pairs

22 The paper does not give any information whether simplification guidelines were provided to the non-experts to unify and support the simplifications.

manually simplified by non-experts. Due to the manual sentence-wise simplification, they could skip the alignment of the documents and sentences. The corpus is openly licensed with an MIT license and available online[23].

Following the automatic readability assessment using Flesch-Reading-Ease (FRE), the original Wikipedia texts can be interpreted as difficult (FRE: 28.1, see Table 4.6) and the simplification as "on average" (FRE: 51.2) so that the simplifications are neither very strong nor mild. Nevertheless, the corpus seems to include many simplification operations as 83 % of the sentence pairs are *n*:*m* alignments (see Table 4.6).

Despite its small size, the corpus has not only been used as a test set (see e.g., Stodden et al. 2023), but has also been divided into development and test sets (see Mallinson et al. 2020) or training and test sets (see Ryan et al. 2023).

			Complex			Simple		
	# Sent. Pairs	*n*:*m*	**FRE↓**	**Sent. Len.↑**	**Word Len.↑**	**FRE↑**	**Sent. Len.↓**	**Word Len.↓**
TextComplexityDE	250	83 %	28.1	27.75	2.08	51.2	14.17	1.9

Table 4.6: Characteristics of the sentence simplification corpus TextComplexityDE. Own calculation. Sentence length in Tokens. Word length in syllables.

4.2.7 GEOLINO

In comparison to the previous corpora, the sentence simplification corpus proposed by Mallinson et al. (2020) called GEOlino is based on a different resource for knowledge acquisition, i.e., a German science magazine for children called GEOlino[24]. The same source has already been used in the context of readability classification of German text, e.g., in Hancke et al. 2012 or Weiss and Meurers 2018. However, in contrast to the data of Mallinson et al. 2020, their data is not available.

Similarly to the TextComplexityDE corpus, the data of GEOlino has been manually simplified. However, instead of crowd-workers, a trained German linguist simplified 20 GEOlino articles, following a simplification guideline that has been especially written for this purpose (see Appendix A in Mallinson et al. 2020). The original articles are written for children between the age of 8 and 14 (similarly to Klexikon articles), while the simplifications are directed at children between 5 and 7 (similarly to Miniklexikon articles). In total, their corpus consists of 1,198 manually simplified sentence pairs, split into development and test data. Following the target groups of both sources and the recalculated Flesh-Reading Ease scores (FRE) (see Table 4.7), I can expect very mild simplifications in the corpus as both sources ($FRE_{original}$: 61.5 (interpretation: simple), $FRE_{simplified}$: 66.0 (interpretation: simple)) are on average written more simple than texts in standard German ($FRE_{standard}$: 40 to 60). Following the FRE scores, the simplified GEOlino texts seem to be similarly complex as the Klexikon texts (FRE: 66.7) even if they are written for older children. The word length of Klexikon (see Table 4.4) and GEOlino (see Ta-

23 https://github.com/babaknaderi/TextComplexityDE [last update: April 8, 2022; last access: July 24, 2024]

24 https://www.geo.de/geolino [last access: July 24, 2024]

ble 4.7) are not comparable, as the first is measured in characters, whereas the latter is measured in syllables.

Even if the simplification extent is smaller in GEOlino than in TextComplexityDE, the GEOlino corpus seems to be of a higher quality considering the expert simplifications. GEOlino is also four times larger than TextComplexityDE and can therefore unhesitatingly be split into a development and test set. Unfortunately, this corpus does not contain training data as it was not required in the original approach, i.e., zero-shot simplification (see Section 6.5). Hence, the corpus is great for evaluation, but still too small for training language models.

Another limitation of the corpus is that even if the texts were simplified per article, only parallel sentence pairs are openly available (dev: 535 pairs, test: 663 pairs). Furthermore, I could not find any information on whether the sentence pairs are extracted manually or automatically from the parallel documents (see component C in the corpus building process in Figure 3.1).

			Complex			Simple		
	# Sent. Pairs	*n*:*m*	**FRE↓**	**Sent. Len.↑**	**Word Len.↑**	**FRE↑**	**Sent. Len.↓**	**Word Len.↓**
GEOlino (test)	663	40 %	61.5	13.31	1.7	66.0	9.94	1.66

Table 4.7: Characteristics of the sentence simplification corpus GEOlino. Own calculation. Sentence length in tokens. Word length in syllables.

4.3 Corpora with News Texts

In addition to Wikipedia texts, news texts are also a popular resource for building text simplification corpora, because some news providers simplify their articles to make them accessible to a wider audience. Their aim is to enable people to participate in information processes (APA – Austria Presse Agentur, 2020) and to create a good basis for forming their own pattern of opinions (capito, 2020). On the one hand, the vast array of news providers and the multitude of news articles published daily means that there is a mass of parallel documents with varying degrees of complexity. On the other hand, the frequency with which news articles are published presents a great opportunity for TS to support translators in their daily work.

4.3.1 Non-German Corpora

In many languages comparable or parallel news articles have been used to build text simplification corpora: e.g., Arabic (Al-Raisi et al., 2018), Danish – DSIM (Klerke and Søgaard, 2012), English – Newsela-EN (Xu et al., 2015) or OneStopEnglish (Vajjala and Lučić, 2018), Spanish – Newsela-ES (Štajner et al., 2017), Italian – READ-IT (Dell'Orletta et al., 2011), Japanese (Goto et al., 2015; Kodaira et al., 2016; Maruyama and Yamamoto, 2018; Katsuta and Yamamoto, 2018), Brazilian Portuguese (Caseli et al., 2009), Dutch (Bulté et al., 2018), Czech – COSTRA (Barancikova and Bojar, 2020), Swedish – LäsBarT (Heimann Mühlenbock, 2008), Finnish (Dmitrieva and Konovalova, 2023) or Urdu – SimplifyUR (Qasmi et al., 2020). In order to build these corpora, publicly available, professionally simplified news have been utilized (e.g., see Bott and Saggion 2011 or Xu et al. 2015) or the original news have been professionally translated especially for the purpose of building a text simplification corpus (e.g., see Caseli et al., 2009, Klerke

and Søgaard, 2012, or Vajjala and Lučić, 2018). Most of these corpora are created for sentence simplification, and unfortunately, the parallel documents of which the sentence pairs have been extracted are often not additionally provided to also support document simplification.

Furthermore, the news articles of some news providers are simplified, but there is no parallel version in the standard version. An example of a collection of these (not necessarily comparable) news data is provided in the SNIML corpus (Hauser et al., 2022). It is a dynamic, monthly updated corpus with simplified news in six languages (i.e., English, Italian, Belgian French, Swiss German, Finnish, and Swedish). However, since no parallel version in the standard language is available, the resources cannot be used to train a sequence-to-sequence text simplification model.

The most prominent English news text simplification corpus is Newsela (Xu et al., 2016) which contains parallel news texts in standard and simplified English. In addition, it also contains data in standard and simplified Spanish. For both languages, the documents are available in 5 to 6 versions of different complexity levels, where one of them is the original resource. In total, it contains roughly 1,000 document pairs for Spanish and roughly 9,000 document pairs for English. All of the simplified versions were written by professional translators of simple language. Furthermore, Xu et al. (2015) also provides sentence pairs which were automatically aligned for the English version of this corpus. This corpus is not publicly available, but can be requested for academic purposes[25]. Jiang et al. (2020) extended the original Newsela-EN corpus in the following way: 1,932 documents are available at 5 different language levels (0 to 4, where 0 is the original text), each of the levels can be aligned with each other (except self-alignment), resulting in a total of 10 different alignment settings. Following this procedure, their corpus consists of 19,320 document pairs. Furthermore, they manually and automatically aligned these document pairs also sentence-wise, resulting in the corpora Newsela-auto and Newsela-manual (Jiang et al., 2020).

4.3.2 German Resources

In German-speaking countries, several newspaper agencies also offer their news articles in simpler versions. For example, Deutschlandfunk[26], Nord-Deutscher Rundfunk[27], Mitteldeutscher Rundfunk[28], Saarländischer Rundfunk[29], taz Verlag[30], kurier.at[31], or infoeasy[32]. However, they cannot be used to train a text simplification model because a corresponding parallel (or neither comparable) version of the same news report is often not available in standard German. Aligning the simplified news with news in standard German by other news agencies does not

25 https://newsela.com/data/ [last access: July 24, 2024]
26 nachrichtenleicht.de [last access: July 24, 2024]
27 https://www.ndr.de/fernsehen/barrierefreie_angebote/leichte_sprache/index.html [last access: July 24, 2024]
28 https://www.mdr.de/nachrichten-leicht/nachrichten-in-leichter-sprache-114.html [last access: July 24, 2024]
29 https://www.sr.de/sr/home/nachrichten/nachrichten_einfach/index.html [last access: July 24, 2024]
30 https://taz.de/tazleicht/!t5425449/ [last update: August 20, 2021; last access: July 24, 2024]
31 https://kurier.at/einfache-sprache [last access: July 24, 2024]
32 https://infoeasy-news.ch/ [last access: July 24, 2024]

seem useful as it is not guaranteed that they contain similar information, and they would neither build a parallel nor comparable resource. In Section 4.8, I will demonstrate how to utilize these simplified monolingual texts despite their inherent limitations of not being parallel.

Since 2017, the Austrian Press Agency (APA)[33] is publishing one news report (including four to five news items) in two language levels on each working day, i.e., news in standard Austrian German and news to be understandable for foreign language learners with CEFR level B1. Starting in November 2018, they added versions in a third language level, i.e., CEFR level A2. APA is obtaining the two professionally simplified versions from the Capito translation agency[34], which are also the editors of the capito corpus (see Subsection 4.1.6). Hence, all APA corpora described in the following are written with respect to the same simplification guidelines as the capito corpus. Each news report includes the following topics: culture, economics, politics, or sports (capito, 2020; Kneil et al., 2020). The news articles in all three language levels are fully parallel, as they were translated upon the standard version. An example of a news report including four news items which are simplified for people with CEFR level B1 and A2 of German is provided in the Appendix.

This resource can and already has been used to build corpora with up to four different combinations of complex-simple variations, i.e., OR-B1 (see German-News-Corpus in Subsection 4.3.3, APA-LHA in Subsection 4.3.4, and APA-RST in Subsection 4.3.5) OR-A2 (see APA-LHA in Subsection 4.3.4, and APA-RST in Subsection 4.3.5), B1-A2 (see APA-RST in Subsection 4.3.5), or OR-B1-A2 (with two references).

4.3.3 German News Corpus

The first paper in which the APA resource was used for text simplification is Säuberli et al. (2020). They propose a TS corpus named "German News Corpus" based on APA news articles that were published between August 2018 and December 2019. They automatically aligned the original articles and the simplification at level B1 on the sentence level with the alignment method called CATS (Štajner et al., 2018). The alignment results in 3,616 sentence pairs.

This resource is also claimed to be part of the multi-lingual TS benchmark MultiSim (Ryan et al., 2023). However, comparing the corpus size stated in Säuberli et al. (2020), and the size stated in Ryan et al. (2023), the latter is much larger. It remains unclear whether Ryan et al. (2023) has used an extension of the German News Corpus or a different German news corpus such as APA-LHA (Spring et al., 2022) (see Subsection 4.3.4), whose size is more close to the size referred to in Ryan et al. (2023).

4.3.4 APA-LHA

In 2021, Spring et al. extended the German News Corpus wrt. several aspects, i.e., publication period, number of news articles, number of language levels, and as a result a higher number of sentence-wise aligned pairs.

The resulting corpus, called APA-LHA, includes ca. 480 news articles (the number is estimated based on the number of documents) that were published between August 2018 and April

33 https://science.apa.at/nachrichten-leicht-verstandlich/ [last access: July 24, 2024]
34 https://www.capito.eu/ [last access: July 24, 2024]

2021. As during this time, news articles in two language levels, i.e., A2 and B1, were made available, they propose two parallel subcorpora: OR-B1 and OR-A2. In addition, the authors describe that their corpus consists of about 2,300 pairs of documents per language level. As the number of documents is higher than the number of possible daily articles (equal to working days in the time duration, which is 144 weeks x 5 working days = 720 working days), I assume that they count their number of documents based on the number of news items (720 news articles x 4 news items = 2,880 documents).

As described in Ebling et al. (2022) and Spring et al. (2023), they manually aligned 134 news items (67 for each level) at the sentence level, resulting in 518 sentence pairs for OR-B1 and 504 sentence pairs for OR-A2. The manually aligned sentence pairs are utilized to evaluate automatic alignment methods in this dataset. In a comparison of 8 alignment methods, including CATS (Štajner et al., 2018), MASSalign (Paetzold et al., 2017), and LHA (Nikolov and Hahnloser, 2019), LHA achieved the best F1-Score. The same holds as previously reported for the translated Wikipedia corpus, LHA (and the other alignment methods) shows again the same problems of many misalignments (lower precision than recall) and missing capability of aligning *n*:*m* sentence pairs (also see percentage of *n*:*m* pairs in Table 4.8 and Table 4.9).

With the best-performing alignment method on the APA data, i.e., LHA, they automatically aligned the remaining document pairs. In total, APA-LHA-OR:B1 results in 10,268 parallel sentence pairs and APA-LHA-OR: A2 in 9,456 sentence pairs (Spring et al., 2021; Ebling et al., 2022). As APA-LHA contains alignments between the original texts and the simplified texts, I expect rather strong simplifications in the corpus. By definition, the simplifications of the original texts in CEFR level A2 are stronger than those in level B1. I could justify this based on FRE-scores: the distance between the FRE scores of the original and simplified texts is higher for OR-A2 (distance: ~24) than for OR-B1 (distance: ~18) (see Table 4.8 and Table 4.9). The automatically aligned sentence pairs are made available via Zenodo[35], however, the document pairs and the manually aligned sentence pairs are not available.

An analysis of the alignments of the APA-LHA alignments by Stodden et al. (2023) has shown that the alignments made by LHA are error-prone and should be used with caution. As the automatically aligned data are part of the training, validation, and test set, also the generated outputs on these test data (even if trained on different training data) should be interpreted with caution.

			Complex			Simple		
	# Sent. Pairs	*n*:*m*	**FRE↓**	**Sent. Len.↑**	**Word Len.↑**	**FRE↑**	**Sent. Len.↓**	**Word Len.↓**
APA-LHA-OR-A2-train	8,455	7.1 %	45.1	19.68	1.92	69.45	11.3	1.75
APA-LHA-OR-A2-test	500	6 %	44.7	20.2	1.92	69.55	11.27	1.78

Table 4.8: Characteristics of the sentence simplification corpus APA-LHA OR-A2. Own calculation. Sentence length in tokens. Word length in syllables.

35 https://zenodo.org/record/5148163 [last update: Septemeber 1, 2021; last access: July 24, 2024]

	# Sent. Pairs	*n*:*m*	Complex			Simple		
			FRE↓	Sent. Len.↑	Word Len.↑	FRE↑	Sent. Len.↓	Word Len.↓
APA-LHA-OR-B1-train	9,268	6.9 %	44.70	19.82	1.93	62.3	12.93	1.82
APA-LHA-OR-B1-test	500	8 %	43.7	20.48	1.93	62.6	12.82	1.83

Table 4.9: Characteristics of the sentence simplification corpus APA-LHA OR-B1. Own calculation. Sentence length in tokens. Word length in syllables.

4.3.5 APA-RST

Hewett (2023) also published a corpus based on news texts from the APA called "APA-RST". The main aim of creating the resource was not primarily to build a training or evaluation corpus for text simplification, but to analyze structural aspects regarding text simplification and text readability in parallel texts. To achieve this, they randomly selected 5 news documents from articles published between 2018 and 2022, each of which includes 5 news articles. Each article is available in three language levels: CEFR level A2, B1, and C2. They manually aligned sentence-wise each of the 25 news items between each language level, i.e., OR-A2, OR-B1, and B1 to A2.

For structural analysis, they also annotated the resulting sentence pairs with the Rhetorical Structure Theory (RST) framework (Mann and Thompson, 1988) (see component F in the corpus building process in Figure 3.1). Overall, APA-RST consists of 75 parallel documents (25 per OR-B1, OR-A2 and B1-A2) and 393 sentence pairs (OR-B1: 128 pairs, OR-A2: 112, B1-A2: 153). From the statistics of this small sample (see Table 4.10), we can infer that, as expected, the texts in B1 and A2 are more similar to each other than either of them to the original texts. Hence, the simplification between B1 and A2 seems to be less strong than between OR and B1 or OR and A2. First, more information has been deleted in the simplification from OR to B1 (415 deletions) and OR to A2 (429 deletions) than from B1 and A2 (26 deletions). And B1 to A2 contains fewer syntactical changes than in the other pairs. Furthermore, simplifications from OR to A2 contain more *n*:*m* alignments than from OR to B1 and from B1 to A2 (see Table 4.10). In future work, it would be interesting to calculate these statistics also for APA-LHA and compare them against the statistics of APA-RST to verify the (automatic) alignment quality of APA-LHA.

Level	1:1 (rephrased)	1:1 (copied)	1:1 (total)	*n*:1	1:*m*	*n*:*m*	*n*:*m* (all)	1:0	0:1	total (original)	total (simple)
APA-RST OR-B1	97	1	98	11	18	1	30	415	36	555	183
APA-RST OR-A2	65	0	65	10	33	4	47	429	54	555	203
APA-RST B1-A2	116	4	120	4	29	0	33	26	21	183	203

Table 4.10: Characteristics of the sentence simplification corpus APA-RST. Extended version of Table 3 in Hewett (2023) with additional own calculations.

4.3.6 20Minuten

20Minuten (Rios et al., 2021) is a Swiss German news corpus for document simplification. It is a news TS corpus which is intended to be used for training text summarization and text simplification models. The split of the data into training, development, and test data is publicly available.[36] Following the FRE scores for the corpus, there is no huge difference between the

readability of the complex and simplified documents (see Table 4.11); the readability of both texts can be interpreted as "on average". In comparison to Klexikon (see Subsection 4.2.5), the sentences in the complex documents are much shorter in 20Minuten than in Klexikon.

This corpus was extended by Kew et al. (2023), but they put the summarization task more in the foreground. For this corpus, no sentence-wise alignments exist; it is predominately designed for document summarization and document simplification.

			complex			simple		
	# Doc. Pairs	$n:m$	**FRE↓**	**Sent. Len.↑**	**Word Len.↑**	**FRE↑**	**Sent. Len.↓**	**Word Len.↓**
20Minuten-test	200	100%	54.45	16.82	1.8	52.1	12.27	1.9

Table 4.11: Characteristics of the document simplification corpus 20Minuten. Own calculation. Sentence length in tokens. Word length in syllables.

4.4 Corpora with Medical & Health Texts

In the previous sections, I have so far only discussed simplification from standard to simplified German. Another purpose of text simplification is to simplify texts from an expert language to standard language. Medical, clinical and health texts represent a prominent example of this type of expert-to-laypeople simplification, given that they often contain a considerable amount of technical jargon and are therefore challenging for many individuals to comprehend (Baumert, 2018). To overcome this, much research has emerged in the direction of simplification of medical texts in recent years. As mentioned above, the task of text simplification also has a high overlap with the task of text summarization. In the genre of medical corpora, both tasks are often combined to *plain text summarization* (Goldsack et al., 2023).

The medical and health domain also distinguishes itself from other domains as it can be described as a "safety critical domain". Wrong content-related translations can have an effect (of different extents) on the health condition of the readers, e.g., if the dose, name or regularity of a medication is wrongly translated or simplified (Canfora and Ottmann, 2020). Furthermore, the relevance and impact of simplification of health information have especially gained awareness during the COVID-19 pandemic. Updating the population of current protection measures or protection by vaccination has been distributed in many different language varieties to reach a wide audience, e.g., in expert language (e.g., for an overview see Abd-Alrazaq et al. 2021), standard German (e.g., see Robert Koch-Institut 2024a or Bundeszentrale für gesundheitliche Aufklärung 2024), German Plain Language (e.g., see Forschungsstelle Leichte Sprache 2023), or German Easy Language (e.g., see Robert Koch-Institut 2024b or Task Force Corona Leichte Sprache et al. 2024).

4.4.1 Non-German Corpora

There are some document simplification corpora in some languages, for example, the CLEAR corpus in French (Grabar and Cardon, 2018), the CLARA-MeD corpus in Spanish (Campillos-

36 https://github.com/ZurichNLP/20Minuten [last update: August 17, 2023; last access: July 24, 2024]

Llanos et al., 2022), the BioLaySumm corpus (Goldsack et al., 2022) or the PLABA corpus which are both in English (Attal et al., 2023). For more information, I direct the interested reader to the original papers.

4.4.2 German Resources

Simplifications of medical texts are also becoming more common in German. For example, during the COVID-19 pandemic many pieces of information were also published in simplified language to inform as many people as possible about the current situation, for example, the latest information on COVID as a disease and its symptoms as well as health rules and suggestions. On the webpage "Corona Leichte Sprache" (Corona German Easy Language)[37], for example, many pieces of information in German Easy Language have been published. However, the intention of this webpage is to inform people in German Easy Language and not to provide a translation of information that was written in standard German. Hence, no parallel documents are available.

On another webpage, i.e., "Apotheken Umschau" (pharmacy review)[38], diseases, symptoms, medications, and the health system are described for easy access to health information. Even if these texts are already written in standard German including jargon, they are also provided in a German Plain Language version to be accessible to more people. The documents in German Plain Language are linked to the standard German documents and, hence, easily retrievable parallel data. This resource has been included in the Simple German Web Corpus '23 (see Subsection 4.1.4), even though it is restricted with copyright.

More parallel health information in standard German and German Plain Language is also provided by the "Bundeszentrum für Ernährung" (BZFE, Federal Center for Nutrition)[39]. Although the parallel articles are also linked with each other and are published under an open license, these data have not yet been included in text simplification corpora.

4.4.3 Simple-Patho

Focusing on medical data and German text simplification, this domain is not a prominent resource in TS as for other languages; I am only aware of one resource. Trienes et al. (2022) introduce the so far only German TS corpus containing medical data, i.e., a document simplification corpus called Simple-patho. The corpus contains 851 clinical notes that were manually simplified by medical students for laypeople and patients. The corpus is aligned on the document level and also on the paragraph level (n =3,280). Following the reported FRE scores in Trienes et al. (2022) (see Table 4.12), in contrast to the previous corpora, the text and sentence length increases through simplification. This effect is due to rephrasing notes into full sentences and explaining medical terminology (Trienes et al., 2022). Following this, the FRE score of the simplified texts (in standard German) is still lower (equals more complex) than the FRE scores of the complex texts of the news or web domains (except TextComplexityDE). Currently, the corpus is not available due to data privacy concerns[40].

37 https://corona-leichte-sprache.de/page/6-startseite.html [last access: July 24, 2024]
38 https://www.apotheken-umschau.de/krankheiten-symptome/
39 https://www.bzfe.de/einfache-sprache/ [last access: July 24, 2024]

Text Unit	# Par. Pairs	n:m	Complex FRE↓	Complex Sent. Len.↑	Complex Word Len.↑	Simple FRE↑	Simple Sent. Len.↓	Simple Word Len.↓
document	851	*n/a*	32.9	14	*n/a*	40.30	16	*n/a*
paragraph	3,280	*n/a*	27.65	13	*n/a*	40.05	16	*n/a*

Table 4.12: Characteristics of the paragraph simplification corpus Simple-Patho. The values are copied from Trienes et al. (2022).

	FRE	Sent. Len.	Word Len.
original	30	16.84	7.16
German Plain Language	39	11.61	7.04
German Easy Language	59	7.44	6.74

Table 4.13: Characteristics of the paragraph simplification corpus Online Participation Corpus. Own calculation.

4.5 Corpora with Political & Legal Texts

Another example of expert-laypeople text simplification is the simplification of political texts. Political and legal texts comprise a wide spectrum including, e.g., laws, contracts, political debates, reports of a parliament, election manifestos, or also governmental content. They all have in common that they contain legal jargon, long and complex sentences, are legally binding, and have a large target audience (Schomacker et al., 2023b). Following Garimella et al. (2022), English legal texts are more difficult than Wikipedia texts wrt. sentence length, parse tree height, and readability. Although legal texts are attested to be very difficult to read, they are rarely simplified and, therefore, no parallel legal simplification corpus has existed in recent years (Schomacker et al., 2023b). A few approaches regarding automatic simplification of legal texts exist (e.g., Garimella et al. 2022 or Cemri et al. 2022), but they are using unsupervised methods and are evaluated with reference-less metrics and human assessments.

4.5.1 German Resources

For German, I am aware of a few resources with political and legal content. The Simple German Corpus (Jach, 2020) contains a few simplified documents for children and simplified to German Easy Language. However, no parallel documents in standard German and jargon are available for these resources. But Klepp (2022a) make of this data in the scope of text simplification even if not using parallel data: their web scraper downloads a dictionary of political terms from the Bundeszentrale für politische Bildung ((German) Federal Agency for Civic Education) in two versions: In German Plain Language[41] and in standard German[42]. Unfortunately, they do not align the documents with each other, and, hence, the resource in its current state cannot be used for text simplification yet.

40 https://github.com/jantrienes/simple-patho [last access: July 24, 2024]

41 https://bpb.de/kurz-knapp/lexika/lexikon-in-einfacher-sprache/ [last access: July 24, 2024]

42 https://bpb.de/kurz-knapp/lexika/politiklexikon/ [last access: July 24, 2024]

Other materials such as election manifestos[43] or the basic law[44] have been manually translated into simplified German and are also available in a parallel version in more complex German, but these texts have not yet been made available as parallel text simplification resources.

4.5.2 ABGB

Bydlinski (2015) and Kaban and Krottmaier (2023) propose a legal corpus of the Austrian General Civil Code (DE: "Allgemeines bürgerliches Gesetzbuch", abbreviation: "ABGB") with manually simplified versions for law students and laypeople. It contains three language versions for most of the ABGB paragraphs, that is, "original text", "suggested text", and "alternative text". The original text contains the legally binding version in complex German partially written in the year 1812. The suggested text is a reformulation of the original text without changes in content but simplified into the Austrian German Plain Language, i.e., Klarsprache, (Bydlinski, 2015). The alternative text is again a simplification of the suggested text with additional content additions and deletions resulting in a lower meaning preservation of the original text.

Meister (2023) have recently made this resource available with sentence-wise alignments[45]. Overall, ABGB contains 448 sentence pairs for which two alternative simplifications are available, making it an ideal resource for evaluating German sentence simplification in the legal domain.

4.5.3 Online Participation

Gutermuth (2020a) have linguistically evaluated the understanding of instructions of an online participation process regarding a German transparency law in three language levels, i.e., standard German, professional simplification into German Plain Language, and professional simplification into German Easy Language. They evaluated the comprehension of different people of the target groups regarding all three text levels. As can be seen in Table 4.13, the complexity wrt. FRE, sentence length and word length decrease for each level. The texts of their analysis Gutermuth (2020b) are aligned per paragraph. Unfortunately, these data have not been used yet for any text simplification experiments.

43 Examples of election manifestos in German Easy Language are: https://csu.de/common/download/KM_Broschuere_Leichte_Sprache_BTW_2021_Ansicht.pdf, https://spd.de/fileadmin/Dokumente/Europa_ist_die_Antwort/SPD_Wahlprogramm_Europa_Wahl_Leicht.pdf, https://cms.gruene.de/uploads/assets/2019_Europawahl-Programm_LeichteSprache.pdf, https://die-linke.de/fileadmin/download/wahlen2019/wahlprogramm_leichte_sprache/wahlprogramm2019_leichte_sprache_neu.pdf, or https://fdp.de/sites/default/files/2021-08/FDP_BTW2021_KWP_leichteSprache.pdf [all last accessed: July 24, 2024].

44 Examples of the basic law in German Easy or Plain Language are: https://bamf.de/SharedDocs/Anlagen/DE/LeichteSprache/leichte-sprache-grundgesetz.pdf, https://hurraki.de/wiki/Artikel_des_Grundgesetzes, or https://nachrichtenleicht.de/das-grundgesetz-100.html, or https://bpb.de/shop/materialien/einfach-politik/236587/das-grundgesetz-die-grundrechte/ [all last accessed: July 24, 2024].

45 https://github.com/MeisterFa/ABGB-TextSimplification-Datasets [last update: September 24, 2023; last access: July 24, 2024]

4.6 Corpora with Narratives Texts

In contrast to the domains described thus far, narrative texts are written with the intention of entertaining the reader, rather than informing them, as is the case with medical or news texts (Mar et al., 2021). Examples of narrative or fictional texts are stories, novels, classical literature, or fairy tales. Following Mar et al. (2021), narrative texts are overall easier to comprehend for adults than informative texts (e.g., news or medical texts) due to their analogy to everyday experiences of people. However, despite being more simple, narrative texts are an interesting domain for text simplification, as through simplification literary classics can be made accessible, for example, for non-native speakers or children. Especially, fairy tales or other stories written in previous centuries might contain vocabulary or sentence structures that are uncommon and difficult to read nowadays.

Furthermore, the simplification process of narrative texts includes other strategies than in other domains. As for many fictional texts, the author's style is an important characteristic; translators try to transfer the style to some extent also to the simplified version. Fictional texts often contain a high degree of implicitness, which is made more explicit during translation. Consequently, narrative TS corpora consist of many 0:1 (addition) and 1:n (splitting or extension) alignment pairs. Therefore, the manual and automatic alignment gets more complicated as the complex and simplified documents are rather comparable than truly parallel. Narrative text simplification contains more strong rewriting than, for example, simplification of news.

4.6.1 Non-German Corpora

There are a few parallel corpora with narrative texts for TS: For example, RuAdapt, a Russian corpus with simplified books (Dmitrieva and Tiedemann, 2021), or Italian novels (Brunato et al., 2015) or Italian stories for children (Barlacchi and Tonelli, 2013). For further details, I direct the interested reader to the original papers.

4.6.2 German Resources

There are some German publishers who release only books which have been manually simplified from books in standard German, e.g., Spaß am Lesen Verlag (EB; simple books)[46], Passanten Verlag (PV)[47], and Kindermann Verlag (KV). Further, the North German Broadcasting Corporation (Norddeutscher Rundfunk, NDR) has simplified a few famous fairy tales into German Easy Language[48]. However, there are many more publishers of simplified language or children's books. Often, these books have been written directly in simpler language and have not been adapted from more complex texts, so no parallel documents are available.

46 https://einfachebuecher.de [last access: July 24, 2024]

47 https://www.passanten-verlag.de/ [last access: July 24, 2024]

48 https://www.ndr.de/fernsehen/service/leichte_sprache/Maerchen-in-Leichter-Sprache,maerchenleichtesprache100.html [last access: July 24, 2024]

4.6.3 GNATS

For German, the document simplification corpus GNATS proposed by Schomacker et al. (2023a) made use of the previously named resources. This corpus contains four subcorpora, i.e., German simplified fairy tales in MILS (Märchen in Leichter Sprache; fairy tales in German Easy Language)[49], and books of three publishers in simplified German, i.e., Spaß am Lesen Verlag (EB; simple books)[50], Passanten Verlag (PV)[51], and Kindermann Verlag (KV). All publishers have different reader groups in their addressed audience: While the fairy tales are simplified for people with learning problems or people with dementia (or German Easy Language target group), EB especially addresses young readers and adults who cannot read well (Spaß am Lesen Verlag, 2024), PV addresses readers who like to read but have reading problems (Passanten Verlag, 2024), and KV simplifies for children (Kindermann Verlag, 2024). However, these publishers only make the simple versions available. The counterparts in standard German are available through the Gutenberg Project: The documents have been manually aligned to the complex-simple document pairs. This corpus does not overlap with any of the web corpora introduced before.

Overall, GNATS contains 33 parallel documents, which are also split into a train (27 documents), development (3 documents), and test set (3 documents). To the best of my knowledge, the documents are not aligned on the sentence level. The corpus as well as the URLs of the resources are available online.[52]

4.7 Non-Parallel Corpora

As previously discussed, for some simplified resources no parallel complex version exists. However, this simplified, non-parallel or monolingual data can also be helpful in the scope of text simplification besides parallel training or evaluation pairs. In this section, I introduce non-parallel resources for TS, i.e., lexical simplification corpora that do not require a parallel format (see Subsection 4.7.1), data augmentation strategies to gather more data for structural simplification (see Subsection 4.7.2), and more resources of monolingual data (see Subsection 4.7.3), which can be helpful for data augmentation.

4.7.1 Lexical Simplification Data

Prominent examples of non-parallel data for text simplification are corpora for lexical simplification. Lexical simplification does not require parallel data, as its subtasks require complex texts with annotations regarding their complexity (e.g., see Shardlow et al. 2021 or Mohtaj et al. 2022) or regarding complex words and their simpler substitutes (e.g., see Yimam et al. 2018 or Shardlow et al. 2024).

49 https://www.ndr.de/fernsehen/service/leichte_sprache/Maerchen-in-Leichter-Sprache,maerchenleichtesprache100.html [last access: July 24, 2024]

50 https://einfachebuecher.de [last access: July 24, 2024]

51 https://www.passanten-verlag.de/ [last access: July 24, 2024]

52 For data see: https://github.com/tschomacker/aligned-narrative-documents [last update: September 16, 2023; last access: July 24, 2024]; for URLs see: Table 2 in Schomacker et al. (2023a)

Regarding complex word identification (CWI), a few shared tasks have been conducted, including new resources for this task, e.g., CWI shared task 2016 (Paetzold and Specia, 2016a) or the CWI shared task 2018 (Yimam et al., 2018). While the first tackles only English CWI, the second considers also German and Spanish. Furthermore, the CWI shared task 2018 aims at the automatic identification of the complexity of words, in more detail if a word within a sentence is complex for native and non-native speakers. However, this shared task has not aimed at predicting substitutions of complex words with simpler synonyms.

For lexical complexity prediction, i.e., estimating the complexity of a sentence on a scale from 0 to 1 for people with special needs (e.g., foreign language learners), there are also a few resources, that is, an English dataset (Shardlow et al., 2020, 2021), a German dataset (Mohtaj et al., 2022), and very recently a multi-lingual dataset (including the following languages: English, Spanish, French, Brazilian Portuguese, Bengali, Sinhala, Filipino, Japanese, Italian, Catalan, and German) (Shardlow et al., 2024).

However, all of these resources cannot be used for simplification as a sequence-to-sequence task, as they do not contain substitutes for the complex words which could be used to build complex-simple sentence pairs. I am not aware of any other German resource that focuses especially on lexical simplification.

4.7.2 Syntactical Simplification Data

As mentioned previously, the BiSECT corpus (see Subsection 4.1.5) is designed for syntactical simplification because each sentence pair includes at least one split or merge operation. I am not aware of any other German resource that focuses especially on syntactical simplification.

4.7.3 Monolingual Data

In recent years, a few monolingual corpora have been proposed that contain news articles in German Easy Language, i.e., LeiKo (Jablotschkin and Zinsmeister, 2020) and SNIML (Hauser et al., 2022). SNIML additionally contains simplified news in other languages, i.e., French, Italian, Finnish, Swedish, and English. Further, Anschütz et al. (2023) and Asghari et al. (2023) also propose corpora which contain only texts in German Easy Language, but their texts originate from web documents with different subdomains. Other corpora with web documents and mixed domains are the corpus proposed by Klepp (2022a), the KED corpus (Jach, 2023), or the corpus proposed by Klöser et al. (2024). An overview of simplified German monolingual corpora can be found in Table 4.14.

Reference	Name	Target Simple	Domain	# Docs	# Sents
Jablotschkin and Zinsmeister (2020)	LeiKo v1.5	German Easy Language	news	216	5,961
Hauser et al. (2022)	SNIML	German Easy Language	news	303+	8,136
Klepp (2022a)	-	German Easy & Plain Language	web	81,928	-
Anschütz et al. (2023)	-	German Easy Language	web	-	544,467
Asghari et al. (2023)	LSWeb23	German Easy Language	web	-	-
Jach (2023)	KED 2.0	simplified German	mix	7,098	-
Klöser et al. (2024)	-	German Easy & Plain Language	8,130	-	

Table 4.14: Corpora with non-parallel simplified German data.

Some resources of monolingual corpora are used in more than one of the named corpora. An overview of how the corpora overlap can be found in Table 4.15.[53] News of the news agencies of nachrichtenleicht[54], and NDR[55] are included in LeiKo, as well as in the corpus by Anschütz et al. (2023), but the latter also includes several other web resources. In addition, the corpora of Klepp (2022a) and Anschütz et al. (2023) overlap in three resources, whereas the KED corpus is quite different from the other corpora, as it contains many other resources.

Comparing the list of web pages of the parallel web corpora with the monolingual corpora, I can also see some overlaps: for example, texts from brandeins, Lebenshilfe Main-Taunus are used in the simple German web corpus '23 (see Subsection 4.1.4) as well as the monolingual selection of Anschütz et al. (2023). Further, the Klexikon articles have been used to build the parallel corpora of lexica and Klexikon (see Subsection 4.2.5) as well as part of the monolingual corpora of Klepp (2022b) and KED 2.0.

In the next section, I will show how these corpora can be used as augmented data for TS.

4.8 Data Augmentation

As shown in the previous sections, in some cases the number of available complex-simple sentence pairs is too low to train a text simplification model, especially if training from scratch and not fine-tuning a pre-trained model. In natural language processing, some strategies are proposed on how to cope with limited data, for example, data augmentation, few-shot and zero-shot learning, or transfer learning[56].

In general, data augmentation is a technique to enrich a training dataset in its size and diversity by adding more training samples, but without labeling (or aligning) more data (Feng et al., 2021). This strategy can counteract the overfitting of an NLP model on a small training set (Feng et al., 2021). Following Yang et al. (2022), general data augmentation or data synthesis strategies are, for example, increasing the existing data but replacing, inserting, or deleting some words, or paraphrasing the existing data by translating the data into another language and back-translating it into the original language (also called round-trip translation or back-and-forth translation Gaspari, 2006).

These approaches can also be applied to TS, for example, Palmero Aprosio et al. (2019) proposed two techniques: oversampling (multiplying the gold data several times), and training a simple-to-complex model on the gold data and using the generated more-complex sentences as the source part of a new complex-simple pair. A method to generate synthetic data for syntactical simplification has been proposed by Maddela et al. (2021), they automatically split sentences into smaller parts and use them as additional 1:*m* sentence pairs. In the following, I summarize strategies which have been applied to augment German TS data.

53 Unfortunately, I can not include all corpora listed in Table 4.14 also in Table 4.15, because information regarding the exact resources for these corpora is missing.

54 https://www.nachrichtenleicht.de/ [last access: July 24, 2024]

55 https://www.ndr.de/fernsehen/barrierefreie_angebote/leichte_sprache/index.html [last access: July 24, 2024]

56 For more information regarding few-shot and zero-shot approaches I refer to Section 6.5 and for transfer learning to Section 6.4.

Subcorpus	Website Simple	Simple	Domain	Description	LeiKo 1.5	Klepp	Anschütz et al.	Jach
Einfach Politik: Lexikon	bpb.de/kurz-knapp/lexika/lexikon-in-einfacher-sprache/	PL	politics	dictionary of political terms		x		
kurier.at	kurier.at/einfache-sprache	PL	news	news			x	
nachrichtenleicht	nachrichtenleicht.de	PL	news	news	x	x	x	
Arbeit & Gesundheit	aug.dguv.de/leichte-sprache/	EL	health	health information		x		
Bibel in EL	evangelium-in-leichter-sprache.de	EL	bible	bible texts		x		
brandeins	brandeins.de/themen/rubriken/leichte-sprache	EL		Translating excerpts from various topics			x	
Einfachstars	einfachstars.info/	EL	gossip	gossip of movie stars, sport stars and fashion		x	x	
Hurraki	hurraki.de	EL	wikipedia	dictionary of many topics		x	x	
Lebenhilfe	lebenshilfe.de	EL	accessibility	Non-profit association for disabled people			x	
Lebenshilfe Main Taunus	lebenshilfe-main-taunus.de	EL	accessibility	Non-profit association for disabled people			x	
MDR Nachrichten	mdr.de/nachrichten-leicht	EL	news	State-funded public broadcasting service		x		
MDR Wörterbuch	mdr.de/nachrichten-leicht/woerterbuch/	EL	wikipedia	dictionary of many topics			x	
NDR Märchen	ndr.de/fernsehen/barrierefreie_angebote/leichte_sprache/Maerchen-in-Leichter-Sprache,maerchenleichtesprache100.html	EL	fiction	Fairytales in EL		x		
NDR Nachrichten	ndr.de/fernsehen/barrierefreie_angebote/leichte_sprache	EL	news	State-funded public broadcasting service	x		x	
Hanisauland	hanisauland.de	children	politics	Texts, encyclopedia entries, book reviews, etc. on political topics				x
Klexikon	klexikon.zum.de	children	Wikipedia	Online encyclopedia for children		x		x
Labbe	labbe.de/lesekorb	children	fiction	Stories, fairy tales, and handicraft & game instructions				x
Oekoleo	oekoleo.de	children	nature	nature and environmental protection				x
rossipotti	rossipotti.de	children	fiction	Online literary magazine				x
SimplyScience	simplyscience.ch	children	science	Texts and encyclopedia entries on scientific topics				x
Rechte-einfach	mixed	mixed	politics	simplified texts on legal topics and laws				x

Table 4.15: Characteristics of simplified German resources per web crawler. PL = German Plain Language, EL = German Easy Language. All URLs have lastly been accessed at July 24, 2024.

4.8.1 Word Replacement

In the word replacement strategy, random words in source texts (and also in a target) are replaced with synonyms to create slightly different sentences (e.g., see Wang et al. 2018). To the best of my knowledge, no word replacement strategies have been applied to create a German TS corpus so far. In my opinion, this strategy seems not to be ideal for text simplification, as we do not know if the alternative word will have the same, higher, or lower complexity than the original word. In the last case, a TS model would learn from this instance how to adapt a text by increasing its complexity, which would result in the opposite of the intended task.

Therefore, before generating synthetic data by replacing words, I recommend evaluating the generated data on the basis of its complexity before using it for training. Only if a complexity measure would give a higher score for the new augmented data than for the simplified text, I would consider this data point as a relevant augmented complex sentence.

4.8.2 Translation & Round-Trip Translation

Similar to problems in machine translation (Saunders, 2022), in text simplification, often monolingual data in the language of interest (here simplified or complex data) are available in large quantities, whereas parallel (here sentence-wise aligned) data is less available. This can be counteracted by translating available TS resources from other languages or by round-trip translation of small parallel corpora to increase them. Data synthesis by translation or round-trip translation has been applied in some German TS studies yet.

Translation In the previous section, I have already introduced some synthetic corpora which have been generated with the help of machine translation, i.e., BiSECT (Kim et al., 2021) (see Subsection 4.1.5), the translated Wikipedia corpus by Ebling et al. (2022) (see Subsection 4.2.3), and the translated ASSET corpus by Schlippe and Eichinger (2023) (see Subsection 4.2.4).

Round-Trip Translation In addition to the previously introduced German News corpus (see Subsection 4.3.3), Säuberli et al. (2020) propose to back-and-forth-translate the simplified sentence (target) of a source-target pair (or complex-simple pair) into another language and back to German. The back-translated target sentence is then used as the source sentence of the pair in which the target sentence remains the same. They call this process BT2TRG. They compare this approach with other data augmentation strategies such as adding simple-simple pairs (they call it TRG2TRG) in which the model sees new simple pairs and learns to copy them, or empty-simple pairs (they call it NULL2TRG). In their experiments with German news data, the BT2TRG and NULL2TRG approaches achieved worse results with respect to SARI and BLEU than the baseline without augmented data. But, the TRG2TRG approach improved the baseline and achieved the best results of all approaches.

However, in an additional human evaluation, they found that in all approaches their models mostly preserved no content. Furthermore, the generations of the system with BT2TRG and TRG2TRG were less fluent than the model trained on only gold complex-simple pairs. The simplicity of the generated sentences was rated similarly for all approaches. More experiments regarding round-trip translation and simple-simple pairs for German TS are required to finally assess whether this augmentation strategy is helpful or harmful for text simplification.

4.8.3 Monolingual Data

As briefly discussed in Subsection 4.7.3, some corpora contain simple German texts, but there are no parallel or comparable versions in standard German (or a more complex version). This data can be used, for example,

- to train or fine-tune a language model to generate texts in simplified language (see Anschütz et al. 2023), or
- as additional simple-simple pairs (TGR2TRG).

Anschütz et al. (2023) used monolingual simplified data to fine-tune an autoregressive language model as a basis for a text simplification model (for more information, see Subsection 6.6.1). For sentence simplification, resources for such kinds of simplified data can be docu-

ment simplification corpora which are not aligned on the sentence level, non-parallel simplified German corpora, or texts which can be scraped from German websites. Examples of web scrapers for monolingual data are summarized in Subsection 4.7.3.

As mentioned above, Säuberli et al. (2020) additionally trained their model on simple-simple pairs (TRG2TRG) by using the same simplification twice, i.e., in a complex-simple pair and the simple-simple pair. However, instead of using the aligned simple sentence, non-aligned simple sentences could also be used for this approach to let the model see more different simple sentences, which need no simplification operation (except copying). To the best of my knowledge, this strategy has not yet been employed.

4.9 Summary & Outlook

In this section, I have described the current state of resources for German text simplification, i.e., 40 TS corpora (see Table 4.16) and additional monolingual data. I introduced ten parallel document simplification corpora (see *n/a* in # Pairs column in Table 4.16) of five different text genres (Wikipedia texts, news texts, web texts, medical texts, and narrative texts) and written for seven different target audiences. For paragraph simplification, I have presented 4 corpora, i.e., simple-patho and three versions of Online Participation. I have also described 30 parallel corpora for sentence simplification (see Table 4.16 with a number in # Pairs column) of seven domains and written for 12 different target groups (when combining mixed target groups into one category).

4.9.1 Challenges & Research Gaps

However, all corpora have some advantages and disadvantages. Considering the most commonly used German sentence simplification corpora (see Figure 4.1), we can see imbalances regarding the size of the corpora (see ZEST vs. APA-LHA), the alignment type of the corpora (see TextComplexityDE vs. APA-LHA) as well as one-domain vs. many-domain corpora (see APA-LHA vs. Simple German Corpus '23). In the following, I briefly summarize the challenges of the corpora that have been identified in the previous presentation of the corpora.

These challenges serve to illustrate the relevance of providing new contributions to my research questions regarding the imbalance of domains and target groups in current resources (see RQ 4-1) and the relevance of new data (see RQ 4-2).

4.9.1.1 Availability

The availability of corpora is a relevant issue (further called Data Challenge B); of all 40 corpora, 16 corpora are not available at all, and 3 corpora are only available upon request due to copyright restrictions. Reasons for restrictions are, for example, data privacy, copyright issues, or link rot. The lack of availability hampers the reproducibility of previous studies as well as the progress in the field of German TS. Therefore, new corpora are required that have an open license or are otherwise accessible.

Reference	Name	Target Simple	Domain	Available	# Docs	# Pairs	Aligned	Split
Klaper et al. (2013)	SGC '13	EL	web	on request	256	1,888	manual	*n/a*
Battisti et al. (2020)	SGC '20	CEFR A2	web	*n/a*	36	1,080	manual	train.+val.
Battisti et al. (2020)	SGC '20	CEFR A2	web	*n/a*	378	n/a	automatic	train.+val.
Toborek et al. (2023)	SGC '23	EL + PL	web	available	39	391	manual	train.+val.
Toborek et al. (2023)	SGC '23	EL + PL	web	available	700	5,942	automatic	train.+val.
Kim et al. (2021)	BiSECT	German learner	web	available	*n/a*	186,237	automatic	train.+val.
Siegel et al. (2019)	leichte-sprache-corpus	mixed	web	available	351	*n/a*	*n/a*	val.
Hansen-Schirra et al. (2021)	GEASY	EL	web	*n/a*	93	*n/a*	*n/a*	*n/a*
Jablotschkin et al. (2024)	DE-Lite	mixed	web	not yet	8,000	*n/a*	*n/a*	*n/a*
Spring et al. (2021)	Capito-B1	CEFR B1	web	*n/a*	12	426	manual	train.+val.
Spring et al. (2021)	Capito-A2	CEFR A2	web	*n/a*	8	412	manual	train.+val.
Spring et al. (2021)	Capito-A1	CEFR A1	web	*n/a*	22	416	manual	train.+val.
Spring et al. (2021)	Capito-B1	CEFR B1	web	*n/a*	1,055	54,224	automatic	train.+val.
Spring et al. (2021)	Capito-A2	CEFR A2	web	*n/a*	1,546	136,582	automatic	train.+val.
Spring et al. (2021)	Capito-A1	CEFR A1	web	*n/a*	839	10,952	automatic	train.+val.
Säuberli et al. (2024)	Capito-A2	CEFR A2	web	*n/a*	*n/a*	*n/a*	*n/a*	train.+val.
Naderi et al. (2019)	TextcomplexityDE	German learner	wiki	available	23	265	simplified	val.
Spring et al. (2023)	Wikipedia-Corpus	CEFR A2	wiki	*n/a*	198	1,382	manual	*n/a*
Ebling et al. (2022)	Wikipedia-Corpus	CEFR A2	wiki	*n/a*	106,126	*n/a*	automatic	*n/a*
Schlippe and Eichinger (2023)	Translated ASSET	*n/a*	wiki	*n/a*	*n/a*	1,000	simplified	train.+val.
Hewett and Stede (2021)	Lexica-klexikon	children 6-12	wiki	available	1,090	*n/a*	*n/a*	*n/a*
Hewett and Stede (2021)	Lexica-miniklexikon	children ≤ 6	wiki	available	1,090	*n/a*	*n/a*	*n/a*
Aumiller and Gertz (2022)	Klexikon	children 6-12	wiki	available	2,898	*n/a*	*n/a*	train.+val.
Mallinson et al. (2020)	geolino	children 5-7	wiki	available	20	1,198	simplified	val.
Säuberli et al. (2020)	German News Corpus	CEFR B1	news	*n/a*	*n/a*	3,916	automatic	train.+val.
Spring et al. (2023)	APA-LHA-OR-A2	CEFR A2	news	*n/a*	67	504	manual	n/a
Spring et al. (2023)	APA-LHA-OR-B1	CEFR B1	news	*n/a*	67	518	manual	n/a
Spring et al. (2021)	APA-LHA-OR-A2	CEFR A2	news	on request	2,300	9,456	automatic	train.+val.
Spring et al. (2021)	APA-LHA-OR-B1	CEFR B1	news	on request	2,300	10,268	automatic	train.+val.
Hewett (2023)	APA-RST	CEFR B1	news	available	25	128	manual	*n/a*
Hewett (2023)	APA-RST	CEFR A2	news	available	25	112	manual	*n/a*
Hewett (2023)	APA-RST	CEFR A2	news	available	25	153	manual	*n/a*
Rios et al. (2021)	20Minuten	general	news	available	18,305	*n/a*	*n/a*	train.+val.
Trienes et al. (2022)	simple-patho	laypeople	medical	not yet	850	3,280	simplified	train.+val.
Meister (2023)	ABGB-non-experts	laypeople	politics	available	1	448	manual	val.
Meister (2023)	ABGB-plain	PL	politics	available	1	448	manual	val.
Gutermuth (2020a)	Online Participation	PL	politics	available	1	13	simplified	*n/a*
Gutermuth (2020a)	Online Participation	EL	politics	available	1	13	simplified	*n/a*
Gutermuth (2020a)	Online Participation	EL + PL	politics	available	1	13	simplified	*n/a*
Schomacker et al. (2023a)	MILS+EB+PV+KV	mixed	narrative	available	33	*n/a*	*n/a*	train.+val.

Table 4.16: Summary of German document, paragraph, and sentence simplification corpora without own work. The lines separate the domains of the corpora. EL = German Easy Language, PL = German Plain Language. All URLs have lastly been accessed at July 24, 2024.

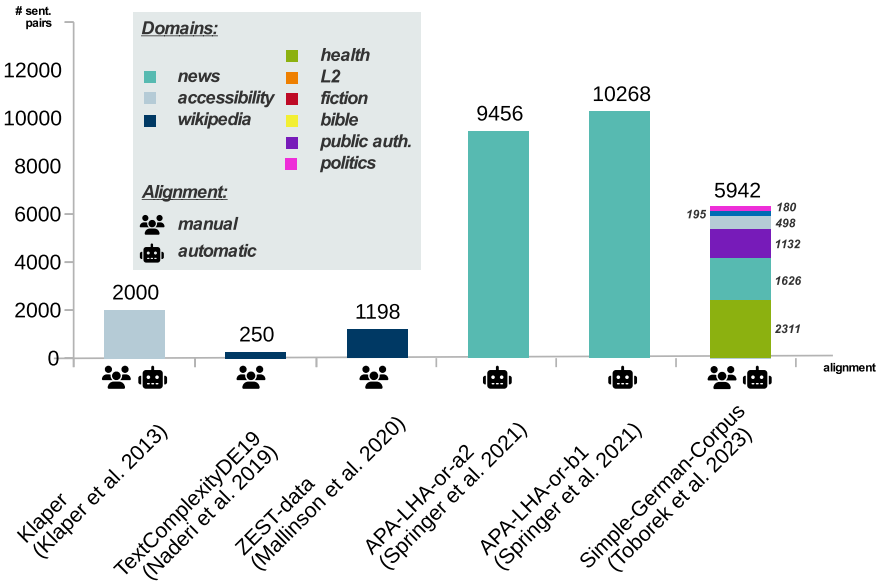


Figure 4.1: Common German sentence simplification corpora including alignment types and domains.

4.9.1.2 Encoding

A minor issue which tackles only one corpus, i.e., BiSECT (see Subsection 4.1.5), is error-prone encoding (further called Data Challenge C): in more detail, in this corpus, all diacritics are missing. Unfortunately, training and evaluation on BiSECT would not result in reliable results for German TS.

4.9.1.3 Size

Another challenge of using the named TS corpora is that their size is often too small to train a deep learning model for TS from scratch (further called Data Challenge D). As can be seen in Table 4.16, most sentence simplification corpora contain less than 5,000 pairs except Simple German Corpus '23 (see Subsection 4.1.4), APA-LHA (see Subsection 4.3.4), BiSECT (see Subsection 4.1.5), and the capito corpora (see Subsection 4.1.6). But, due to huge advances in the last years of NLP, the small corpora can be nowadays used to fine-tune or prompt pre-trained language models. Hence, the knowledge of pre-trained regarding language and other pre-trained tasks can be transferred to text simplification models.

Some of the document simplification corpora are of comparatively small size in terms of samples, but their size in terms of sentences and complex-simple sentence pairs would result in a comparatively large size for sentence simplification. To achieve this, a higher quality of automatic alignment or more support for manual simplification is required. Prior to our work, no corpus has been existing which is available on the document and sentence level and containing the same data (further called Data Challenge E).

4.9.1.4 Parallel vs. Comparable Content

However, not all corpora are well suited for document simplification and additionally sentence simplification. For example, in some corpora, the simple document is written independently of the complex document. In this case, the documents are comparable as they might contain similar content, but are not parallel, which would mean that the simple document was simplified with the complex document used as a prototype (further called Data Challenge F). Comparable documents might be more suitable for summarization or other related tasks than for simplification (further called Data Challenge H). Examples of comparable corpora are Klexikon (see Subsection 4.2.5) or Wikipedia translation (see Subsection 4.2.3).

As a consequence, if the documents are only comparable, the content between both versions might be so different that they cannot be aligned on the sentence level. This might also explain the worse results of automatic alignment methods on some corpora, e.g., on Wikipedia Translation (see Subsection 4.2.3).

4.9.1.5 Alignment

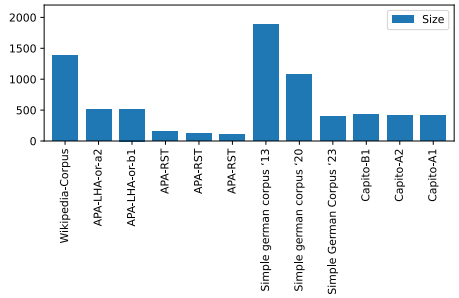
For corpora on which sentence-wise alignment is applicable, approaches on manual and automatic alignment have been shown (see step C in the corpus building process). For manual alignment, annotation guidelines are often missing or are not specifying which alignment types (only 1:1, only 1:m, or n:1, or all n:m?) are considered. In addition, technical support is required to facilitate manual alignment of n:m pairs (see Building Challenge B).

However, the high variation in alignment options makes alignment difficult, not only for humans, but also for algorithms (further called Data Challenge G). I showed that many algorithms are limited to some of the alignment types (e.g., only $1:1$ alignments). Previous work on building (German) TS corpora has already shown that

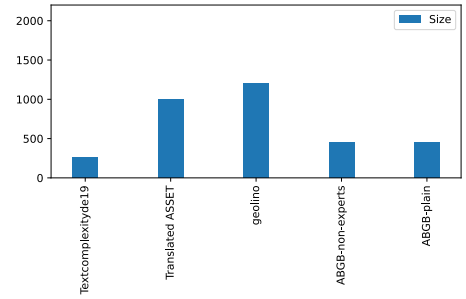
- alignment methods perform only well on nearly identical sentence pairs (Stajner, 2021),
- automatic alignment results in low accuracy even when aligning mild simplifications (see Spring et al. 2023), or
- a sentence-wise alignment is not suitable to some corpora as the documents are simplified independently and, hence, sentence-pairs with parallel meaning can not be identified (Aumiller and Gertz, 2022),
- but, manual alignment is very costly in human resources.

So far, newer alignment methods, such as the neural CRF approach of Jiang et al. (2020), have not yet been applied to German data.

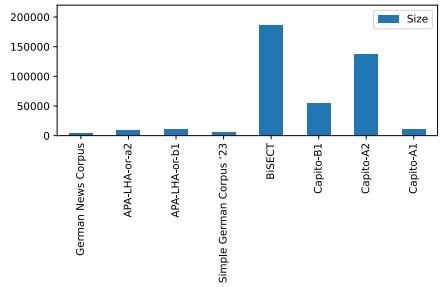
Although automatic alignment is often criticized, it is applied to many corpora (see Table 4.16) because it is a much faster way of aligning many sentence pairs than aligning them manually. In comparison of the alignment techniques, automatically aligned corpora are 100 times larger (up to 200,000 pairs) than manually aligned corpora or manually simplified corpora (up to 2,000 pairs; see Figure 4.2; pay attention to different scale sizes). On the one hand, to ensure high quality of large, automatically aligned corpora, more reliable and more flexible alignment methods are available. On the other hand, larger manually aligned sentence simplification corpora could also address the gap of missing large, high-quality corpora.



(**a**) Manually aligned.



(**b**) Manually simplified.



(**c**) Automatically aligned.

Figure 4.2: Corpus sizes of German sentence simplification corpora.

4.9.1.6 Own Simplification

Another strategy of building simplification corpora of non-parallel documents has been presented in a few corpora, i.e., to instruct crowd-workers or professional translators to simplify complex texts for the purpose of building a new simplification corpus (see step C in the corpus building process). Examples of corpora built following this approach are TextComplexityDE (see Subsection 4.2.6), GEOlino (see Table 4.7), or Online Participation (see Subsection 4.5.3).

However, the guidelines on how the translators have simplified the texts are often not available. Also, the quality of the translations is often not verified (e.g., for TextComplexityDE (see Subsection 4.2.6) or ABGB (see Subsection 4.5.2)). Hence, it is also not clear whether the simplifications are simple enough for a specific target group and if they show a variety of possible simplification operations (see Building Challenge D).

4.9.1.7 Simplification Operations

Following the corpus descriptions, only one corpus exists which especially addresses selected simplification operations, i.e., splitting and merging in BiSECT. However, the other parallel corpora do not contain further annotation except for the language level, domain, and sentence-wise alignment.

Annotation regarding the simplification quality or simplification operations included in the corpus is missing (see step F in the corpus building process). That is why the quality, the extent,

and the variety of the simplifications of the corpora are unclear (further called Data Challenge J). More insights on the extent of simplification, or operations made during the simplification, can be of high gain for the corpus (see e.g., Cardon and Bibal 2023 or Heineman et al. 2023).

For some corpora, especially the web corpora, the proxy is made that sentences written for children or non-native speakers are simpler than the counterpart in Standard German, but this is often not verified. In some cases, the readability is estimated with often criticized readability metrics. Following this, more annotations on simplification corpora are demanded.

Previous to this work, no German corpus has existed that includes the annotation of simplification operations and simplification quality per sentence pair in order to get more insights regarding the gold data.

4.9.1.8 Data Augmentation

I have also discussed a few strategies on how to deal with small corpora, e.g., by adding augmented or synthetic data. Previous studies have experimented with only a few strategies on a few datasets. Hence, no conclusions can be drawn yet about whether data augmentation or which data augmentation strategy is most helpful for German TS (see Model Challenge A). Nevertheless, when using additional datasets for training TS models, the quality of the simplification as well as the addressed target group should always be considered.

4.9.1.9 Target Group

Further, it should be paid attention to not mix target groups within a corpus or add texts of another target group as additional training data, because each target group requires different simplification operations (Siddharthan, 2014) (further called Data Challenge I). However, five of the 40 presented corpora mix the target group or language varieties (see Figure 4.3) as otherwise the number of simplification pairs would have been too few. The following corpora are examples with mixed data, e.g., Simple German Corpus '23 (see Subsection 4.1.4), leichte-prache-corpus (see Subsection 4.1.7), or GNATS (see Subsection 4.6.3).

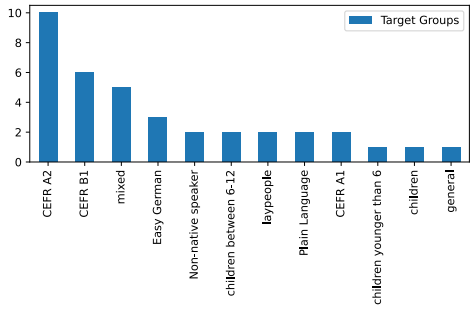


Figure 4.3: Target groups of all German TS corpora.

Furthermore, in the German TS corpora, German language learners are overrepresented with 27 of 40 corpora (CEFR A2: 10, CEFR B1: 6, non-native speaker: 2, CEFR A1: 2; German Plain Language: 2, mixed: 5; see Figure 4.3) whereas other target groups of simplification,

e.g., laypeople (laypeople: 2; simplified from jargon), or people with cognitive impairments (German Easy Language: 3, mixed: 5) are underrepresented. Following this, more and larger corpora with the focus on one target group or one language variety are required.

4.9.1.10 Evaluation Data

A few of the corpora are too small for training and fine-tuning, and they are only useful for validation: i.e., 5 of 30 sentence simplification corpora (with $< 1,2000$ complex-simple pairs) and 5 of 10 document simplification corpora (with $< 1,100$ complex-sentence pairs). For further 11 corpora, it is not clear whether and how they have been used for TS yet; they were originally built for other purposes, e.g., evaluation of automatic alignment methods for TS.

Furthermore, in comparison to English TS evaluation sets, which include multiple alternative simplifications for better evaluation, in German TS there are only three corpora with more than one simplification per original sentence (see Evaluation Challenge F), i.e., APA-RST (see Subsection 4.3.5), Online Participation (see Subsection 4.5.3), and ABGB (see Subsection 4.5.2). However, they all contain exactly two simplifications, which is comparably low to up to 10 references in ASSET (see Section 4.2). In other corpora with more than one simplified version (e.g., APA-LHA (see Subsection 4.3.4) or lexica-corpus (see Table 4.5)), no sentence pairs are available in which the simplified versions are aligned to the same complex sentence, and hence, they cannot be used to evaluate against several references.

As previously argued, automatic alignment is error-prone and hence should (if necessary) only be used for training a TS model but not for evaluating it. Unfortunately, for some corpora, only automatic alignments are (publicly) available (e.g., BiSECT (see Subsection 4.1.5, APA-LHA (see Subsection 4.3.4, or translated ASSET (see Subsection 4.2.4)). If these data are used to evaluate a TS model, for example, in case it is the only available for the target group of interest, an additional manual evaluation is highly encouraged to check whether the automatic scores are due to bad alignment of the gold data or bad simplification of the TS model.

4.9.2 Outlook

In this section, I have introduced several German text simplification corpora and shown how they are related to the previously introduced approaches to building TS corpora (see Chapter 3).

In the next section, I will introduce the next part of the text simplification process, i.e., evaluation (see Chapter 5). I will introduce methods on how to evaluate text simplification models, before I combine this knowledge with knowledge regarding the (training and test) corpora to discuss the capability of German text simplification models (see Chapter 6).

Chapter 5

Text Simplification Evaluation

After finishing the data creation or selection part (see step A to F in text simplification workflow; see Figure 2.1 in Subsection 2.2.2), a text simplification model can be trained (see step G) and evaluated on this data (see step H). After the first evaluation in the development phase against automatic metrics, the parameters of the models will be tuned to achieve better performance in the given task, here text simplification. To better understand TS models' performance, I am explaining in this section how and why TS models are evaluated.

It is essential to evaluate the automatically simplified texts to ensure that TS systems are effective in terms of simplicity and meaning preservation (see step H). A good simplification should be grammatically correct, simpler, and more readable than the original text, but preserve the original meaning of it by only omitting the least important information. Evaluation during the development phase of a text simplification model is usually conducted automatically, but during the final phase manual and/or automatic evaluation are consulted to verify the models' capabilities. In the development and the final phase, a model is evaluated against the development and test set of a corpus (see Chapter 4).

On the one hand, computational methods allow for a fast and efficient automatic evaluation of large amounts of text. Therefore, metrics have been designed that are used to automatically measure the quality of simplified texts wrt. simplicity, grammaticality, and meaning preservation compared to the original texts and gold simplifications (Alva-Manchego et al., 2020b). On the other hand, in manual evaluation, people are asked to carefully read and rate the extent of these three aspects for the generated simplification with respect to the original sentence. However, manual evaluation is very time-consuming; hence, for a first quality check, automatic metrics are nearly always used. But, manual evaluation is more reliable than automatic evaluation because humans can understand and verify the accuracy of the meaning in a way that machines may miss. Additionally, humans can ensure that the text remains clear and easy to comprehend.

In the following, I will provide more details on manual (see Section 5.1) and automatic evaluation (see Section 5.2), including their benefits and challenges. Unless otherwise stated, I will describe the evaluation procedure of sentence simplification as there is sparse research regarding the evaluation of document simplification. I also include much information regarding

the evaluation of English TS as non-English TS evaluation studies reuse parts of the English evaluation process.

I then summarize how German TS studies have been evaluated so far, again including manual and automatic evaluation methods (see Section 5.3) before I finally conclude the section (see Section 5.4).

5.1 Manual Evaluation

Manual evaluation is the most reliable method to determine the quality of system generated simplifications (Alva-Manchego et al., 2020b). As summarized in Subsubsection 2.2.2.3, the usual process of manual evaluation is to ask a few people to rate the system output regarding grammaticality (also called fluency), meaning preservation (also called adequacy) and simplicity. However, detailed best practices on how to manually evaluate text simplification are lacking.

In the following, I will discuss current approaches to manual text simplification evaluation. First, I introduce two concepts of evaluation, i.e., intrinsic and extrinsic evaluation (see Subsection 5.1.1). I then focus on the intrinsic evaluation and explain the evaluation aspects against which the text simplification outputs are evaluated (see Subsection 5.1.2). After that, I describe how to evaluate the overall quality of the simplification as one score when combining all aspects of the intrinsic evaluation (see Subsection 5.1.3). Finally, I present available (English) datasets with manual judgments (see Subsection 5.1.4).

5.1.1 Intrinsic vs. Extrinsic Evaluation

Manual evaluation of a natural language generation task, e.g., text simplification or question answering, can be measured with intrinsic or extrinsic evaluation methods (Belz and Reiter, 2006; van der Lee et al., 2019; Stajner, 2021). van der Lee et al. (2019) define both methods as follows:

> "Intrinsic approaches aim to evaluate properties of the system's output, for instance, by asking participants about the fluency of the system's output in a questionnaire. Extrinsic approaches aim to evaluate the impact of the system, by investigating to what degree the system achieves the overarching task for which it was developed." (van der Lee et al., 2019, p.356)

On the one hand, in recent years, intrinsic evaluation has been conducted much more often than extrinsic evaluation in general in natural language generation (van der Lee et al., 2019), and also in automatic text simplification (Stajner, 2021; Säuberli et al., 2024). As mentioned previously, generated simplifications are most often intrinsically evaluated on a questionnaire on the evaluation aspects of fluency, meaning preservation, and simplicity. In the next section, I will describe these evaluation aspects in more detail.

On the other hand, extrinsic evaluation in the scope of text simplification contains analysis with the target group by comparing the reading speed or the comprehension of a text on complex and simplified texts. The assumption is that for well-simplified texts, the reading

time would decrease but the comprehension of the text would increase. This kind of analysis can be performed by asking comprehension questions, measuring reading time, or recording eye movements (Stajner, 2021). For measuring the comprehension of a text, some content-related questions are asked and afterwards the accuracy of responses is measured (Angrosh et al., 2014); if the target group can answer more questions correctly after reading the simplified texts, these texts appear to be good simplifications and to be helpful to better understand a text (Alva-Manchego et al., 2020b). However, only in a few TS studies has extrinsic evaluation been performed: For example, Saggion et al. (2015) and Angrosh et al. (2014) conducted comprehension-based evaluation with the target group of the simplification, Alonzo et al. (2021) measured the reading speed of people of the target group on the simplified texts, or Rello et al. (2013) analysed eye movements of people of the target group.

Extrinsic evaluation has the advantage over intrinsic evaluation that it is able to directly evaluate the task for which it was designed (Amidei et al., 2018) and is more goal-oriented (Alva-Manchego et al., 2020b), i.e., higher comprehensibility of the text by the target group. In contrast, in intrinsic evaluation conclusions are drawn from questions that are not standardized and answered by the participants who are often not part of the addresses of the simplified texts. Further, intrinsic evaluation contains self-indicated complexity assessment, which is often not as reliable as actually measuring the comprehension of text with comprehension questions as part of extrinsic evaluation (Säuberli et al., 2024). Nevertheless, it is more costly to create and execute extrinsic evaluation than intrinsic evaluation because finding participants of the target group is more difficult than finding general participants. Additionally, the questions of intrinsic evaluation can be reused across different evaluation studies, whereas comprehension questions must be designed dependently on each simple-complex text pair.

5.1.2 Evaluation Aspects (Intrinsic)

Following the overall aim of text simplification to generate a grammatically correct simplification of a complex text that retains the original meaning, the most used criteria for the evaluation of generations are: grammaticality (also called fluency), meaning preservation (also called adequacy), and simplicity (Alva-Manchego et al., 2020b). An overview of the description of these three aspects per study is presented in Table 5.1.[1] Although the three aspects are used in several human evaluation studies, there is no agreement on scale sizes, aspect naming, or scale descriptions.

The most common rating scales are Likert-scales, which have been introduced by Likert (1932) to measure attitudes of people. These scales often include several items, which are all verbalized by a question or statement and accompanied with answer options on a scale ranging from disagreement to agreement. However, the length of the scale is not fixed and can be adapted to the needs of the researcher. In text simplification, mostly 5 point Likert-scales are chosen from 1 to 5 (e.g., see Maddela et al. 2021), a scale from -2 to $+2$ (e.g., see Sulem et al.

1 The table does not aim at completeness, but highlights the variety of different usages of the evaluation aspects. I show the shortened version of the scale labels of Yamaguchi et al. (2023) in the table. The long description of each scale point would exceed the available space in this table. For full scale descriptions, see e.g., Table 6 in Yamaguchi et al. (2023).

	Grammaticality		Meaning Preservation		Simplicity	
	Scale	**Description**	**Scale**	**Description**	**Scale**	**Description**
Siddharthan (2006)	yes/no	grammaticality	0 to 3	meaning preservation		
Woodsend and Lapata (2011)	1 to 5	Is the target sentence grammatical	1 to 5	Does the target preserve the meaning of the source?	1 to 5	Is the target sentence simpler than the source?
Wubben et al. (2012)	1 to 5	The extent to which a sentence is proper, grammatical English	1 to 5	The extent to which the sentence has the same meaning as the source sentence	1 to 5	The extent to which the sentence was simpler than the original and thus easier to understand.
Narayan and Gardent (2014)	0 to 5	Is the simplified output fluent and grammatical?	0 to 5	Does the simplified sentence(s) preserve the meaning of the input?	0 to 5	Does the generated sentence(s) simplify the complex input?
Štajner et al. (2014)	1) ungrammatical, 2) minor problems with grammaticality, 3) grammatical		1) meaning is seriously changed, or [...], 2) some of the relevant information is lost but [...], 3) all relevant information is kept [...]			
Štajner et al. (2016a)	good, okay, bad		good, okay, bad		good, okay, bad	
Xu et al. (2016)	0 (worst) to 4 (best)		0 (worst) to 4 (best)			
Nisioi et al. (2017)	1 to 5 (very bad to very good)		1 to 5 (very bad to very good)		+2 – much simpler; +1 – some-what simpler; 0 – equally difficult; −1 – some-what more difficult; −2 – much more difficult	
Kriz et al. (2019)	1 to 5	Is the simplified sentence written in well-formed English?	1 to 5	Preserves the simple sentence the meaning of the original sentence?	1 to 5	Is the sentence simpler than the complex sentence?
Alva-Manchego et al. (2020a)	1 (strongly disagree) to 100 (strongly agree)	The Simplified sentence is fluent, there are no grammatical errors.	1 (strongly disagree) to 100 (strongly agree)	The simplified sentence adequately express the meaning of the Original, perhaps omitting the least important information.	1 (strongly disagree) to 100 (strongly agree)	The Simplified sentence is easier to understand than the Original Sentence.
Cooper and Shardlow (2020)	1 (poor grammar) to 10 (perfect grammar)		1 (poor meaning preservation) to 10 (very good meaning preservation)			
Dong et al. (2019)	1 to 5	Is the output grammatical?	1 to 5	How much meaning from the original sentence is preserved?	1 to 5	Is the output simpler than the original sentence?
Maddela et al. (2021)	1 (strongly disagree) to 5 (strongly agree)	The simplified sentence is fluent.	1 (strongly disagree) to 5 (strongly agree)	The simplified sentence adequately express the meaning of the original sentence.		
Yamaguchi et al. (2023)	1) completely unintelligble, 2) partially understandable, 3) understandable, 4) non-native speaker level fluent, 5) native speaker level fluent	grammaticality / fluency	1) unintelligible, 2) completely different, 3) partially preserved, 4) mostly preserved, 5) adequately preserved	meaning preservation / adequacy	1) unintelligible, 2) more difficult, 3) the same simplicity, 4) simpler, 5) much simpler	simplicity

Table 5.1: Summary of scales, their descriptions and names per rating aspect in TS human evaluation studies.

2018c), or a continuous scale from 0 to 100, (e.g., see Scialom et al. 2021; Alva-Manchego et al. 2020a).

On the one hand, Alva-Manchego et al. (2020a) argue that a continuous scale leads to more consistency in the agreement between annotators in the evaluation of text simplification, as already proved for machine translation. On the other hand, Sulem et al. (2018c) prefer a Likert-scale with negative to positive scale points, including a neutral middle point. Following Chyung et al. (2017); Nadler et al. (2015), annotators interpret the middle point as, e.g., "undecided", "neutral", or "no opinion". Although both scales, i.e., -2 to $+2$ and 1 to 100, have a middle point, only in the first setting it is clearly intended as a neutral value. Depending on the scale understanding by the participants, in the continuous scale, 50 could also be used as a neutral value, which the study developers might not have intended to use when designing the study.

Following Cohen et al. (2007) a Likert-scale can be verbalized with a question or a statement. However, the verbalization should be written in the form of a multiple choice question and not a dichotomous question. In text simplification evaluation, Likert-scales are often introduced using a question such as "how much does the simplified sentence ..." or "to which extent ..." (e.g. see Sulem et al. 2018b; Zhang and Lapata 2017). Typical answer options or items of a Likert-scale are "strongly disagree", "disagree", "neither disagree nor agree", "agree", or "strongly agree" (e.g., see Cohen et al. 2007, Alva-Manchego et al. 2020a, or Maddela et al. 2021). However, in some studies the items are detailed comments regarding the statement (e.g., see Wubben et al. 2012). Further, in some paper, a closed question is asked (which expects a binary answer), but the answer options are on a scale from 1 to 5 (e.g., see Dong et al. 2019; Kumar et al. 2020).

Furthermore, Alva-Manchego et al. (2020b) have raised the questions of whether the typical three evaluation aspects of grammaticality, meaning preservation, and simplicity are enough to evaluate the quality of text simplification. Especially with a view on document simplification, other aspects such as coherence of the text might be interesting during evaluation. Cumbicus-Pineda et al. (2021) suggest a detailed checklist to evaluate text simplification outputs including sub-aspects of the three main aspects and ethical aspects as well. The works of Devaraj et al. (2022) and Yamaguchi et al. (2023) go in a similar direction; they propose annotation schemata for manual error annotation in system generations. Currently, these schemata have just been applied to English TS, and it remains an open question whether the same or other categories would be relevant to evaluate German TS.

I also want to pay attention to the usage of singular and plural in the scale descriptions. In some descriptions (or scale questions), it is intended that a simplification or the original contain only one sentence, which is not true for sentences which have been split or merged. However, only a few studies include the plural in their description, e.g., Narayan and Gardent (2014) or Narayan and Gardent (2016).

In the following, I provide more details on these three main aspects, i.e., meaning preservation (see Subsubsection 5.1.2.1), grammaticality (see Subsubsection 5.1.2.2), simplicity (see Subsubsection 5.1.2.3), and also on additional aspects which are used to intrinsically evaluate TS system outputs, i.e., structural & lexical simplicity (see Subsubsection 5.1.2.4), and coherence (see Subsubsection 5.1.2.5).

5.1.2.1 Meaning Preservation

Meaning preservation aims, as the name suggests, to assess how much of the original meaning of the complex text is retained in the generated simplified text. Most human evaluation studies do not specify whether all or most of the content should be preserved (see Table 5.1). In contrast Alva-Manchego et al. (2020a) specifies that the least important information is allowed to be omitted to still achieve the highest meaning preservation score, whereas Štajner et al. (2014) and Yamaguchi et al. (2023) penalize if the meaning is only partially preserved by explicitly addressing it in their scale point descriptions.

Following Sulem et al. (2018b), the aspect of meaning preservation can be further split into the sub-aspects of information deletion and information addition (possible answers in questionnaire: no,maybe, and yes). In a simplification, new information could be gained, for example, when implicit information is correctly made more explicit, when complex terms are correctly explained, or if wrong information is added. Zhang et al. (2020b) also includes the sub-aspect of information gain by asking if the text introduces new facts (dichotomous; answers: yes/no).

However, in line with the manifestations of meaning preservation, in the manifestation of both versions of information gain, it is not specified whether additional information is expected or whether it was added correctly or wrongly. Stajner (2021) urges that people of the target groups of text simplification rely on the simplified texts and can be harmed by incorrectly added or omitted information. Therefore, she claims that meaning preservation should be checked manually before making simplified texts publicly available. To manually evaluate the correctness of information insertion, deletion, or additionally its substitution, Devaraj et al. (2022) propose an error annotation schema regarding the extent of the wrong change, e.g., no or trivial change, non-trivial change which preserves the main idea, and change which does not preserve the main idea. Yamaguchi et al. (2023) propose a similar idea and add error categories in their annotation schema regarding inappropriate deletion, inappropriate addition, and inappropriate paraphrase.

Overall, to rate how much meaning is correctly preserved or wrongly changed, the annotators need to understand both texts; hence, native speaker could be asked, but people from the target group are not ideal judges for this task.

5.1.2.2 Grammaticality

As syntactical changes are expected outcomes of (structural) simplification, measuring the grammaticality of the generated simplified text is important. Stajner (2021) argues that if simplified texts contain grammatical errors, wrong sentence structures, or incorrect word forms, the target group (e.g., children and language learners) would learn these errors. Therefore, in nearly each TS study with human evaluation, the outputs of TS systems are checked if they contain grammatical errors, e.g., including non-grammatical sentence structures or wrong inflection of words. If the system generations contain too many errors, they could be manually corrected during post-editing (Stajner, 2021).

A term or criterion that is often combined with grammaticality is fluency (e.g., see Narayan and Gardent 2014 or Alva-Manchego et al. 2020a). However, even if most studies refer to the same concept for the evaluation of grammaticality and fluency, the questions or statements for

the rating scales are phrased differently. It ranges from dichotomous questions (e.g., "Is the output grammatical?" Dong et al. 2019), via simple statements with scale points of strongly agree and disagree (e.g., Maddela et al. 2021), to detailed scale labels (e.g., see Štajner et al. 2014 or Katsuta and Yamamoto 2018).

People with high language skills in the language of interest should judge grammatical correctness (Alva-Manchego et al., 2020b), as they are aware of the grammar rules of the language.

5.1.2.3 Simplicity

The evaluation of the simplicity or simpleness of a generated text is the most obvious criterion for the evaluation of text simplification. However, simplicity and simple have different meanings in the scope of the evaluation of text simplification. Simpleness measures the readability level or also if a specific readability level has been reached with the simplification (e.g., see Martin et al. 2018), in more detail, if a text is simple enough to be understandable by the target group of simplification. The simpleness of a text can be measured by determining whether it has reached a specific readability level through simplification. In other words, the simpleness of a text should match the readability skills of the target group for whom it has been simplified. For example, very high simpleness for the target group of German Easy Language or high simpleness for the target group of German Plain Language. In contrast, simplicity measures the extent of the simplification no matter the start and end point, it just measures the distance (e.g., see Nisioi et al. 2017 or Yamaguchi et al. 2023). For example, if a source text would be written in standard language and the target text should be written in easy language, the simplicity would be much higher for this pair than for a plain-language easy-language pair.

Simpleness is more suitable to measure if the needs of participants are solved whereas simplicity is more suitable to measure the capabilities of a text simplification system. The latter often neglects the target group, i.e., not focusing on whether a sentence is simple enough for the target group or if more simplification is required. Often the simplicity is measured between the target texts and the generated texts to measure the distance between the simpleness of the generated text and the gold simplification.

In text simplification evaluation, simplicity is more often considered than simpleness. However, there is no agreement on how to measure it, options are, e.g., to quantify i) the simpleness of the generated text without a reference to the original text (e.g., see Martin et al. 2018), ii) the increase of simpleness in comparison to the original text (e.g., see Wubben et al. 2012 or Alva--Manchego et al. (2020a)), iii) or the extent of the change wrt. simpleness including also more complex variants (e.g., see Nisioi et al. 2017 or Yamaguchi et al. 2023).

In some studies, the simplicity is evaluated on a dichotomous scale (whether the text is more simple or not, e.g., see Surya et al. 2019 or Kriz et al. 2019) or on a continuous scale referring to the number of successful lexical or syntactical simplified phrases (also called simplicity gain, e.g., see Xu et al. 2016). Koptient and Grabar (2020) propose a pairwise comparison by asking which sentence is simpler, the original, or the simplified text.

Overall, the approach most often used for measuring simplicity is to rate it on a Likert-scale giving the following phrase : "The simplified sentence is easier to understand than the original sentence." (e.g., see Wubben et al. 2012, Alva-Manchego et al. 2017, Vu et al. 2018,

Alva-Manchego et al. 2020a, Scialom et al. 2021, or Lin and Wan 2021). However, this version does not specify if the goal is to measure only the extent of the simplification (where 0 means of same or higher complexity; as used in e.g., Alva-Manchego et al. 2020a) or if the goal is to measure the whole complexity scale starting from more complex and ending in more simple (as used in e.g., Nisioi et al. 2017).

Furthermore, following existing human evaluation studies, there is no common thread regarding how many people and people with what skills to ask to rate the simplicity of the generated texts. Usually, the participants or evaluators are not part of the target group of the target texts. In contrast, they are often, for example, native speakers, students with a high literacy level, crowd workers, or a non-specified group of annotators (Stajner, 2021). Compared to the previous evaluation aspects, for the aspect of simplicity it is important that the participants are addressees or part of the target group of the simplified texts, because complexity estimation is very subjective (e.g., see Siddharthan 2014, Štajner 2018, or Gooding and Tragut 2022) and people with high literacy levels often do not recognize much difference in the complexity of the source text and simplified texts (Štajner and Nisioi, 2018). Hence, people with high literacy levels are often not qualified to estimate the simplicity of a text on behalf of the people of the target group.

5.1.2.4 Structural & Lexical Simplicity

Depending on the included or intended simplification operations in a TS system, more fine-grained evaluation of simplicity regarding structural or lexical simplicity can be included in the evaluation study. Sulem et al. (2018b) were the first to add the aspect of "structural simplification" ("Is the output simpler than the input, ignoring the complexity of the words?") to the human evaluation procedure. They first evaluated it with the answer options "no", "yes", and "maybe", but they extended it in 2018c to a 5-point Likert-scale ranging from -2 to $+2$ (Sulem et al., 2018c) including the option of higher complexity (-2) and the same complexity (0).

Another option of evaluating structural simplicity has been proposed by Zhang et al. (2020b), they ask the annotators whether a given text is split wrongly and whether it should be further split into smaller parts (dichotomous; answers: "yes" and "no"). Maddela et al. (2021) have slightly modified this evaluation item by asking on a 5-point Likert-scale if "the simplified sentence undergoes correct sentence splitting".

To the best of my knowledge, the aspect of lexical simplification has not been explicitly included in any manual evaluation study prior to this work.

5.1.2.5 Coherence

Another aspect of how to evaluate system-generated simplifications is coherence (Siddharthan, 2006; Shardlow, 2014; Vásquez-Rodríguez et al., 2023) which can be especially relevant for the evaluation of sentences with structural simplification (Siddharthan, 2006) or simplification at the paragraph or document level (Vásquez-Rodríguez et al., 2023). Following Siddharthan (2006), coherence in light of discourse simplification can concern conjunctive (usage of conjunctions to connect sentences) and anaphoric cohesion (usage of referential expressions, e.g., pronouns). If sentences are well connected and, hence, coherent, they are easier to understand

wrt. faster reading time and higher recall (Kintsch et al., 1975; Myers et al., 1987). In addition, the resolution of the anaphora can make a text less ambiguous.

Following Vásquez-Rodríguez et al. (2023), manual evaluation of coherence is challenging, which might be the reason why it is currently rarely included in TS human evaluation (e.g., see Siddharthan 2006).

5.1.3 Overall Quality of Simplicity

The previous aspects of TS evaluation can be combined into one huge evaluation aim or criterion, further called "overall simplicity quality". It is important to note that the terms "simplicity" and "overall simplicity quality" are related but not synonymous: In contrast to "simplicity", "overall simplicity quality" aims at evaluating fluent, adequate, and more simple system outputs all at once.

To the best of my knowledge, Maddela et al. (2023) are the first who have collected human judgments regarding the overall simplicity quality of complex-simple pairs. Similarly, as for the other evaluation aspects, participants are asked to give a continuous rating between 0 (worst) and 100 (best), but here following an extensive definition of simplification[2]. Further, they provide for each quartile a scale point description to facilitate the rating as follows (according to Maddela et al. (2023)):

- "0 – The sentence is completely unreadable."
- "25 – The sentence is equivalently simple, still has some fluency but the meaning is lost."
- "50 – The sentence is simpler, somewhat fluent and the meaning is similar to the original sentence."
- "75 – The sentence is somewhat simpler, mostly fluent and the meaning is close the original sentence."
- "100 – Only when the sentence is fully simplified, entirely fluent and preserves the core meaning of the original sentence."

The extent of the aspects regarding simplicity, grammaticality and meaning preservation are all simultaneously increasing for each quartile. However, in this combined scale it is unclear how to rate a sentence which accomplishes only two of the three aspects, e.g., a more simple and fluent version which greatly changes the meaning of the original.

In addition, Maddela et al. (2023) extend the usual "simplicity" scales by asking for complete simplification, i.e., the highest possible version of simplification, no more simplification operations can be applied. In usual simplicity scales, only the extent is measured without addressing the maximum possible simplification. However, it is unclear whether a "fully simplified sentence" is assessed with respect to the target level of the simplification or the simplification purpose.

Furthermore, they explicitly mention how to address the same simplicity (and, therefore, implicitly lower simplicity) which has not been expressed in the aspect regarding only "simplicity". To evaluate the quality of this newly proposed scale, more investigation is required regarding the scale interpretation by the participants.

2 Unfortunately, this definition is not publicly available.

5.1.4 Datasets with Human Judgments

However, as evaluation sets often contain several hundred or thousands of items, only random samples of the sentence pairs are picked and manually evaluated (e.g., 50 to 100 pairs). The results of the human evaluation studies can be stored together with the complex-simple pairs in "datasets with human judgements". Unfortunately, even if the evaluation procedure of some studies has been well described, the human judgments, which have been collected in these studies, are only available for a few studies: e.g., QATS (Štajner et al., 2016b)[3], HSplit (Sulem et al., 2018c)[4], PWKP test (Sulem et al., 2018b)[5], ASSET (Alva-Manchego et al., 2020a)[7], Web-Split, Wiki-BM, & Cont-BM (Zhang et al., 2020b)[8], human-likert & system-likert (Scialom et al., 2021)[9], Fusion (Schwarzer, 2018; Schwarzer et al., 2021)[10], Simplicity-DA (also called Wiki-DA) (Alva-Manchego et al., 2021)[11], Newsela-Likert (Maddela et al., 2021)[12], and SimpEval (Maddela et al., 2023)[13]. Comparing the evaluation dimensions, scale size, and number of raters per dataset (see Table 5.2), there are many differences in the realizations of the human evaluation.

Overall, most of these datasets are relatively small, which hampers the construction of trainable metrics on these scores. Beauchemin et al. (2023) merged a few of the datasets with very similar evaluation setups (that is, ASSET, Simplicity-DA, SimpDA Maddela et al. 2023, System-Likert, and Human-Likert) to get a larger data set called "Continuous Scale Meaning Dataset (CSMD)"[14]. However, it remains an open question whether these datasets can be smoothly merged or whether the differences regarding the evaluation setup (e.g., training of annotators or different quality checks of annotators) are too huge to build a homogeneous dataset.

3 https://qats2016.github.io/shared.html [last access: July 24, 2024]

4 https://github.com/eliorsulem/simplification-acl2018 [last update: September 6, 2018; last access: July 24, 2024]

5 Due to a currently dead link to the system outputs of the sentence pairs, you can instead copy the system outputs provided in EASSE (Alva-Manchego et al., 2019a) in the given order. However, the sentence pairs of the 2 system outputs could not be found. Hence, this version of the dataset contains only 500 sentence pairs. The human judgments are available online.[6]. The original sentences and system outputs are available in EASSE https://github.com/feralvam/easse/tree/master/easse/resources/data [last update: October 13, 2021; last access: July 24, 2024].

7 https://github.com/facebookresearch/asset/tree/master/human_ratings [last update: September 16, 2022; last access: July 24, 2024]

8 https://developer.ibm.com/exchanges/data/all/split-and-rephrase/ [last update: March 24, 2021; last access: July 24, 2024]

9 http://dl.fbaipublicfiles.com/questeval/simplification_human_evaluations.tar.gz [last access: July 24, 2024].

10 The data will be available here https://cs.pomona.edu/~dkauchak/simplification/ [last access: July 24, 2024]. Currently it is only available on request of the authors.

11 https://github.com/feralvam/metaeval-simplification [last update: July 29, 2022; last access: July 24, 2024]

12 Newsela-Likert is not available online, but Maddela et al. (2023) edited it and made the updated version available as SimpLikert: https://github.com/Yao-Dou/LENS/tree/master/data [last update: July 11, 2023; last access: July 24, 2024].

13 https://github.com/Yao-Dou/LENS/tree/master/data [last update: July 11, 2023; last access: July 24, 2024]

14 The CSMD is available online: https://github.com/GRAAL-Research/csmd [last update: March 23, 2024; last access: July 24, 2024].

Dataset	Meaning Preservation		Grammaticality		Simplicity		Source	# Raters
	Definition	Scale	Definition	Scale	Definition	Scale		
QATS	-	1 (bad), 2 (ok), 3 (good)	-	1 (bad), 2 (ok), 3 (good)	-	1 (bad), 2 (ok), 3 (good)	EventS, EncBrit, LSLight	-
HSplit	Does the output preserve the meaning of the input?	1 to 5	Is the output fluent and grammatical?	1 to 5	"Is the output simpler than the input?"	-2 to $+2$	TurkCorpus	3 experts
PWKP test	Does the output add information, compared to the input? Does the output remove important information, compared to the input?	1 (no), 2 (maybe), 3 (yes)	Is the output grammatical?	1 (no), 2 (maybe), 3 (yes)		1 (no), 2 (maybe), 3 (yes)	PWKP	5 experts
ASSET	The simplified sentence adequately express the meaning of the original, perhaps omitting the least important information.	0 ("strongly disagree") to 100 ("strongly agree")	The Simplified sentence is fluent, there are no grammatical errors.	0 ("strongly disagree") to 100 ("strongly agree")	The simplified sentence is easier to understand than the original sentence.	0 ("strongly disagree") to 100 ("strongly agree")	TurkCorpus	15 CW
Human-Likert & System-Likert	The simplified sentence adequately express the meaning of the original, perhaps omitting the least important information.	0 ("strongly disagree") to 100 ("strongly agree")	The Simplified sentence is fluent, there are no grammatical errors.	0 ("strongly disagree") to 100 ("strongly agree")	The simplified sentence is easier to understand than the original sentence.	0 ("strongly disagree") to 100 ("strongly agree")	TurkCorpus	$12-35$ CW
Fusion	Sentence 2 preserves the meaning of sentence 1	1 to 5			How much simpler is sentence 2 than sentence 1	-2 (much less simple) to $+2$ (much simpler)	Newsela	3 CW
Simplicity-DA	The simplified sentence adequately express the meaning of the original, perhaps omitting the least important information.	0 ("strongly disagree") to 100 ("strongly agree")	The Simplified sentence is fluent, there are no grammatical errors.	0 ("strongly disagree") to 100 ("strongly agree")	The simplified sentence is easier to understand than the original sentence.	0 ("strongly disagree") to 100 ("strongly agree")	ASSET & TurkCorpus	15 CW
Newsela-Likert	The simplified sentence adequately expresses the meaning of the original sentence.	1 ("strongly disagree") to 5 ("strongly agree")	The simplified sentence is fluent.	1 ("strongly disagree") to 5 ("strongly agree")	The simplified sentence is easier to understand than the original sentence.	1 ("strongly disagree") to 5 ("strongly agree")	Newsela-Auto	5 CW
SimpEval-Past & SimpEval-2022	-	-	-	-	0 (the sentence is completely unreadable) to 100 (Only when the sentence is fully simplified, entirely fluent and preserves the core meaning of the original sentence.		TurkCorpus & ASSET &new Wikipedia articles	$3-5$ experts

Table 5.2: Evaluation dimensions, scales, raters (CW = crowd workers), and number of sources per dataset. Extended version of Table 1 in Stodden (2021c).

5.2 Automatic Evaluation

With the increasing mass of evaluation data from different model approaches, it becomes challenging to manually evaluate this large number of generated texts. Automatic evaluation methods have been employed to improve and facilitate the assessment process by providing quick measures of the performance of text simplification systems (Alva-Manchego et al., 2020b). Compared to manual evaluation methods, automatic evaluation methods facilitate a quick assessment of the output of various text simplification models, making it feasible to compare and iterate on different approaches efficiently. In addition, automatic evaluation methods allow researchers to scale up their assessments to handle large datasets effectively (Alva-Manchego et al., 2020b).

One common approach to automatic text simplification involves comparing the generated simplified text not only against the original text, but also against one or more reference texts, which are manually (sometimes also verified) simplified versions of the original text (Alva-Manchego et al., 2020b). Most metrics that are used to automatically measure the quality of simplified texts mirror the manual intrinsic evaluation[15]: The metrics are built to correspond and correlate to one or more of the manual evaluation aspects, e.g. meaning preservation (see Subsubsection 5.1.2.1), grammaticality (see Subsubsection 5.1.2.2) or simplicity (see Subsubsection 5.1.2.3). As these aspects are also relevant for related tasks of TS, such as machine translation or text summarization, for automatic evaluation of TS, often metrics of these related tasks are utilized, e.g., BLEU (Papineni et al., 2002) (see Subsubsection 5.2.1.1), BERTScore (Zhang* et al., 2020) (see Subsubsection 5.2.1.1, or FKGL (Flesch, 1948) (see Subsubsection 5.2.1.3).

Only a few metrics exist which are invented for the simplification purpose, i.e., SARI (Xu et al., 2016) and SAMSA (Sulem et al., 2018b)(see Subsubsection 5.2.1.4). But, currently a group of metrics is used to evaluate TS systems, i.e., SARI, BERTScore, BLEU, and FRE, because they have all different advantages and disadvantages: they are either not entirely suitable, focus only on one part of text simplification, or do not sufficiently correlate with the human ratings (Alva-Manchego et al., 2021).

In the following, I will present the metrics used for text simplification sorted by the manual evaluation aspects, which they should represent or correlate to (see Subsection 5.2.1). I then introduce a metric (called LENS) which tackles all aspects of simplification evaluation at the same time (i.e., "overall quality of simplicity"; see Subsection 5.2.2), and metrics which do not require references (see Subsection 5.2.3). In addition, I will present an evaluation framework in which several TS metrics are implemented for an easy comparison of TS systems (see Subsection 5.2.4).

To exemplify the usage of the most relevant metrics (i.e., BLEU, SARI and BERTScore), I am providing the following example, including an original sentence, one gold or reference simplification, and three examples for possible system-generated simplifications:

(1) *Original Sentence:*
Veröffentlichungspflichtig sind alle Informationen, die bei den Behörden vorliegen.

15 To the best of my knowledge, no automatic metric focuses on replicating extrinsic evaluation approaches.

'Required to be published are all information, which is available with the authorities.'

(2) *Reference Simplification in German Plain Language:*
Alle vorhandenen Informationen der Behörden müssen veröffentlicht werden.
'All existing information of the authorities must be published.'

(3) *System Simplification 1):*
Alle vorliegenden Informationen der Behörden sind veröffentlichungspflichtig.
'All available information of the authorities are required to be published.'

(4) *System Simplification 2):*
Alle vorhandenen Informationen der Behörden sind veröffentlichungspflichtig.
'All existing information of the authorities is required to be published.'

(5) *System Simplification 3):*
Alle vorliegenden Informationen der Behörden müssen veröffentlicht werden.
'All available information of the authorities must be published.'

(6) *System Simplification 4):*
Behörden müssen alle ihre vorhandenen Informationen veröffentlichen.
'Authorities must publish all their available information.'

In system simplification 1 and 2, the long adjective "versicherungspflichtig" is copied from the original sentence and shows deviations (see red lines) from the reference simplification where this word is split. In system simplification 3, the word is correctly split (see green lines). In system simplification 4, the adjective is also split. Additionally, the grammatical voice has correctly changed from passive to active following simplification guidelines. However, this change has not been applied to the reference simplification, meaning that the output of system simplification 4 does not exactly overlap with the reference simplification (see yellow lines).

5.2.1 Evaluation Aspects

5.2.1.1 Meaning Preservation

I have already previously discussed in Subsubsection 5.1.2.1 that meaning preservation is a crucial evaluation aspect of simplified texts. As it is also a relevant aspect in related tasks, such as machine translation and text summarization, the metrics of these tasks are utilized for text simplification evaluation, e.g., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or BERTScore (Zhang* et al., 2020). Recently Fruth et al. (2024), have proposed a new approach on automatically measuring meaning preservation based on similarity measurement and sentence-wise alignment via BERTScore. Due to its novelty, more time and research are required to verify the approach. Furthermore, to the best of my knowledge, no automatic metric has been applied to text simplification that considers information gain.

BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002) is a string-based similarity metric that measures the overlap of n-grams (1 to 4-grams) between the system output and at least another text. Hence, BLEU is a language-independent metric. In early text simplification

studies, e.g., Specia (2010); Zhu et al. (2010); Woodsend and Lapata (2011), BLEU has been utilized to evaluate sentence simplification outputs and, despite many criticisms (e.g., Wubben et al. 2012, Sulem et al. 2018a or Zhang et al. 2020b), it is still a common evaluation metric in the latest studies, e.g., Ryan et al. (2023); Wu et al. (2023); Heineman et al. (2023).

The original version of BLEU has been designed for bilingual machine translation comparing the n-grams of the system output with those of the target text, as only those are in the same language. Instead, in text simplification, the n-grams of the system output could be compared with the reference simplifications or the original text as all are written in the same language but different varieties of it. However, the usual BLEU approach in TS research is to evaluate the system generation against several (but at least one) reference simplifications.

Looking at a concrete scoring example of BLEU with item 1 to item 6[16], the systems would be ranked as follows: first system 3 (see item 5; BLEU: 75.06), second system 2 (see item 4; BLEU: 47.75), third system 1 (see item 3; BLEU: 20.61), and as last system 4 (see item 3; BLEU: 14.57). Because the first system output has a higher n-gram overlap than the other with the reference simplification (see item 2), it is ranked first. Even if system 4 has a high lemma overlap with the gold reference, the BLEU is low due to the different word order and different inflection of the words.

Although many studies show a correlation of BLEU and human judgments regarding meaning preservation and grammaticality (see, e.g., Xu et al. 2016, Alva-Manchego et al. 2020a, or Zhao et al. 2023), in recent years, some criticism regarding BLEU has been proposed: In contrast to previous studies, Wubben et al. (2012) found that BLEU significantly correlates in their experiment with grammaticality and simplicity, but not with meaning preservation. Sulem et al. (2018a) (evaluated on HSplit) and Zhang et al. (2020b) (evaluated on WebSplit, Wiki-BM, and Cont-BM) further argued that BLEU is not suitable to evaluate syntactically simplified sentences. They could find no or only low correlation between BLEU scores and human judgments regarding grammaticality and meaning preservation in complex-simple pairs which include a sentence split.

Although, in many current studies, the criticism against BLEU is briefly discussed, the same studies still include the score with the justification to be comparable with previous work (e.g., see Ma et al. 2022). Only a few studies refer to the issues and refuse to include the BLEU score in their evaluation (e.g., see Kim et al. 2021; Sun et al. 2023).

Extensions of BLEU have been introduced several times in previous years and they have also been utilized (or designed) for text simplification in some evaluation studies, e.g., T-BLEU (Rios et al., 2011) in Štajner et al. (2014), iBLEU (Sun and Zhou, 2012) in Mallinson et al. (2020), BLEURT (Sellam et al., 2020) in Lu et al. (2023), or FK-BLEU (Xu et al., 2016) in Xu et al. (2016).

One prominent example, which is also relevant for further TS evaluation metrics, is iBLEU by Sun and Zhou (2012). iBLEU is a metric originally designed for paraphrase generation evaluation. It includes the n-gram overlap of the system generation and a simplified reference as in the original BLEU, but it additionally considers also the n-gram overlap of the system genera-

16 The BLEU scores reported here are generated with EASSE-DE Stodden 2024a and the following settings: lowercase: false; language: DE; tokenizer: SpaCy.

tion and the original text. It has been used in a few TS studies, e.g., Xu et al. (2016), Mallinson et al. (2020) or Zhao et al. (2023).

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) is a metric commonly used to assess the quality of automatic text summarization. Specifically, it focuses on measuring the overlap of content between an automatically generated summary and one or more reference summaries. Similarly to BLEU, ROUGE operates by comparing the overlap of n-grams between the generated summary and the reference summaries. The three main types of ROUGE scores are ROUGE-1 or ROUGE-2, which consider unigrams or bigrams, and ROUGE-L, which looks at the longest common sub-sequence of words in both texts. Therefore, ROUGE is language independent, as is BLEU.

ROUGE can be useful to some extent in evaluating how well a simplified text preserves the meaning or retains the information from the original text. Beauchemin et al. (2023) (evaluated on CSMD) found a weak but significant correlation between ROUGE-2 and meaning preservation but not for ROUGE-L or ROUGE-1, which supports the previous statement. Currently, ROUGE has been primarily utilized to evaluate document simplification systems, e.g., as in the following studies: Rios et al. (2021) or Trienes et al. (2022).

Criticism against ROUGE tackles the same points as also holds for BLEU (Ganesan, 2018): they rely on exact word matches, which can be sensitive to the choice of vocabulary. Hence, this approach might not capture semantic similarity effectively, as it does not consider synonyms or paraphrasing.

BERTScore (Zhang* et al., 2020) also evaluates the meaning preservation of a text pair – including the system output and at least one simplified reference – with contextualized word embeddings such as BERT embeddings (Devlin et al., 2019). The proposed BERTScore considers semantic similarity and not only exact-string overlaps; hence, it might support word substitution more than BLEU or ROUGE. Furthermore, BERTScore can compare long words at once, while BLEU is limited to comparisons of n-grams (mostly 4-grams) (Zhang* et al., 2020).

According to Zhang* et al. (2020) as cited in Alva-Manchego et al. (2021), in this method, you get three scores: BERTScore-Recall checks how well the words in the reference match with the ones in the system output. BERTScore-Precision is looking at it the other way around: it tells how well the words in the system output match the words of the reference(s). BERTScore-F1 is the combination of BERTScore-Recall and BERTScore-Precision. If there are multiple references, the system output is compared with all of them, and the highest score is returned.

Looking at a concrete scoring example of BERTScore with item 1 to item 6, again system 3 would be ranked better than system 3, 2, and 1.[17] As can be seen in Table 5.3, system 1 and system 2 achieve a very similar F1 score, whereas system 3 nearly achieves the highest possible score of 1. The only difference between system 1 and 2 are the words "vorhanden" vs. "vorliegen", where the first is also in the reference text. However, unlike the BLEU score (see Subsubsection 5.2.1.1), some synonymy of both words is recognized, so the scores for both sys-

17 The BERTScore scores reported here are generated with EASSE-DE Stodden 2024a and the following settings: lowercase: false; language: DE; tokenizer: SpaCy; BERT-model: bert-base-multilingual-cased.

tems are almost identical. Because the same words are the only difference between the output of system 3 and the gold reference, the score for that system is that high.

In comparison, the output of system 4 receives a higher precision but lower recall than system 1 and 2, because there is a higher ratio of word overlap between system 4 and the gold reference (only "ihre" not fully matching) than between system 1 and 2 and the reference (precision; "sind veröffentlichungspflichtig" not fully matching). However, starting the comparison from the gold reference (i.e., recall) system 4 has a lower overlap with the reference than system 2 with the reference because "müssen veröffentlicht werden" is not matching with the reference, but it is semantically closer to "sind veröffentlichungspflichtig" than "der" and "werden" from system 4.

If we would consider the use of capital and small initial letters, system 4 (BERTScore-F1: 0.5977) would be ranked worse than system 2 (BERTScore-F1: 0.6276) because the uppercase version of "alle" is present in the gold reference and the output of system 2, but the output of system 4 contains the lowercase version.

	Precision	**Recall**	**F1**
System 1	0.4132	0.7049	0.5499
System 2	0.4344	0.7452	0.5793
System 3	0.9034	0.9496	0.9264
System 4	0.5588	0.6450	0.6017

Table 5.3: Scoring example of BERTScore of four system outputs (see item 1 to item 6).

In related works, the strengths and weaknesses of BERT-Score have been investigated: Alva-Manchego et al. (2021) (evaluated on Simplicity-DA), Maddela et al. (2023) (evaluated on SimpEval, Simplicity-DA, and Newsela-Likert), and Beauchemin et al. (2023) (evaluated on CSMD) showed that BERTScore-Precision is a suitable metric to evaluate text simplification: they found that the score correlates with human judgments regarding meaning preservation and simplicity for English TS. Maddela et al. (2023) also show that BERTScore weakly correlates with overall simplicity even if analyzing only simplification pairs with paraphrases or only with sentence splits. BERTScore is applicable to English and also other languages for which transformer-based word embeddings exist; the original English embeddings of RoBERTa (Liu et al., 2019) can be replaced with multi-lingual embeddings of BERT (Devlin et al., 2019) or monolingual embeddings in the language of interest to evaluate texts in other languages. As BERTScore has been rather recently introduced and tested for its capabilities regarding TS evaluation, only some TS studies include this metric so far, e.g., Alva-Manchego et al. (2021), Maddela et al. (2023) or Zhao et al. (2023).

However, Maddela et al. (2023) also revealed some criticism against the BERTScore-Precision: The score is also high if a complex sentence is just copied and nothing is simplified, or if generated sentences are error-prone, e.g., contain wrong sentence splits or random word removal. Further, Alva-Manchego et al. (2021) argues that a low BERTScore value indeed corresponds to a low-quality simplification, but if the score is high it becomes less reliable and more information regarding simplification operations and their correctness should also be considered. In contrast to previous findings, Zhao et al. (2023) (evaluated on PWKP) found that BERTScore still correlates with meaning preservation, but not with simplicity (measured

as a combination of simplicity and structural simplicity). Furthermore, they found a correlation between human judgments regarding fluency and BERTScore. More analysis is required to verify these contradictory findings.

MeaningBERT (Beauchemin et al., 2023) is a BERT model with a regression head that predicts a meaning preservation score between 0 and 100. The model is fine-tuned with complex-simple pairs of CSMD as input and the meaning preservation value of the human assessments as a label. They achieved their highest correlation with human judgments when augmenting their data with identical complex-simple pairs (with the label of 100) and totally unrelated complex-simple pairs (with the label 0).

Overall, on their test dataset they received a nearly perfect correlation (i.e., 0.928) between their proposed metric MeaningBERT and the human judgments. More analysis is required to check if the metric overfits to their data or if the high correlation also holds when evaluating on other datasets. Due to the recent publication date of this metric, it has not yet been used in other text simplification studies to verify if the results also hold on other datasets.

The trained model of MeaningBERT score is only applicable to English TS evaluation as an English BERT model has been trained with English sentence pairs as input. However, the approach could be tested on other languages and other TS sets if enough human ratings are available.

5.2.1.2 Grammaticality

As discussed in the previous section, in TS research, BLEU or BERTScore are also utilized to measure the fluency or grammaticality of the system outputs. However, Sulem et al. (2018a) (evaluated on HSplit) and Alva-Manchego et al. (2021) have shown that other metrics, e.g., the negative Levenshtein distance (Levenshtein, 1966) or the proportion of deleted words, correlate more with human judgments regarding grammaticality. However, BLEU is still the common metric to evaluate the grammaticality of TS outputs (Alva-Manchego et al., 2020b).

Furthermore, although promising reference-less evaluation metrics for grammar error correction are applied to related NLG tasks, e.g., Napoles et al. (2016); Asano et al. (2017) and in text summarization, e.g., Hardy et al. (2019), automatic error detection or correction has not been applied to text simplification evaluation yet.

5.2.1.3 Simplicity

The evaluation of simplicity can again be separated into measuring the simpleness of a generated text, increase of simpleness, and extent of change wrt. simpleness as introduced in Subsubsection 5.1.2.3. Readability metrics are often utilized to automatically measure the simpleness of a text, e.g., FKGL (see Subsubsection 5.2.1.3), or FRE (see Subsubsection 5.2.1.3). In addition, the newly introduced SLE score could also be described as a simpleness metric (see Subsubsection 5.2.1.3). However, the increase of simpleness and its extent of change is not easy to be separated in the scope of automatic metrics. In order to measure them, I am aware of two metrics: FKBLEU (see Subsubsection 5.2.1.3) and FRE-BLEU (see Subsection 5.3.3).

Further, to the best of my knowledge, no metric is designed with the intention of mirroring the extent of simplicity of the generated simplification considering the original text. However, all metrics that significantly correlate with human simplicity judgments could be interpreted as metrics to evaluate simplicity, e.g., BERTScore-precision (see Subsubsection 5.2.1.1), or SARI (see Subsubsection 5.2.1.4).

FKGL & FRE & OTHER READABILITY METRICS have been proposed upon the two prominent readability metrics, i.e., Flesch-Kincaid Grade Level (FKGL) and Flesch Reading EaseFlesch. FRE and FKGL are often used to evaluate the simpleness of generated simplifications. The scores rely on language-dependent constants, average word length, and average sentence length. The results are on a scale between 0 to 100 (in the case of FRE, where the higher the score the easier the text), or a value usually between 0 and 13 (in the case of FKGL) to estimate the school grade which is required to understand the text.

Readability metrics have been developed for many languages, as they are language-dependent due to their constants, and, hence cannot be applied to other languages without adaptation. For example, FKGL and FRE scores were originally designed to estimate the readability of English texts, but Amstad (1978) also proposed a version of FRE that is suitable for German texts. To automatically calculate the readability scores of texts, the Python package textstat[18] is often used because it comprises implementations of readability metrics of eight languages (i.e., English, German, Spanish, French, Italian, Dutch, Polish, and Russian).

Although these scores were originally designed to measure the readability of documents and manually written texts, they are frequently applied to assess the readability of automatically generated sentences in TS evaluation (Alva-Manchego et al., 2021). Following Xu et al. (2016) (evaluated in TurkCorpus), Alva-Manchego et al. (2021) (evaluated on Simplicity-DA) and Maddela et al. (2023) (evaluated on SimpEval, Simplicity-DA, & Newsela-Likert), simplicity gain and the extent of simplicity weakly correlate with FKGL, while Zhao et al. (2023) (evaluated on PWKP) could not even find a significant correlation between the simplicity gain and FKGL.

Tanprasert and Kauchak (2021) argue that FKGL is not a suitable metric for the evaluation of text simplification because it is easily manipulable with small post-processing adaptations. Following this, they argue that FKGL should no longer be used to evaluate TS outputs. Furthermore, (Alva-Manchego et al., 2020b) warn about the missing correlation between FKGL and grammaticality as a text with a high FKGL score can include many grammatical errors.

FKBLEU (Xu et al., 2016) aims to measure meaning preservation and simplicity at the same time. FKBLEU is the product of iBLEU (see Subsubsection 5.2.1.1) and FKGL (see Subsubsection 5.2.1.3). iBLEU is intended to measure the meaning preservation considering the original sentence, the manual simplified reference(s) and the system generation. The degree of simplicity is measured with the difference between the FKGL score of the system output and the FKGL score of the original text. The higher the score, the higher the simplification quality. Following the evaluation study of Xu et al. (2016) (evaluated on TurkCorpus), FKBLEU sig-

18 https://github.com/textstat/textstat [last update: June 3, 2024; last access: July 24, 2024]

nificantly and moderately correlates with grammaticality and meaning preservation, but only weakly with simplicity gain. However, the correlations with FKBLEU have a lower coefficient than the correlations with BLEU (strong correlation with meaning preservation and grammaticality) and with SARI (moderate correlation with simplicity), hence, the other scores seem to be more reliable (at least in this study). These findings are partially confirmed by Zhao et al. (2023) (evaluated on PWKP), but in their study FKBLEU negatively and weakly correlates with simplicity gain. Alva-Manchego et al. (2020b) argue that FKBLEU is less reliable in comparison to other metrics because it is built on top of FKGL and inherits its problems.

A German adaptation of FKBLEU is FRE-BLEU, instead of FKGL it contains the German FRE version by (Amstad, 1978).

SLE (Simplicity Level Estimate) (Cripwell et al., 2023c) is an extension of FKBLEU. Cripwell et al. (2023c) propose to calculate the differences between the system-generated simplification and the original text by predicted simplicity levels. They have fine-tuned a RoBERTa model (Liu et al., 2019) with sentences from the Newsela corpus (Xu et al., 2015) and its corresponding simplicity level. Therefore, the model is mostly applicable to evaluate sentence simplification (and not document simplification). To measure the degree of a system-generated simplification, they calculate the difference in the simplicity level of a system generation and of the original text.

Cripwell et al. evaluated the reliability of their SLE score by measuring the correlations between human judgments on simplicity and the SLE scores on two datasets, i.e., Human-Likert Scialom et al. 2021 and Simplicity-DA Alva-Manchego et al. 2021. SLE correlates significantly with simplicity on both datasets and obtained higher correlation coefficient values than for FKGL, BERTScore, or SARI, but slightly lower values than LENS (see Subsection 5.2.2).

On the one hand, this score has the advantage over LENS and SARI that it is a reference-less metric. On the other hand, it also includes the following major limitations: i) it has been fine-tuned on human generated data, but it is used to test machine generated texts, which are error-prone, and ii) it tackles only the extent of simplicity and no other aspect of simplification evaluation or simplification operations. Further scores in the same directions are readability assessment tools, e.g., see Deutsch et al. (2020), Weiss et al. (2021), Lee and Vajjala (2022), or Imperial and Tayyar Madabushi (2023). Due to its recent publication, SLE has not yet been used in any text simplification study.

5.2.1.4 Structural & Lexical Simplicity

Besides general simplicity, its measurement can also be split into structural and lexical simplicity. SAMSA (Sulem et al., 2018b) is a metric especially designed for syntactic simplification; I will explain this metric in more detail below. For lexical simplification, to the best of my knowledge, there is no automatic metric which has been specially designed to evaluate it. However, SARI (Xu et al., 2016) could be considered as a lexical simplification metric even if originally designed to measure the fineness of three edit-based operations in the simplification pair, i.e., add, keep, and delete. I will also explain SARI below in more detail.

SAMSA (Simplification Automatic evaluation Measure through Semantic Annotation) (Sulem et al., 2018b) is a metric especially designed for syntactic simplification. It focuses on the evaluation of syntactical text simplification by comparing the existence and movement of event structures in the source and the automatically simplified sentence. Therefore, they utilize semantic structures, such as UCCAs (Abend and Rappoport, 2013) or abstract meaning representation (Banarescu et al., 2013). The highest SAMSA score can be achieved when all event structures of an input sentence result in a separated sentence that includes all minimal core elements and arguments.

SAMSA is a reference-less metric because it considers only the system output and the original sentence. Its aim is to break down the original sentence into minimal propositions, which are automatically annotated by a semantic parser; hence no reference or manual simplification is required for the evaluation. In principle, this enables evaluation of TS systems on complex texts for which no manual simplification exists, e.g., in under-resourced domains of English TS. In reality, SAMSA still might not be suitable for low-resource languages as its major limitation is a good semantic parser for the language of interest, which often does not exist. Further, SAMSA is not a metric for measuring the overall quality of system-generated simplification as it focuses only on structural simplification. Sulem et al. (2018b) argue that structural and lexical simplification can be separated during evaluation and that an overall simplification score is not required.

Alva-Manchego et al. (2021) (evaluated on Simplicity-DA) show in their study that SAMSA receives low correlation scores with human judgments of structural simplification, although it is designed for it. They argue that this might be due to the TS systems selected for evaluation because the original study only evaluated structural simplification systems, whereas they evaluated with general TS systems. More research is required to analyze the behavior of SAMSA wrt. different TS models, TS evaluation data sets, semantic structures, and languages. Due to more insights regarding the evaluation capacity of SAMSA, it has been rarely used for TS evaluation in recent years (e.g., see Niklaus et al. 2019b; Alva-Manchego et al. 2019a). An extension of SAMSA, known as SEMA, has been proposed by Zhang et al. (2020). But, as with SAMSA, it has yet to be thoroughly analyzed to provide more information about its advantages.

SARI (Comparing **S**ystem output **A**gainst **R**eferences and the **I**nput sentence) (Xu et al., 2016) is a n-gram-based metric developed especially for the evaluation of text simplification. Putting it simply, it compares a generated simplification with the source sentence and at least one simplified reference to estimate the quality of the generated text wrt. (lexical) simplification. SARI is comprised of the evaluation of three edit-based operations, i.e., add, keep, and delete, which are added to a score ranging from 0 to 100, where 100 is intended to be a perfect simplification.

I explain the score by giving more details for one of the operations, i.e., the 'add' operation before explaining the combination of all three operations: SARI_{add} measures the F1-score_{add} of the added n-grams which is the harmonic mean of the precision_{add} and the recall_{add}. In this case, the recall is the ratio of the correctly added n-grams (n-grams that occur in the system generation and in at least one of the reference texts, but not in the original text) to all added n-grams in the reference texts (n-grams that occur in the reference texts, but not in the original text). The precision is here the ratio of correctly added n-grams (n-grams that occur in the

system generation and in at least one of the reference texts, but not in the original text) to all added n-grams (n-grams that occur in the system generation but neither in the original nor any of the reference texts). Simply put, SARI penalizes all words that are added that should *not* have been added, but rewards words that are added and should have been added.

The F-Score$_{add}$ is measured for all variants of n-grams (usually up to 4-grams) and then added up.[19] The same is repeated for all three edit operations, resulting in F1-Score$_{add}$, F1-Score$_{keep}$, and F1-Score$_{delete}$. SARI is then the arithmetic average of the two F1-Scores of the add and keep operation, as well as the precision score of the delete operation (Xu et al., 2016).

Looking at a concrete scoring example of item 1 to item 6, system 1 (see item 3) generates the worst simplification wrt. SARI (45.48) as too few correct n-grams have been added and kept (see Table 5.4).[20] System 1 has slightly added more correct n-grams, hence it is ranked slightly better than system 4 (SARI: 46.76). System 2 (see item 4) achieves an even better SARI score (SARI = 56.24) because "vorliegen" is replaced with "vorhanden" which is also expected in the reference simplification. The best simplification wrt. SARI (66.37) or the simplification most close to the reference is the simplification of system 3 (see item 5): "veröffentlichungspflichtig" is correctly simplified, but, following the reference, "vorliegen" could be still more simplified by replacing it with "vorhandenen".

Comparing the SARI scores to BERT-Score and BLEU, the small textual difference between system 1 and 2 has again a higher effect in the score (such as for BLEU) than for BERT-Score, which is due to their n-gram approach vs. BERT-Score embedding approach. Also, BERTScore-F1 and SARI would rank the systems differently: system 4 would be second best wrt. BERTScore-F1 or BERTScore-Precision but last wrt. SARI. Furthermore, system 4 is ranked worst wrt. SARI although the sentence structure might be better readable for the plain language target group due to the rewriting of passive to active voice. The reasons for this discrepancy are, on the one hand, that in this example, the system output can be only evaluated against one gold reference (with fewer syntactical changes) and, on the other hand, that SARI focuses more on lexical and less on syntactical simplification.

	add	**keep**	**delete**	**SARI**
System 1	20.41	21.43	98.33	46.76
System 2	48.96	21.43	98.33	56.24
System 3	74.11	25.00	100.0	66.36
System 4	16.67	21.43	98.33	45.48

Table 5.4: Scoring example of SARI of four system outputs (see item 1 to item 6).

In the following, I will go into more details of SARI's strengths and weaknesses as reported in related works. SARI's capability for TS evaluation is measured by correlations against human judgments of meaning preservation, grammaticality, and simplification gain (number of correct simplification operations). Xu et al. (2016) (evaluated on TurkCorpus) found a significant

19 (Alva-Manchego, 2019) identified a difference between the procedure described in the article and the code provided to replicate the score. It is unclear whether the scores per n-gram are averaged before or after calculation of the F-Score. However, I describe here the procedure as it is mentioned in the SARI paper.

20 The SARI scores reported here are generated with EASSE-DE Stodden, 2024a and the following settings: lowercase: false; language: DE; tokenizer: SpaCy.

moderate correlation of the SARI scores with human judgments regarding all three. However, BLEU reaches a higher correlation (moderate) for meaning preservation and grammaticality, while SARI has the highest correlation with simplicity gain in comparison to all other scores.

SARI has also been evaluated in other evaluation studies in comparison to newer metrics which also tackle simplicity: On the one hand, Alva-Manchego et al. (2021) (evaluated on Simplicity-DA) could confirm and extend the findings that values also moderately correlate with human simplicity and structural simplicity assessments. But, on the other hand, they mitigate the findings because the correlations seem to hold only for the part of the system generations with low quality. If a generated simplification is manually rated as high quality, the SARI scores are not reliable because the correlations are low.

Maddela et al. (2023) additionally report moderate correlations of SARI and overall simplicity (evaluated on SimpEval), meaning preservation, grammaticality, and simplicity (evaluated on Simplicity-DA). However, their newly proposed metrics (called LENS, see Subsection 5.2.2) could reach higher correlation than SARI.

In addition, Zhao et al. (2023) (evaluated on PWKP) analyzed the effect of the number of references included in measuring SARI. If using multiple references, they could find low to moderate correlation between SARI and meaning preservation, grammaticality, and simplicity gain, but evaluating against only one reference resulted in non-significant correlations with grammaticality and meaning preservation and only low significant correlation with simplicity gain. However, they found their newly proposed metrics BETS-simp (see Subsection 5.2.2) to correlate higher with simplicity judgments, although not comparing to any reference simplification. Following this, the quality and capability of SARI can be doubted in order to evaluate text simplification in lower resource languages or lower researched domains.

In addition to criticism against the correlation of human judgments, SARI has further disadvantages as it only considers three simplification transformations that predominantly tackle the lexical simplification (Alva-Manchego et al., 2020b), although many more lexical and syntactical transformations could be applied, as shown in Subsection 2.1.4. Indeed, all operations which consider $n : m$ alignments are neglected as SARI expects the transformations to be within 1:1 alignment pairs. Hence, it neglects other alignment types and focuses only on short-distance edits (Alva-Manchego et al., 2020b).

Furthermore, SARI has again the same advantage and disadvantage as BLEU and ROUGE as they all are n-gram-based metrics. On the one hand, the metrics are all language-independent. On the other hand, substituted words can be wrongly penalized when the substitute does not occur in any of the references even if the substituted word is a correct simplification, e.g., a simpler synonym of a complex word.

In conclusion, SARI is intended to measure how often the simplification operations of adding, deleting, and retaining are correctly applied by ignoring other simplification operations and neglecting wrong or missing simplifications. Despite these aspects, SARI is still the most widely used metric for the evaluation of text simplification, even if mainly in combination with other scores such as BLEU or BERTScore.

5.2.1.5 Coherence

The evaluation of coherence is also often neglected in the automatic evaluation of TS, which is also due to less progress in document than sentence simplification research. For example, Todirascu et al. (2013) found that the cohesion and coherence features of the text are helpful in predicting the readability of a text. In addition, Vásquez-Rodríguez et al. (2023) propose a method on how to evaluate coherence in the simplification of English documents. Overall, more investigation is required on how to integrate coherence into TS evaluation.

5.2.2 Overall Quality of Simplicity

In contrast to the previously named metrics, a metric for overall simplicity quality should combine several aspects of TS evaluation, e.g., aspects of the system output (independently of the source or the reference) such as readability and grammaticality. Further, it should also evaluate the quality of the output with respect to the source and the target sentence, e.g., meaning preservation and usage of simplification operations. A few recent metrics tackle the automatic measurement of overall simplicity quality: e.g., LENS and BETS.

LENS (**L**earnable **E**valuation **M**etric for **T**ext **S**implification) (Maddela et al., 2023) is a recently proposed method to evaluate ATS models regarding overall simplification quality. Following Maddela et al. (2023) (evaluated on SimpEval), none of the previously described metrics, for example, FKGL, BLEU, SARI or BERTScore, at least moderately correlate with overall simplicity ratings. Hence, they propose their own metric to close this gap.

Their metric is trained on human assessments regarding meaning preservation, fluency, and overall simplicity of English TS system outputs. Maddela et al. report moderate to strong correlations of LENS with grammaticality, meaning preservation, and simplicity ratings (evaluated on SimpEval). However, LENS is only trained with English TS outputs and its assessments; hence, the current version is not applicable to other languages than English. Furthermore, the score is difficult to reproduce for other languages as human assessments are rare, especially in non-English languages, and if available, they are rated following different evaluation aspects or different item descriptions of them, which might hamper the reproduction.

BETS (BERT Embedding-based evaluation for Text Simplification) (Zhao et al., 2023) is a recent reference-less evaluation metric to estimate grammaticality, simplicity, and meaning preservation at the same time. Based on contextualized BERT embeddings (see Devlin et al. 2019) of the original text and the system output, they measure their similarity to estimate meaning preservation. Furthermore, they fine-tune BERT to judge whether a word of the original text is more complex than that of the simplified text (Zhao et al., 2023). Then, they combine both resulting scores into a score which should mirror the overall quality of simplicity. Following, their evaluation (evaluated on their own data), BETS has a higher significant correlation with the compared score of grammaticality, simplicity, and meaning preservation than SARI, SAMSA, BERTScore, or FKBLEU. As LENS and BETS have been simultaneously but independently published, no comparison between both metrics exists yet.

Currently, BETS is only available for English due to the required embeddings of the English BERT model. However, the approach could be transferable to other languages when changing the pre-trained model to a multi-lingual model.

5.2.3 Reference-less Metrics

A few presented metrics do not rely on the simplified references (the gold simplification) to evaluate the system outputs; these metrics are also called reference-less metrics. Examples of this type of metrics are: FKGL, FRE, SLE, or SAMSA. Further approaches on reference-less evaluation are proposed by Martin et al. (2018).

Following the criticism of the scores regarding low interpretability, Martin et al. (2018) argue to not only focus on the scores, but to evaluate linguistic characteristics of the simplifications generated by the system. Therefore, they propose a feature extraction toolkit (further called TS-eval) to get more insights regarding the linguistic changes of the system simplification in comparison to the original texts. Their toolkit includes, for example, word concreteness, statistics on the number of characters and syllables, the position of the words in a frequency table, and some of the previously introduced metrics (e.g., BLEU, ROUGE or FKGL). However, their framework focuses only on English, and in the current version, it is not applicable to other languages.

5.2.4 EASSE: Evaluation Framework

The EASSE framework (Alva-Manchego et al., 2019a) is designed for the ease of evaluation of English automatic sentence simplification. It contains the implementation of automatic evaluation metrics, including SARI, BLEU, SAMSA, FKGL, and BERTScore. In addition, the tool can be used to build an evaluation report on all specified metrics of all specified TS models to facilitate the entire evaluation process. They have built evaluation reports with several metrics, including several system outputs and test sets, which are also part of the framework.

Even if EASSE contains several English resources, it can also be used with other custom English TS resources or TS resources in other languages. Hence, it is also often used to evaluate models of other languages, e.g., Spanish (e.g., see Gonzalez-Dios et al. 2022 or Holmer and Rennes 2023), French (e.g., see Cardon and Grabar 2020), or for the evaluation of a multi-lingual TS benchmark (Ryan et al., 2023). However, it raises some problems when using it for a different language, e.g., 1. the BERTScore is evaluated on an English-only BERT model, 2. the tokenizer is not adapted to the language, and 3. the readability scores are also only designed for English.

5.3 German Evaluation Studies

In the previous section, I have focused on the evaluation of English TS because most of the research regarding TS evaluation is conducted on English. For German TS, only a few evaluation studies exist; most of them include automatic evaluation, but a few also contain manual evaluation. In this section, I will introduce the procedure for German TS evaluation starting first with manual evaluation studies (see Subsection 5.3.1) and continuing with an overview of Ger-

man datasets that include human judgments regarding TS evaluation (see Subsection 5.3.2). I conclude this section with automatic evaluation studies (see Subsection 5.3.3).

5.3.1 Manual Evaluation

A similar picture, as presented for manual evaluation of English TS, also emerges for human evaluation of automatic German text simplification. First, because of its high effort and long duration time, human assessments regarding outputs of German TS systems are collected rarely. However, for German TS, out of overall 17 German TS studies, I am aware of the following studies that include human evaluation (see Table 5.5), i.e., Säuberli et al. (2020), Mallinson et al. (2020), Trienes et al. (2022), Anschütz et al. (2023), Deilen et al. (2023), Schlippe and Eichinger (2023), Fruth et al. (2024), Klöser et al. (2024), and Säuberli et al. (2024).

5.3.1.1 Manual Evaluation of Automatic Text Simplification – Intrinsic

In most of the German manual evaluation studies, an intrinsic evaluation has been conducted by asking questions regarding a few evaluation aspects as introduced in Subsection 5.1.2. Only Suter et al. (2016), Trienes et al. (2022), Deilen et al. (2023), and (Fruth et al., 2024) differ from the common evaluation approach since they report qualitative observations after manual inspections (and no quantitative scores as a result of an intrinsic evaluation).

	Meaning Preservation	**Grammaticality**	**Simplicity**	**Simplicity (simple)**	**Fluency**
Suter et al. (2016) *					
Niklaus et al. (2019a) *					
Siegel et al. (2019) *					
Mallinson et al. (2020)	meaning adequacy	grammaticality (simple)	simplicity	-	-
Säuberli et al. (2020)	content preservation	fluency of output	relative simplicity	-	-
Rios et al. (2021) *					
Trienes et al. (2022) *					
Spring et al. (2021) & Ebling et al. (2022)					
Anschütz et al. (2023)	-	grammaticality (pairwise)	-	-	-
Deilen et al. (2023) *					
Ponce et al. (2024) *					
Ryan et al. (2023) *					
Schlippe and Eichinger (2023)	content	grammar	simplification	comprehensibility	fluency
Schomacker et al. (2023a) *					
Fruth et al. (2024) *					
Klöser et al. (2024)	content similarity	-	-	-	-
Säuberli et al. (2024)	-	-	-	difficulty	-
Stodden and Kallmeyer (2022)	meaning preservation	grammaticality	overall simplicity	simplicity (simple)	-
Stodden et al. (2023)	meaning preservation	grammaticality (simple)	overall simplicity	simplicity (simple)	-
Stodden (2024b) *					

Table 5.5: Names of evaluation aspects per German TS study. Last part shows own contributions. Studies marked with * do not contain intrinsic evaluation of German TS.

The German studies show the same inconsistency of intrinsic human evaluation studies as for English TS evaluation. As summarized in Table 5.5, 6 of 17 German TS studies include intrin-

sic manual evaluation. Furthermore, 3 of the 6 studies evaluate the aspects of meaning preservation, grammaticality, and simplicity (see Mallinson et al. 2020, Säuberli et al. 2020, Schlippe and Eichinger 2023). Schlippe and Eichinger (2023) and Säuberli et al. (2024) further evaluate the simplicity of only the simplified sentence in addition to the extent of the simplicity between the original and the simplified text. In one study, i.e., Anschütz et al. (2023), only evaluates grammaticality, but they do not rank it on a scale; instead, they compare the grammaticality pair-wise between the original sentence and the system output. The grammaticality is also split into analysis of grammar and fluency in another study, i.e., Schlippe and Eichinger (2023).

	Meaning Preservation	**Grammaticality**	**Simplicity**	**Simplicity (simple)**	**Fluency**
Mallinson et al. (2020)	1 to 5	1 to 5	1 to 5	-	-
Säuberli et al. (2020)	0 to 3 †	0 to 3 †	0 to 3 †	-	-
Schlippe and Eichinger (2023)	1 to 5	1 to 5	1 to 5	1 to 5	1 to 5
Säuberli et al. (2024)	-	-	-	1 to 5	-
Klöser et al. (2024)	1 to 3	-	-	-	-
Stodden and Kallmeyer (2022)	1 to 5	−2 to +2	−2 to +2	1 to 5	-
Stodden et al. (2023)	1 to 5	−2 to +2	−2 to +2	1 to 5	-

Table 5.6: Scale points per German human TS evaluation study. All scales are described as Likert-Scales except for the ones marked with †, for these each scale point is labeled content-wise. Last part shows own contribution.

When grouping the German evaluation aspects per study to the previous categorization aspects (see Subsection 5.1.2), we again see various names for the same aspect (see Table 5.5) and different scale sizes (see Table 5.6). Regarding naming, e.g., meaning preservation is named as "content preservation", "meaning adequacy", "content similarity", or simply "content". Regarding scale size, e.g., Säuberli et al. (2020) evaluates on a scale with even scale points (from 0 to 3), while the other studies all evaluate on a scale with 5 points. Following this, the scales do not always contain a neutral element (e.g., no, 0 or 3). However, similar to English scales (see Subsection 5.1.2), it is not specified how the neutral element should be interpreted.

Furthermore, statements (or questions) per aspect are not consistently verbalized (see Table 5.7). For example, the aspect of simplicity is once verbalized as the extent of simplification (see Schlippe and Eichinger 2023 in Table 5.7) and once in a dichotomous question whether the original text has been simplified (see Mallinson et al. 2020). Mallinson et al. (2020) also adds information gain as a restriction to the aspect of meaning preservation, whereas Schlippe and Eichinger (2023) does not explicitly address neither information gain nor information loss.

In all TS study papers, only an English version of the verbalization is provided. It is not clear whether the statements have been translated into English to be better understandable by the research community or if the English version has also been shown to participants during the evaluation study.

In summary of the previously named differences, each study measures the quality of the simplifications slightly differently, which makes the results incomparable (Štajner, 2018; Stajner, 2021).

When comparing the group of raters and the number of samples manually evaluated, more disagreement is found (see Table 5.8). In all studies, the annotators are German native speakers who might give reliable judgments for meaning preservation and grammaticality because they

	Meaning Preservation	Grammaticality	Simplicity	Simplicity (simple)	Fluency
Mallinson et al. (2020)	To what extent is the meaning expressed in the original sentence preserved in the output, with no additional information added?	Is the output grammatical and fluent?	Is the output a simpler version of the input?	-	-
Säuberli et al. (2020)	*n/a*	*n/a*	*n/a*	-	-
Schlippe and Eichinger (2023)	How well does the simplified text reproduce the original text in terms of content?	How correct is the simplified text in terms of grammar?	How well is the text simplified?	How understandable is the simplified text?	How smoothly does the simplified text read?
Säuberli et al. (2024)	-	-	-	*n/a*	-
Stodden and Kallmeyer (2022)*	The simplified sentence adequately expresses the meaning of the original sentence, perhaps omitting the least important information.	The simplified sentence is fluent, there are no grammatical errors.	The simplified sentence is easier to understand than the original sentence.	The simplified sentence is easy to understand.	
Stodden et al. (2023)*	The simplified sentence adequately expresses the meaning of the original sentence, perhaps omitting the least important information.	The simplified sentence is fluent, there are no grammatical errors.	The simplified sentence is easier to understand than the original sentence.	The simplified sentence is easy to understand.	-

Table 5.7: Statements and questions per evaluation aspect and German human TS evaluation study. Studies marked with * provide German (and English) statements. Last part shows own contributions.

know the German grammar and can understand both texts (i.e., the original and the simplified text) well. However, native German speakers without reading problems are mostly not part of the target group of the simplifications (except for expert-laypeople simplification); hence, even if they state that the sentences are better readable than before, it does not mean that the readability is also enhanced for non-native speakers (Štajner, 2018; Gooding et al., 2021) or other target groups. Following suggestions from Stajner (2021), the target group should be more involved in the process of evaluating text simplification, e.g., by extrinsic evaluation (see Subsection 5.1.1).

	n (total)	n (per item)	test set size
Säuberli et al. (2020)	1	1	50 pairs per model
Mallinson et al. (2020)		5	100 pairs per model & gold simplification
Schlippe and Eichinger (2023)	105	*n/a*	5 pairs per model
Säuberli et al. (2024)	36	*n/a*	12 pairs
Klöser et al. (2024)	*n/a*	*n/a*	135 pairs
Stodden et al. (2023)	2	1	430 pairs per gold simplification

Table 5.8: Number of participants and test set size per German human TS evaluation study. The participants in all studies are German native speakers. Last part shows own contributions.

5.3.1.2 Manual Evaluation of Automatic Text Simplification – Extrinsic

The evaluation of Säuberli et al. (2024) stands out in comparison to the other evaluation studies, as they conducted an extrinsic evaluation on the outputs of the system. 18 participants of the target group, i.e., "persons with intellectual disabilities", and 18 participants of a control

group (i.e., native German speakers) read texts either in standard German, manually simplified German, or automatically simplified German. They asked the participants multiple-choice questions regarding the overall content and content details of the texts to verify if they understood them correctly and to assess the texts' difficulty on a rating scale.

They also measured reading speed, the time to answer the questions, and scrolling interactions. They found that the results from the target group are partially contradictory to the results of the control group, e.g., the target group self-assessed the automatically simplified texts as most simple, whereas the control group self-assessed them as most difficult. However, a comparison between the different evaluation approaches shows that this type of measurement, i.e., difficulty rating, is the least accurate method for evaluating the results of the target group. Nonetheless, difficulty rating is the most frequently used measurement in other evaluation studies.

Following their results, the most reliable approach when evaluating with people with intellectual disabilities is the accuracy of the answers regarding the comprehension questions; for the control group, the reading time measurement was most successful and all other methods were of nearly equal quality. Overall, these findings underline the importance of evaluating text simplification outputs by the target group and call again into question the reliability of current approaches regarding rating scale-based evaluation.

Following upon this research, recently Carrer et al. (2024) suggest using a mixed-method design for evaluation of automatically generated texts, for example, combining results of scale-based evaluation and behavioral data of reading or comprehension studies. Additionally, they propose another method, i.e., evaluation based on effects during post-editing of professional translators.

5.3.2 German Dateset with Human Judgments

Unfortunately, only one dataset of the manual evaluation studies presented before is publicly available for further research purposes, i.e., the judgments used for evaluation in Säuberli et al. (2024). These data might be helpful for future research regarding a correlation analysis of automatic metrics and the judgments or training (or fine-tuning) automatic metrics on these judgments.

In addition, some resources exist regarding the annotation of complexity levels of a given text, e.g., see Subsection 4.7.3. These resources could be used to rebuild the SLE (see Subsubsection 5.2.1.3) for the German TS evaluation. Furthermore, TextComplexityDE (see Subsection 4.2.6) contains human judgments on the complexity of manually simplified German sentences. A shared task has been conducted to automatically assess the complexity of these sentences (see Subsubsection 5.3.3.2).

5.3.3 Automatic Evaluation

I have previously discussed previous work on manual evaluation of German TS including many challenges, e.g., missing best practices for German (see Subsection 5.3.1) as well as lack of publicly available datasets with human judgments (see Subsection 5.3.2). Automatic metrics are usually dependent on manual evaluation, e.g., by correlation analysis of metrics and human

judgments to prove that the automatic metrics are reliable (to some extent). However, as no such data exist for German TS, the current approach for German automatic TS evaluation is to rely on metrics for English TS. The common metrics for English TS (e.g., SARI and BLEU) have been applied several times to German TS evaluation yet despite missing justification for their reliability. Indeed, some studies fully rely on automatic scores by neglecting the previously named criticism.

To the best of my knowledge, prior to our work, in no German TS study have additional readability metrics or linguistic feature analysis been performed to obtain more insights into the quality of the generated simplification. In this section, I am now describing the automatic evaluation setup of German TS studies.

5.3.3.1 TS Evaluation Studies

All German TS studies have evaluated their TS models with SARI and BLEU (see Table 5.9). Mallinson et al. (2020) additionally include more versions of BLEU, i.e., I-BLEU (see Subsubsection 5.2.1.1) and FRE-BLEU. TS studies which tackle paragraph or document simplification also evaluate with ROUGE (see Rios et al. 2021, Trienes et al. 2022 and Anschütz et al. 2023). Only one study includes the BERTScore (see Ponce et al. 2024); although the BERTScore has recently been shown to be helpful for the evaluation of TS in the year 2021 (see Alva-Manchego et al. 2021), TS studies published after 2021 could already have included the BERTScore.

	SARI	BLEU	I-BLEU	FRE-BLEU	FRE or LIX	R-1	R-2	R-L	BERT-Score	TS-eval
Suter et al. (2016)					x					
Niklaus et al. (2019a) *										
Siegel et al. (2019) *										
Mallinson et al. (2020)	x	x	x	x						
Säuberli et al. (2020)	x	x								
Rios et al. (2021)	x	x						x		
Trienes et al. (2022)	x	x				x	x	x		
Spring et al. (2021) & Ebling et al. (2022)	x	x								
Anschütz et al. (2023)	x	x						x		
Deilen et al. (2023) *										
Ponce et al. (2024)	x	x							x	
Ryan et al. (2023)	x	x								
Schlippe and Eichinger (2023)	x									
Schomacker et al. (2023a)		x						x	x	
Fruth et al. (2024)	x				x					
Klöser et al. (2024)	x	x								
Säuberli et al. (2024) *										
Stodden et al. (2023)	x	x			x				x	x
Stodden (2024b)	x	x			x				x	x

Table 5.9: Automatic scores used per German TS study. R-1 = Rouge-1, R-2 = ROUGE-2, R-L = ROUGE-L. Studies marked with * have not been automatically evaluated. Last part shows own contributions.

Some German TS papers (i.e., Trienes et al. 2022, Ryan et al. 2023, and Ponce et al. 2024) describe that they have evaluated their systems using the implementation of the metrics in the EASSE framework, although it is mainly designed to evaluate English TS (see Subsection 5.2.4).

Anschütz et al. (2023) did not use EASSE as it does not include ROUGE. Therefore, they use the implementations of BLEU, SARI, and ROUGE provided in Huggingface (Wolf et al., 2020). In other papers, e.g., Spring et al. (2021) or Rios et al. (2021), it is not mentioned which implementation of SARI or BLEU has been used. This points up another issue of automatic evaluation: the scores of one study might not be comparable to other studies as they are measured with different implementations of the metrics.

5.3.3.2 Complexity Assessment Shared Task

The complexity or simplicity of German texts has been primarily analyzed with readability formulas. However, in the research field of "text readability", a few trainable approaches beyond readability formulas have been proposed for German readability assessment, e.g., the works of Weiss and Meurers (2022), Klepp (2022a), or Seiffe et al. (2022). Unfortunately, none of those approaches have yet been used in the scope of text simplification evaluation.

In the same research direction, but with a greater focus on the simplification of German sentences, Mohtaj et al. (2022) have organized a shared task on the assessment of the complexity of German sentences. In more detail, participants have been asked to predict a mean opinion score, which is more or less the average complexity assessment of a German sentence rated by non-native German speakers. Hence, the task can also be described as predicting the subjective complexity of a sentence for a specific target group, here for non-native German speakers.

However, in comparison to the previous metrics, the evaluated sentences are manually generated and not system-generated; hence, it is not clear whether the resulting models can also predict the complexity or simplicity of system outputs well. According to the results of Alva-Manchego et al. (2021), the correlations between TS metrics and human judgments can highly vary with respect to the system-generations of varying TS models; Therefore, I assume that similar or even stronger findings could be revealed when comparing automatically and manually simplified sentences. To the best of my knowledge, the results of these shared tasks have not yet been used to evaluate text simplification output, which might be due to the previously named limitations.

5.4 Summary & Outlook

In summary, in this section, I have presented and discussed current evaluation methods for text simplification including manual and automatic evaluation. I have shown that human evaluation can be conducted extrinsic (e.g., measuring the usefulness for the target group) or intrinsic (e.g., measuring the readability and meaning preservation). With automatic metrics, it is tried to cover the aspects of intrinsic evaluation regarding simplification quality (e.g., SARI, BERTScore, or FKGL) or measure all at once (e.g., LENS or BETS).

In more detail, I can summarize from the state-of-the-art presented regarding human evaluation of German text simplification that the procedure varies in its design regarding i) the evaluation aspects, ii) design of the scale, iii) group of participants, iv) size of the test data, and v) availability of the ratings. Differences in the design of the evaluation study, especially different scales and annotator groups, make the results of the studies often unreliable and in-

comparable (Stajner, 2021). Best practices on how to manually evaluate text simplification are highly demanded for German as well as for other languages. Due to missing best practices and its high costs with respect to time and effort, human evaluation is rarely conducted for German as well as English TS systems.

An overview of the latest automatic metrics for TS is presented in Table 5.10. Overall, 8 of 15 metrics are source-dependent (hence, 7 source-free), 6 reference-based (hence, 9 reference-free), and 8 language-dependent. Currently, many metrics are used together to evaluate text simplification systems because they have all different advantages and disadvantages (as previously discussed), i.e., SARI, BERTScore, BLEU, and FRE. However, the metrics are criticized regarding i) being not suitable, e.g., BLEU or Flesch Reading Ease (Wubben et al., 2012; Xu et al., 2016; Sulem et al., 2018a), ii) considering only syntactical simplification, e.g., SAMSA (Sulem et al., 2018b), or lexical simplification, e.g., SARI (Xu et al., 2016), or iii) having low correlation with human judgments (e.g., see Martin et al. 2018 or Xu et al. 2016).

5.4.1 Challenges

For human and automatic evaluation of TS, I repeated or identified many challenges which also hold for evaluation of German TS. In the following, I concisely name the challenges and research gaps in the evaluation of German text simplification, i.e., varying scale design and scale interpretation (see Subsubsection 5.4.1.1), insufficient manual evaluation (see Subsubsection 5.4.1.2), reliability of automatic metrics (see Subsubsection 5.4.1.3), interpretability of automatic metrics (see Subsubsection 5.4.1.4), different settings and scores (see Subsubsection 5.4.1.5), and availability of human judgments (see Subsubsection 5.4.1.6).

These challenges also show again the relevance of finding answers to my research questions regarding robustness and reliability of evaluation (see RQ 5-1), their transferability to other languages than English (see RQ 5-2), and finding and improving the key aspects for manual and automatic evaluation (see RQ 5-3).

5.4.1.1 Scale Design & Scale Interpretation of Manual Evaluation

As mentioned above and also in previous work (e.g., see Stajner 2021), a unified approach for manual evaluation is required to make the results reliable and better comparable, as already mentioned in previous work (further called Evaluation Challenge A). For example, it is currently not clear which scale size is most suitable for extrinsic evaluation or if full scale descriptions are more favorable than scale statements. In addition, no analysis has yet been conducted regarding the quality of human judgments and the quality of the scale of human judgments. In some papers, inter-annotator agreement is used to assess the consistency of the ratings of the participants, but less attention is paid to the evaluation criteria. As shown previously, several different terms and definitions are used to evaluate the same concept. For example, "meaning preservation" is often also called "adequacy" or "content similarity", but only sometimes does the concept also include measuring the information gain. However, the different concepts can only be identified by a closer look at the terms or the questions used for evaluation. Additionally, in some studies, the terms are not further explained or defined. This makes it

Name	Authors	Aspects				Original-dependent	Reference-dependent	Language-dependent
		Meaning Preserv.	Grammaticality	Simpl.	Overall Simpl.			
BLEU	Papineni et al. (2002)	x	x				x	
iBLEU	Sun and Zhou (2012)					x	x	
ROUGE	Lin (2004)	x					x	
BERTScore	Zhang* et al. (2020)	x	x	x	x	x	x	x
MeaningBERT	Beauchemin et al. (2023)	x						
Levenshtein distance	Levenshtein (1966)		x				x	
FKGL & FRE	Flesch (1948)			x				x
FKBLEU	Xu et al. (2016)	x		x		x		x
SLE	Cripwell et al. (2023c)			x		x		x
SAMSA	Sulem et al. (2018b)			x		x		
SARI	Xu et al. (2016)	x	x	x	x	x	x	
LENS	Maddela et al. (2023)	x	x	x	x			x
BETS	Zhao et al. (2023)	x	x	x	x			x
Quality Estimation Metrics	Martin et al. (2018)		x	x		x		x
REFeREE	Huang and Kochmar (2024)	x	x	x		x		x

(**a**) Relevance of metrics per aspect. I marked a metric as relevant for an aspect, if ≥ 1 study supports it.

Name	Authors	Acceptance	Criticism
BLEU	Papineni et al. (2002)	Xu et al. (2016), Alva-Manchego et al. (2020a), Zhao et al. (2023)	Wubben et al. (2012), Sulem et al. (2018a), and Zhang et al. (2020b)
iBLEU	Sun and Zhou (2012)	Xu et al. (2016), Mallinson et al. (2020), Zhao et al. (2023)	
ROUGE	Lin (2004)	Beauchemin et al. (2023)	Ganesan (2018)
BERTScore	Zhang* et al. (2020)	Alva-Manchego et al. (2021), Maddela et al. (2023), Beauchemin et al. (2023)	Maddela et al. (2023), Zhao et al. (2023)
MeaningBERT	Beauchemin et al. (2023)		
Levenshtein distance	Levenshtein (1966)	Sulem et al. (2018a), Alva-Manchego et al. (2021)	
FKGL & FRE	Flesch (1948)	Xu et al. (2016), Alva-Manchego et al. (2021), Maddela et al. (2023)	Zhao et al. (2023), Tanprasert and Kauchak (2021), Alva-Manchego et al. (2020b)
FKBLEU	Xu et al. (2016)	Zhao et al. (2023)	Alva-Manchego et al. (2020b)
SLE	Cripwell et al. (2023c)		
SAMSA	Sulem et al. (2018b)	Zhang et al. (2020), Alva-Manchego et al. (2021)	Alva-Manchego et al. (2021)
SARI	Xu et al. (2016)	Alva-Manchego et al. (2021), Maddela et al. (2023), Beauchemin et al. (2023), Zhao et al. (2023)	Alva-Manchego et al. (2020b), Alva-Manchego et al. (2021), Maddela et al. (2023), Zhao et al. (2023)
LENS	Maddela et al. (2023)	-	-
BETS	Zhao et al. (2023)	-	-
Quality Estimation Metrics	Martin et al. (2018)	-	-
REFeREE	Huang and Kochmar (2024)	-	-

(**b**) Acceptance and criticism per metric.

Table 5.10: Summary of automatic metrics. The vertical lines separates the metrics by their aspects they belong to most: 1) meaning preservation, 2) grammaticality, 3) simplicity, 4) overall simplicity, 5) other.

very challenging, or even impossible, to comprehend which specific concept is employed for the evaluation.

To reduce the vagueness of the manual evaluation procedure, many points should also be discussed in the TS community, for example, whether simplicity should either measure the extent of simplification (where the lowest valuable determines no-simplification and more complex) or measure also the extent of making a sentence less simple (which is more complex). I cannot give any recommendations yet regarding which human evaluation procedure is most reliable or most promising, as further analysis of each aspect, statement, and scale length is required to make the evaluation solid. However, I can endorse including all relevant information from the human studies (including, e.g., questions per aspect, number of annotators per sentence pair, or demographics of annotators) in the report and publish the human ratings in conjunction with the system outputs. Overall, this would be beneficial for the progress in evaluation research and facilitate more accurate interpretation and replication of the judgments.

5.4.1.2 Insufficient Manual Evaluation (Intrinsic vs. Extrinsic)

As previously mentioned, the current manual evaluation aspects might not be sufficient to adequately evaluate sentence and document simplification (Alva-Manchego et al., 2020b) (further called Evaluation Challenge B). To better evaluate document simplification, a few studies rely on extrinsic rather than intrinsic evaluation. However, this approach is even more costly, so it is likely that extrinsic evaluation will remain rare in the future. To reduce the workload for researchers when building comprehension tests, Säuberli and Clematide (2024) have recently proposed a method to automatically generate items for these tests. Future will tell if this approach facilitates extrinsic evaluation or if still more work is required in this direction. Furthermore, a better and more extensive evaluation protocol is required for intrinsic evaluation, e.g., sub-aspects of grammaticality or simplicity. As proposed by Cumbicus-Pineda et al. (2021), the sub-aspects could be combined in an extensive evaluation checklist for TS.

In addition, manual annotation of system-generated errors during the simplification or classification of correctly performed simplification operations, e.g., see (Yamaguchi et al., 2023), could be beneficial to better interpret black-box TS models. These two approaches could be valuable to get more and better insights regarding the quality of system-generated simplifications besides simple scores such as SARI, BLEU, or BERTScore. Unfortunately, no annotation schema of simplification operations and simplification errors in German TS has existed before our work.

5.4.1.3 Reliability of Automatic Metrics

As a consequence of non-sufficient manual evaluation, automatic metrics might also not be reliable, as their quality is always evaluated against the human judgments (further called Evaluation Challenge C). I have previously outlined that automatic metrics have frequently been criticized regarding their quality for English TS evaluation (see Sulem et al. 2018a, Tanprasert and Kauchak 2021, or Alva-Manchego et al. 2021). Although the scores are criticized and were only evaluated against human annotations of English annotations (which correlations are not yet reproduced or repeated in other languages), these metrics are still the most commonly used

metrics to estimate the quality of TS models in any language including German. For evaluation of German TS models, these metrics are not yet evaluated and might be of even less quality than for English.

In conclusion, automatic evaluation especially helps in the development phase of text simplification systems, as it facilitates researchers to quickly assess which settings of their TS model (e.g., comparison of fine-tuned models, comparison of hyperparameters, or hand-crafted rules) are most or least successful. Nonetheless, in the test phase, it is recommended to perform a manual evaluation with the target group (Alva-Manchego et al., 2020b). Reasons for this are that automatic metrics indeed offer efficient evaluation, but manual evaluation is essential to capture nuanced aspects of text quality and comprehensibility. Furthermore, manual evaluation can consider various dimensions of simplification quality at the same time that might be challenging for automated methods yet.

5.4.1.4 Interpretability of Automatic Metrics

In addition to being not reliable or suitable, the automatic metrics for TS are also not interpretable (Alva-Manchego et al., 2021) (further called Evaluation Challenge D). Following Alva-Manchego et al. (2021), a high SARI does not always refer to a good simplification: the score might be influenced by other text edits than simplification. Therefore, more information about the simplification operations performed (Vásquez-Rodríguez et al., 2021; Cardon et al., 2022), the change in linguistic characteristics (Martin et al., 2018), or the errors that occurred in the generation of the system might be helpful to increase the interpretability of the metrics and the systems' quality (Devaraj et al., 2022; Yamaguchi et al., 2023). However, prior to our work, these approaches have not been transferred or applied to German TS.

Although these metrics and scores may be difficult to interpret, automatic scores can be less biased than human judgments (Alva-Manchego et al., 2020b). Simple algorithm-based automatic evaluation metrics (such as readability scores or n-gram-based metrics) can also provide quantifiable measures that help reduce human subjectivity from the assessment process (Alva-Manchego et al., 2020b). However, some automatic metrics (such as trainable metrics on large language models) are less transparent and do not allow inference on the decisions, e.g., explanations of why a score is higher or lower.

5.4.1.5 Varying Settings and Scores

Even if TS evaluation metrics would be reliable and interpretable, another challenge would remain (further called Evaluation Challenge E): the settings of the metrics would have only been optimized for English TS. It is not clear whether the same settings of a metric are suitable for English as well as for German, e.g., does tokenization or case-sensitivity have an effect on the metrics? Furthermore, in current studies, different implementations of the same metric are used (e.g., for ROUGE or SARI). The EASSE evaluation framework has unified the evaluation of English TS models and so nearly solved the problem for English TS, but the problems still exist for other languages, because EASSE only supports, for example, English readability scores or English linguistic feature analysis. An adaptation of EASSE to other languages would fa-

cilitate the evaluation of TS and make it more comparable. Prior to our work, something like transferable evaluation frameworks to other languages than English has not been available.

5.4.1.6 Availability & Number of References

Another challenge of TS evaluation is the lack of available human judgments (further called Evaluation Challenge F). To the best of my knowledge, no human judgments (e.g., regarding meaning preservation or simplicity) on complex-simple pairs are available yet. Human judgments on German TS data could help verify the quality of current automatic metrics for German TS or could be used to build new metrics for German TS.

Another problem of automatic evaluation for German compared to English is that most test sets contain only one gold reference (further called Evaluation Challenge F). Current TS metrics, e.g., SARI, are designed to be evaluated using multiple references; hence, the quality of SARI for German TS might be even lower than for English TS due to the lack of additional references.

5.4.2 Outlook

The current chapter contains a summary of the state-of-the-art on how to evaluate text simplification models as well as discussion on its challenges.

In the next chapter, Chapter 6, I introduce models and approaches which have been used for German text simplification. I will discuss TS models with respect to the size of their training data, as well as the target groups and domains on which they are trained and evaluated (see Chapter 4). In order to interpret the capabilities of the TS models, I will refer to the manual and automatic evaluation methods introduced in this chapter.

Chapter 6

German Text Simplification Models

In previous chapters, I have introduced corpora and evaluation methods for (German) text simplification. Building on this, in this chapter, I will describe the last missing component in the academic text simplification workflow, i.e., approaches towards automatically simplifying (German) complex texts (see step G in Figure 2.1 in Subsection 2.2.2). I will first start with a short timeline regarding the development of English TS models in the last years (see Section 6.1). Then, I will focus on German text simplification models or multi-lingual TS models which include German texts either in training or evaluation .[1] Most of the approaches presented are either rule-based or neural sequence-to-sequence models. Therefore, I split the last category more fine-grained as follows. The lines of research regarding German TS models can be split into:

(i) rule-based models, e.g., rule-based model by Suter et al. (2016), DISSIM (Niklaus et al., 2019a), and HDA-ETR (Siegel et al., 2019) (see Section 6.2),

(ii) training sequence-to-sequence models, e.g., sockeye-APA-LHA (Spring et al., 2021) and other sockeye variants (Ebling et al., 2022) (see Section 6.3),

(iii) fine-tuning pre-trained sequence-to-sequence models, e.g., mT5-MultiSim (Ryan et al., 2023) (see Section 6.4),

(iv) prompting with zero-shot and few-shot learning, e.g., ZEST (Mallinson et al., 2020), BLOOM in Ryan et al. (2023) or Ponce et al. (2024), or ChatGPT in Deilen et al. (2023) (see Section 6.5),

(v) combining auto-regressive language models and sequence-to-sequence models, e.g., custom-decoder-ats (Anschütz et al., 2023) (see Section 6.6), and

(vi) proprietary TS models (see Section 6.7).

In the following, I will briefly summarize and discuss the capabilities of text simplification systems, with a particular focus on the strategies and approaches applied to these systems.

1 Due to the fast pace of development of new model architectures or new pre-trained models, I am aware that this overview might not be up-to-date for a long time (especially in comparison to the existing evaluation methods and simplification corpora), but I hope that it can be a good starting point for new researchers in the field of German text simplification.

The models will be grouped according to their strategy or approach, with the oldest models appearing first and the newest models appearing last. In contrast to the grouping approach taken in Chapter X, where corpora were sorted according to their domains, the grouping in this section allows for a more straightforward comparison of models using the approach. This is because it is often challenging to make meaningful comparisons between models that have been evaluated on different test sets, even if their texts are from the same domain. Therefore, I will discuss the differences of the models per approach, rather than the results per domain. However, where possible, I will refer to similar models that have been tested on the same test set.

6.1 Chronicle of English TS Models

Early approaches of automatic text simplification have relied on hand-crafted rules to automatically simplify texts. For example, Carroll et al. (1998) proposed a TS model for English which first syntactically and then lexically simplifies a sentence following hand-crafted rules. For only structural simplification, Chandrasekar and Srinivas (1997) proposed rules to split appositions, relative, or subordinating clauses of nested sentences into separate sentences. The rule-based TS model of Siddharthan (2011) includes additionally the rules according to coordination and passive constructions.

In the next era of text simplification models, data-driven approaches have been the most prominent choice. In data-driven approaches, a model learns how to automatically simplify a text directly from complex-simple pairs. In contrast to the previous approaches, for this line of research parallel text simplification corpora are required. Following Alva-Manchego et al. (2020b), these approaches are statistical machine translation models (e.g., see Zhu et al. 2010 or Wubben et al. 2012), models with induction of synchronous grammars (e.g., see Feblowitz and Kauchak 2013 or Paetzold and Specia 2013), semantics-assisted models (e.g., see Narayan and Gardent 2016 or Štajner and Glavaš 2017), and neural sequence-to-sequence models (e.g., see Nisioi et al. 2017, Zhang and Lapata 2017, or Martin et al. 2020).

Some of the neural sequence-to-sequence models distinguish from the others because they are unsupervised (e.g., see Surya et al. 2019 or Martin et al. 2022) or use a sequence labeling approach (e.g., see Alva-Manchego et al. 2017, Dong et al. 2019 or Omelianchuk et al. 2021). In sequence labeling approaches, words or sequences are first labeled with, e.g., REPLACE, MOVE or KEEP, and then the named task is applied so that the simplification transactions are known and the result is easier to interpret (Alva-Manchego et al., 2017).

Most recent trends in neural text simplification are, on the one hand, to design models which are controllable so that the simplification can be adapted to the user's need (e.g., see Bingel et al. 2018, Nishihara et al. 2019, or Yanamoto et al. 2022). On the other hand, large language models are tested on how much they have learned about simplification during pre-training by few-shot learning (e.g., see Section 6.5) or prompting (e.g., see Ryan et al. 2023).

For a more extensive overview of text simplification models (including English, German, and other languages), I refer to Alva-Manchego et al. (2020b), Al-Thanyyan and Azmi (2021), or Espinosa-Zaragoza et al. (2023).

6.2 Rule-based Models

In order to build TS models with much control, including much expert knowledge and a high level of interpretability, rule-based models are a good choice (Liu et al., 2017). For this TS approach, experts write hand-crafted rules that can be used to automatically adapt a complex text to make it more readable. The engineering approach in this line of research focuses on finding the best rules or features to generate the best simplifications (Liu et al., 2023).

In this section, I will present three rule-based approaches for German TS, i.e., the rule-based model by Suter et al. (2016) (see Subsection 6.2.1), HDA-ETR by Siegel et al. (2019) (see Subsection 6.2.2), and DISSIM by Niklaus et al. (2019a) (see Subsection 6.2.3).

6.2.1 Rule-based Model by Suter et al. (2016)

To the best of my knowledge, the first German text simplification system was proposed by Suter et al. (2016). In their approach, they try to transfer the German Easy Language guidelines by Maaß (2015b) into an automatic rule-based simplification system. Their rules address the levels of characters & words (e.g., visually splitting compound nouns or replacing special characters and digits), sentences (e.g., cutting off (semi-)colon constructions or splitting subordinate clauses), and text & layout (e.g., adding word explanations from Hurraki[2]).

They evaluated their system with one short news article by calculating a readability score (i.e., LIX score Björnsson 1968; Heimann Mühlenbock 2013) on the original and simplified document. Indeed, their model could reduce the LIX score (from "difficult" to "fairly difficult") and, hence, enhance the readability to some extent. Additionally, they manually inspected the simplified data for errors and found that complex terms could have been simplified further, but are better readable due to visual segmentation.

Unfortunately, the implementation of the rules is not available, and the descriptions of the rules are not detailed enough to reproduce the model.

6.2.2 HDA-ETR

Similar to Suter et al. (2016), Siegel et al. (2019) also implements some rules of German Easy Language. They include their rules in LanguageTool[3], a rewriting tool that assists in giving recommendations on how to correct or improve a given input text. The tool of Siegel et al. (2019), named HDA-ETR, extracts passages in the text that do not conform to the Leichte Sprache guidelines and explains the reasons, e.g., if the text includes genitive, passive voice, numbers in words, abbreviations, conjunctive, indirect voice, long words, or special characters. Due to the complexity of some tasks, they only provide automatic rewriting of the difficult text segments regarding lexical substitution and compound splitting. However, their rules are applicable to simplify documents as well as sentences. Their code is available online[4]. But, unfortunately,

2 https://hurraki.de/ [last access: July 24, 2024]

3 https://languagetool.org/ [last access: July 24, 2024]

4 https://github.com/hdaSprachtechnologie/easy-to-understand_language [last update: March 16, 2022; last access: July 24, 2024]

the original paper does not include an evaluation or error analysis. Prior to our work, it has also not been used or evaluated in other related work.

6.2.3 DISSIM

Moreover, Niklaus et al. (2019a) propose a framework for discourse-aware simplification by splitting sentences into smaller ones (called DISSIM). The sentences are recursively split by applying hand-crafted grammar rules to the original text. The paper describes that the framework can be applied in English as well as in German; however, it is not clear which of the proposed rules are suitable for German. DISSIM was applied in a few articles on text simplification (see Niklaus et al. 2019c, or Niklaus et al. 2023), but unfortunately, the results are only reported on English text simplification datasets. Hence, I cannot include more insights regarding DISSIM's capabilities on German TS.

6.3 Training Sequence-to-sequence Models – Sockeye

However, rule-based models require a lot of effort, time, and expert knowledge of the simplification process to create good simplification rules. In comparison, data-driven models can directly and simultaneously learn multiple simplification operations from parallel data without relying on explicitly defined rules (Alva-Manchego et al., 2020b). Common neural data-driven TS models rely on an attention-based encoder-decoder architecture (Bahdanau et al., 2015) or its advancement, i.e., a transformer architecture (Vaswani et al., 2017).

Encoder-decoder models are also called sequence-to-sequence models (Jurafsky and Martin, 2024c), they can generate a text based on input texts where both texts can have a different length. Simply put, in the encoder the input sequence (or here the source text) is projected into a continuous and contextualized vector representation. This representation is passed to a decoder which then generates the output sequence (or here the target text) (Jurafsky and Martin, 2024c). The transformer architecture (Vaswani et al., 2017) stands out from other encoder-decoder architectures due to its self-attention mechanism, which enables capturing long-range dependencies more efficiently, facilitating parallel processing of input tokens, and allowing the model to learn contextual relationships within the sequence without recurrent connections (Jurafsky and Martin, 2024a). This design has significantly improved the performance and scalability of NLP tasks compared to traditional recurrent neural network-based architectures.

In order to learn a new task, such as text simplification, on the one hand, a sequence-to-sequence model can be trained from scratch with parallel (or labeled) data such as complex-simple pairs. On the other hand, a model can be first pre-trained on massive unlabeled text data and, afterwards, fine-tuned on a specific task such as text simplification. In this chapter, I will concentrate on the first approach while the next section (see Section 6.4) continues with the latter approach. Training from scratch requires many training samples, but it enables control over which data has been seen during training, e.g., controlling the vocabulary or controlling which bias the model might contain. These kinds of TS model approaches include techniques wrt. finding the best model architecture, its best hyper-parameters (Liu et al., 2023), and a high-quality and huge parallel training dataset to generate the best simplifications.

An example of a Python toolkit that facilitates the training and building of one's own sequence-to-sequence models is the open-source framework called Sockeye (Hieber et al., 2018; Domhan et al., 2020; Hieber et al., 2022). It is especially designed for training machine translation models, and the latest versions also include transformer architectures (Domhan et al., 2020; Hieber et al., 2022).

In recent years, researchers from the computational linguistic department of the University of Zurich have published text simplification models trained with Sockeye, e.g., Sockeye-Benchmarking by Säuberli et al. (2020) (see Subsection 6.3.1) and Sockeye-APA-LHA by Spring et al. (2021) (see Subsection 6.3.2). Ebling et al. (2022) summarizes their work by reporting the results of their best performing model on German sentence simplification, i.e., training a sequence-to-sequence model with a transformer architecture (Vaswani et al., 2017) using the Sockeye framework (Domhan et al., 2020).

6.3.1 Sockeye-benchmarking

Säuberli et al. (2020) experimented with the German News Corpus (see Subsection 4.3.3; target group: German language learners), and Sockeye 1 (Hieber et al., 2018). Säuberli et al. report results of their base Sockeye architecture as well as additional experiments with, e.g., smaller batch sizes (Batch1K) or extension with linguistic features (BASE+LingFeat). They also experimented with data augmentation strategies, i.e., adding non-parallel simplifications (NULL2TRG), adding identical pairs with the simplifications on both sides of the pair (TRG2TRG), and adding pairs including back-and-forth translated simplifications and original simplifications (BT2TRG).

Compared to their base sentence simplification model, adding more sentence pairs decreases their SARI and BLEU scores, except for the TRG2TRG approach, which is their overall best performing system. As far as I know, these are the only experiments with augmented data for German text simplification, where the data are augmented from the original corpus and without additional resources (e.g., as in Anschütz et al. 2023). Unfortunately, the experiments cannot be reproduced as neither the corpus, the models, the code, nor enough details regarding building the models are available. Hence, I cannot give more information regarding this model.

6.3.2 Sockeye-APA-LHA

Spring et al. (2021) decided on a model with 5 layers, 4 attention heads, and 512 hidden units.[5] They have trained their model on the parallel APA-LHA news corpus (see Subsection 4.3.4; target group: German language learners) and the parallel capito corpus (see Subsection 4.1.6; target group: German language learners) with different target language levels, i.e., A1, A2, and B1. For their baseline model, they have trained the model on all language levels at once, but for their best performing model, they add the language level of the target sentence as an indication tag to the source sentence. Additionally, they conducted experiments on pre-training the model on translation from German to English and English to German, or experiments on adding a copy-label to the source if the simplified sentence is identical to the source.

5 For more details on the hyper-parameters I refer to Spring et al. (2021).

Spring et al. have evaluated their models with the automatic evaluation metrics SARI and BLEU on the APA-LHA corpus and the capito corpus. Furthermore, they have analyzed how often their models perform no simplification, i.e., when the model has just copied the source text without any changes. Their baseline model could not be significantly improved with their data augmentation strategies. Spring et al. (2021) have identified the copying behaviors of their models as their major challenge (minimum 70% of copying in all their approaches). However, adding the copy-label has slightly reduced the number of system outputs without any changes (e.g., for CEFR level A2: from 79.73% to 70.59%).

Looking more closely at the capabilities of the models regarding simplification strength: The model trained on APA-LHA achieves higher SARI and BLEU scores when simplifying into CEFR level A2 than B1. Hence, the model seems to be more capable of stronger simplifications. In contrast, the models trained on the capito corpus achieve higher scores on CEFR level B1. Further investigation is required to find reasons for this contrary behavior, e.g., different data quality or unreliable evaluation metrics.

As mentioned above, the same architecture and framework of the model, but in a different version, with less data, and with presumably different hyperparameters, was used to build the German TS benchmark of Säuberli et al. (2020) (see Subsection 6.3.1). The extension of the corpus (in training and test set) or changing the Sockeye settings, has improved the SARI and BLEU scores (comparing BASE+LingFeat of Säuberli et al. 2020 and Sockeye-APA-LHA (also called APA-multi) by Spring et al. 2021).

In conclusion, I assume that Sockeye-APA-LHA have still much room for improvement. When increasing the number of samples (i.e., roughly 10,000 sentence pairs) or the quality of the samples (many misalignments) the results and the quality of training a model from scratch with the Sockeye framework could get better. I assume that the model might not learn well how to simplify on the too small corpus with too many misalignments. As the model checkpoints nor the system outputs are available, a reproduction of Sockeye-APA-LHA[6] (using the code provided by the authors) is necessary to fully understand the strengths and weaknesses of this model. Evaluating the reproduced model on other datasets or evaluating other TS models on the APA-LHA data could also give more insights regarding the capability of Sockeye-APA-LHA. To the best of my knowledge, neither the code nor the data has been used in related work prior to our work.

6.4 Fine-tuning Sequence-to-Sequence Models

As previously discussed, training from scratch requires much labeled data and (depending on the model architecture and training data size) also high computational resources and much time. Hence, this approach may not be well suited for some text simplification purposes with low resources such as simplification of medical or narrative German texts. In addition, an advantage as well as a disadvantage of training language models for text simplification from scratch is that they are limited to the data seen during training. On the one hand, the model might perform very well on the test set of the corpus it has been trained on. But, on the other

6 A reproduction of Sockeye-capito is not possible as the capito is not publicly available.

hand, it might be overfitting to this corpus and might underperform on other data even if from the same domain. Hence, depending on the size and balance of the training data, it is expected that the model will have challenges when simplifying texts from different domains, unseen topics, or unseen simplification operations.

Nevertheless, *transfer learning* can assist in overcoming these challenges. Transfer learning is a technique used in NLP to first acquire knowledge by pre-training a language model on massive (unlabeled) data. Then this knowledge will be transferred to a downstream task when fine-tuning the pre-trained model with new task-specific data (Raffel et al., 2020; Jurafsky and Martin, 2024b). The intuition of this approach is that during the pre-training phase, the model learns, e.g., representations of word meanings and of text structure, which enables the model to more easily adapt to the requirements of another downstream NLP task (Jurafsky and Martin, 2024b) or the same task in another language. Following the transfer learning objective, these pre-trained models require fewer additional training samples to build a good NLP system than when training a model from scratch (Howard and Ruder, 2018; Clark et al., 2018), e.g., depending on the task, a few thousand versus at least ten thousand.

This approach is especially helpful for tasks with low resources such as parallel corpora for German text simplification for specific purposes, such as expert-laypeople simplification. Following this approach, the main engineering part of the models is to find the best pre-trained model, the best hyperparameter, and high-quality training data with a sufficient number of samples.

In the following, I will describe German TS models that are based on fine-tuning of two pre-trained models, i.e., mBART (see Subsection 6.4.1) and mT5 (see Subsection 6.4.2).

6.4.1 mBART

mBART is a sequence-to-sequence model which is pre-trained on data of either 25 (Liu et al., 2020), i.e., mbart-large-cc25[7], or 50 languages (Tang et al., 2021), i.e., mbart-large-50[8], where both models include German. It extends the transformer-based neural network model called BART (Bidirectional and Auto-Regressive Transformers) (Lewis et al., 2020) for multi-lingual text. In addition, mBART is the first multi-lingual encoder-decoder model that has been pre-trained on multi-lingual documents by reconstructing masked tokens in them (also called de-noising) (Liu et al., 2020).

mBART has been developed primarily for sentence and document machine translation (Liu et al., 2020), and therefore it is suitable for short and long text pairs (maximum input and output length of 1024 tokens). In order to specify the language of the source and the target text, a special language ID token is added to both. Given that text simplification is an intra-lingual translation task, mBART appears to be also suitable for text simplification.

For German TS, mBART has been utilized in a few studies, i.e., mBART-20min & mBART-APA+capito (see Subsubsection 6.4.1.1), mBART-SimplePatho (see Subsubsection 6.4.1.2), mBART-GNATS (see Subsubsection 6.4.1.3), and mBART-capito (see Subsubsection 6.4.1.4). In

7 https://github.com/facebookresearch/fairseq/tree/main/examples/mbart [last update: January 28, 2021; last access: July 24, 2024]

8 https://huggingface.co/facebook/mbart-large-50 [last update: March 28, 2024; last access: July 24, 2024]

all studies, the fine-tuning approach of mBART has been slightly changed, e.g., broadening the input and output length or previously adapting the domain, as well as they have been trained and evaluated on completely different data, e.g., documents vs. paragraphs vs. sentences, and news vs. medical vs. narrative vs. web texts. Hence, it is not possible to make a detailed comparison regarding the capabilities of these TS models, e.g., wrt. model settings, domains, or simplification strength. But, combining the findings of all German mBART models, fine-tuning mBART seems to be appropriate for document, paragraph, and sentence simplification on various domains.

6.4.1.1 mBART-20min & mBART-APA+capito

Rios et al. (2021) are the first who have used mBART (Liu et al., 2020) for German document simplification. The main improvements of their approach compared to the standard mBART are to maximize the input length (to 4,096), reduce the vocabulary to the 20,000 most frequent German tokens, and add a special language tag to specify the target language level (de_A1, de_A2, or de_B1). Another adaptation of their mBART approach is the required resources; while mBART is very resource-intensive, the small mBART also works on smaller GPUs by making use of the Longformer attention mechanism (Beltagy et al., 2020). For both models, similar to Spring et al. (2021), they use a language tag to specify the language level of the source and target documents.

Rios et al. have trained both model versions on two datasets, i.e., 20min (see Subsection 4.3.6; domain: news texts; target group: n/a) and APA+capito (see Subsection 4.3.4; domain: news and web texts; target group: German language learners). Interestingly, the small mBART model can outperform the general mBART model on all language-level test sets of the APA+capito corpus when evaluating with ROUGE-L, SARI, and BLEU. In addition, the small mBART model also performs very similarly to the general mBART model on the 20Minuten corpus.

The small mBART-20min model has been already used as a reference TS model in at least one other German TS study, i.e., Anschütz et al. (2023) (see Subsection 6.6.1). In comparison, small mBART achieves a better ROUGE but lower SARI score (ROUGE-L: 19.96; SARI: 33.29) than the other model (ROUGE-L: 17.93; SARI: 42.74). However, small mBART has been trained on ten times more parameters than the other model.

6.4.1.2 mBART-SimplePatho

Trienes et al. (2022) has adapted the approach and the implementation of Rios et al. (2021) for the simplification of clinical notes Compared to previous approaches, Trienes et al. propose to simplify neither documents nor sentences, but paragraphs for their specific simplification purpose and domain, i.e., expert-laypeople simplification of medical texts (see Subsection 4.4.3; target group: laypeople). Similarly to Rios et al. (2021), they also add a language tag to the source and target texts for their mBART model.

For comparison purposes, they report results on the identity baseline, a BERT2BERT model, and a BERT2Share model. For the last two models, they applied BERT (Devlin et al., 2019) as the encoder and decoder of the model, but only in the last one, the encoder and decoder share the weights. Because no comparable model or dataset has previously existed for the simplifica-

tion of German medical texts, they have built their own evaluation protocol by evaluating only on their own corpus with ROUGE-1, ROUGE-2, ROUGE-L, BLEU, and SARI. With regard to all their reported metrics, their mBART model performed better than the BERT models. Unfortunately, due to copyright issues, they were not yet allowed to share their dataset or their model checkpoints.

6.4.1.3 mBART-GNATS

The Longformer mBART approach of Rios et al. (2021) has also been transferred to the simplification of narrative documents. Schomacker et al. (2023a) propose a document simplification model, called mBART-GNATS, following the approach of Rios et al. (2021), but without vocabulary reduction. However, they have changed their approach by first fine-tuning mBART on 60 narrative texts to adapt the domain of the model, and then fine-tuned on 28 complex-simple narrative document pairs (see Section 4.6; target group: mixed).[9]

They found that, overall, mBART performs best without domain adaptation and fine-tuning wrt. BERTScore-F1, ROUGE-L, and BLEU. In their experiments, they achieved the best results with fine-tuning on domain-specific data for just 1 epoch. However, this fine-tuning step has not shown a significant improvement in comparison to omitting the domain adaptation step. Fine-tuning on the non-parallel texts for more than 1 epoch has even worsened the result. Following Schomacker et al. (2023a), on the one hand, these effects might be due to catastrophic forgetting of the model. On the other hand, the tasks the model has been pre-trained on (inter-language translation) and the tasks the model has been fine-tuned on might be too different (intra-language translation).

6.4.1.4 mBART-capito

Recently, mBART has also been used to build a TS system for simplifying complex German documents into texts for non-native speakers (CEFR level: A2) (Säuberli et al., 2024). They also followed the method of Rios et al. (2021) and fine-tuned and evaluated mBART on a part of the capito corpus (see Subsection 4.1.6; domain: web texts; target group: German language learners)[10].

However, the objective of their study is not to evaluate the quality of their model, but to compare different manual evaluation approaches with each other using the generated simplifications of the mBART model. Hence, they do not evaluate their mBART with automatic metrics such as SARI or BLEU, but manually evaluate it with extrinsic approaches (e.g., comprehension tests) and intrinsic approaches (e.g., self-assessment regarding complexity and comprehensibility). They compared the behavior of two groups, i.e., people with intellectual disabilities (target group of German Easy Language) and a control group with student participants, on manually and automatically simplified texts.

9 The code for fine-tuning and domain adaptation is available, but the model checkpoints are not.

10 It remains unclear to which extent the following corpora overlap: the capito corpus used here, the capito+APA corpus used by Rios et al. (2021), and the APA-LHA corpus used by Spring et al. (2021). Therefore, the results of their TS approaches can only be compared to each other with caution. The checkpoints of their model or code for reproduction are not available

Following the results of Säuberli et al., the target group of people with intellectual disabilities can read the automatically simplified texts as fast as manually simplified texts and perceive the automatically generated simplification as more simple than the original or manually simplified texts. In contrast, they could not significantly answer comprehension questions regarding the content better for the automatically simplified texts than for the original texts. Following this, the document simplification of mBART is not simple enough for the target group. More investigation is required to identify the reasons for this, e.g., if the texts still contain too many complex words or too complex structures.

6.4.2 mT5

mT5 (Xue et al., 2021) is another multi-lingual pre-trained encoder-decoder model, which is the extension of the only-English T5 model ("Text-to-Text Transfer Transformer") (Raffel et al., 2020). In contrast to BART, T5 is not limited to only sequence-to-sequence NLP tasks (e.g., machine translation or summarization). Additionally, T5 is capable of other tasks which can be cast into a text-to-text format, e.g., text classification (e.g., sentiment analysis) or sequence labeling (e.g., named entity recognition). This has the advantage that the model has been pre-trained on a combination of all these tasks by using the same data format, and, hence, the same model is capable of approaching many NLP tasks. Further, to tackle a new downstream task, it is sufficient to fine-tune the T5 model with only a few examples or use it for zero-shot experiments since it is intended that T5 can transfer the knowledge of related tasks to new tasks. A new task can be added via fine-tuning with task-specific data and an instructive prefix (e.g., "summarize", "translate to xy").

However, T5 is limited to English only as it has been pre-trained on only English data. In order to extend T5 to multiple languages, I describe two approaches: mT5 by Xue et al. (2021) and flan-T5 by Chung et al. (2024). For mT5, Xue et al. (2021) have slightly adapted the training process of T5 and trained the model with the unlabeled multi-lingual dataset mC4 including 101 languages (also German). Unfortunately, the available version of mT5[11] is not trained on any text-to-text task and yet requires fine-tuning on relevant tasks before applying for downstream tasks. For flan-T5, Chung et al. (2024) have followed the pre-training approach of T5, but have changed the training data: flan-T5 is trained on 60 languages and on overall 1,836 tasks. In contrast to mT5, the available checkpoints for flan-T5[12] are pre-trained on many tasks and, therefore, can directly be used on new downstream tasks without additional fine-tuning.

Text simplification is neither part of the pre-trained tasks of mT5 nor flan-T5, but it can be easily added by applying a new task prefix and task-specific data during fine-tuning. This has been done previously for German TS by Ryan et al. (2023) (mT5-MultiSim; see Subsubsection 6.4.2.1) and Schlippe and Eichinger (2023) (flan-T5-translated; see Subsubsection 6.4.2.2). Both TS models differ regarding their T5 versions as well as regarding their training data (both multi-lingual, but original vs. augmented data). Hence, a comparison of their results would not be reasonable. Overall, mT5-MultiSim and flan-T5-translated both appear suitable for German

11 https://github.com/google-research/multilingual-t5 [last update: December 15, 2022; last access: July 24, 2024]

12 https://github.com/google-research/t5x/blob/main/docs/models.md#flan-t5-checkpoints [last update: July 17, 2024; last access: July 24, 2024]

TS. In the remainder of this Subsection, I describe in more detail the approaches and results of both TS models.

6.4.2.1 mT5-MultiSim

Ryan et al. (2023) have recently fine-tuned mT5 (Xue et al., 2021) on their multi-lingual corpus called MultiSim (domain: mixed; target group: mixed). MultiSim is a multi-lingual corpus for sentence simplification containing overall 653,468 complex-simple pairs for training and evaluation including texts in 12 languages and 8 domains. Their corpus includes a GermanNews corpus (domain: news texts; target groups: German language learners), the GEOlino (see Subsection 4.2.7; domain: knowledge acquisition texts; target group: children), and the TextComplexityDE dataset (see Subsection 4.2.6; domain: Wikipedia-based texts; target group: German language learners). However, as for the two latter, only evaluation data are available. They have re-split the corpora and included a small portion of them into the multi-lingual training set. Therefore, their version of the test sets is different from the one presented in Chapter 4. In total, MultiSim consists of 2,329 German sentence pairs for training. Following Ryan et al. (2023), the amount of data is too small for fine-tuning only on German data or only one of the German corpora; hence, during fine-tuning, they used the German data only in combination with the remaining multi-lingual data.

For evaluation, they report scores per mono-lingual dataset, i.e., for German on the GermanNews corpus, TextComplexityDE and GEOlino, regarding SARI and BLEU scores. They compare their results with two baselines, that is, the identity baseline (source-to-source, no change) and the truncation baseline (removing the last 20% of words from the source text). mT5-MultiSim can outperform both baselines on all three test sets wrt. SARI, but wrt. BLEU only on the GEOlino test set. They further compared the capabilities of their mT5 approach with few- and zero-shot approaches using BLOOM (see Section 6.5); the fine-tuned mT5 significantly outperforms their other approaches on all three German test sets wrt. SARI. The same also holds for the test sets of other languages they have evaluated on, meaning that mT5 seems capable of learning how to simplify in multiple languages.

Comparisons between test sets of different languages do not seem reasonable, as the scores are highly dependent on the quality, the size, and the number of references in the test set. For this kind of evaluation, a translation of one high-quality test set in the languages of interest would be required to have a more fair comparison. Furthermore, due to the re-split of the German test sets, unfortunately, the reported results by Ryan et al. (2023) cannot be compared to other results on the same test set (e.g., see Subsection 6.5.1) as they have different sizes. For comparisons with other German TS models, either mT5-MultiSim has to be reproduced[13] and tested on other German test sets, or other German models have to be evaluated on the shortened test sets of Ryan et al. (2023).

13 Currently, the model checkpoints are not available, but the code is available here https://github.com/XenonMolecule/MultiSim [last update: September 27, 2023; last access: July 24, 2024].

6.4.2.2 FLAN-T5-TRANSLATED

Schlippe and Eichinger (2023) also used a T5 model to train their German sentence simplification model, but they use the multilingual flan-T5 model (Chung et al., 2024). In order to test flan-T5's capabilities regarding simplification (although not being trained on this task), they made use of the prefixes the model has been trained on instead of introducing a new prefix such as "simplify" via fine-tuning. As prefixes, they have selected prefixes of similar tasks, i.e., "translate", "paraphrase", and "summarize". For their corpus, they selected random data from TurkCorpus (Xu et al., 2016) and ASSET (Alva-Manchego et al., 2020b) and translated complex-simple pairs with Google Translate into 40 languages (see Subsection 4.2.4; domain: Wikipedia-based texts; target group: n/a). They have experimented with fine-tuning flan-T5 either with 1,000 pairs of translations of one language of this corpus (e.g., German), or with each 1,000 pairs per each of the 40 languages.

For the setting with 1,000 pairs, their system yields the best results on the German test set wrt. SARI using the instruction of "translation". The same does not hold for the test sets of the other languages, e.g., "paraphrase" yields the best results for English, French, and Portuguese. For the setting of 40,000 pairs, they only report results with the instruction of "paraphrase" which are better than all approaches with just 1,000 pairs across all test sets of all languages. Hence, on the one hand, the more data (even if from another language) the better the simplification wrt. SARI. On the other hand, more analysis on different test sets is required to verify whether "paraphrase" can also yield better results on German test sets. Furthermore, all training and test data have been automatically translated into the languages of interest (except for English), hence, more investigation is necessary to check whether the results also hold for manually simplified gold data. Unfortunately, neither their code, translated data, nor their model checkpoints are available for further studies on their TS approach.

Schlippe and Eichinger further analyzed the ability of their model in another domain by manually evaluating system generations on texts of 5 social science books. They found that their German TS model performs well wrt. fluency, comprehensibility, grammaticality, and simplification (scores higher than 4 on a scale from 1 (worst) to 5 (best)). Furthermore, in this evaluation, flan-T5-translated achieves always better results than their approach using ChatGPT (see Subsection 6.5.5) wrt. simplicity and fluency, but worse wrt. meaning preservation.

6.5 Prompting with Zero- & Few-shot Learning on Auto-regressive Models

In recent years, another common approach of making use of pre-trained language models besides fine-tuning a model on a specific task (see Section 6.4) is to directly infer a generated text (here a generated simplification) from the model by formulating the task as a textual prompt (Liu et al., 2023). A prompt is a structured input for a pre-trained language model to perform the formulated task of the input, it can perform either on an unseen task without any examples (zero-shot learning) or on an introduced task by including a few examples in the prompt (few-shot learning) (Dang et al., 2022). One main advantage of this approach is

that no or only a few complex-simple pairs are required, which facilitates simplification for purposes with lower resources or simplification into languages with fewer resources.

In this line of research, the task is more focused on finding the best formulation of a prompt (prompt engineering) and the best pre-trained model than finding the best training data or best hyperparameter to predict the best simplifications (Liu et al., 2023). Examples for auto-regressive large language models (also called decoder-only models) are, e.g., Gemini (Gemini Team et al., 2024), Llama 2 (Touvron et al., 2023), BLOOM (BigScience Workshop et al., 2023), or ChatGPT (OpenAI, 2024).

In the following, I will present German TS approaches based on multi-lingual transfer learning (see ZEST, Subsection 6.5.1), an unsupervised model (see GUTS, Subsection 6.5.2), prompting BLOOM (see Subsection 6.5.3 and Subsection 6.5.4), and prompting ChatGPT (see Subsection 6.5.5). All models rely on different auto-regressive models, different prompts, and evaluation on different domain and test sets. Hence, I can only present particular results per model and cannot give recommendations for or against one of the TS models. Reasons for that are missing references to previous work in the German TS studies and a lack of a comparable evaluation set-up for these models. However, overall, BLOOM as well as ChatGPT seem to be suitable to simplify German texts to some extent.

6.5.1 ZEST

Mallinson et al. (2020) propose a zero-shot cross-lingual sentence simplification model called *ZEST*. The model can be described as a multi-task multi-lingual and transfer learning model which combines machine translation, text simplification, and next token prediction (language modeling). Put simply, the model jointly learns to translate a sentence from a first language into a second language (using WMT19 Barrault et al. 2019, domain: mixed) and to simplify sentences within the first language (using WikiLarge; see Section 4.2; domain: Wikipedia-based texts; target group: English language learners). They propose that the learned knowledge regarding the simplification of the first language and the translation from the first to the second language can also be transferred to the simplification of sentences in the second language.

This architecture can overcome the scarcity of parallel simplified data, e.g., in a low-resource language or for a low-resource simplification purpose, because for this approach, parallel simplified data is only required in another language, which can be a high-resource language such as English. However, still, some parallel translation data is required from the high-resource to the low-resource language.

They tested their approach with English as a high-resource language and German as a low-resource language and evaluated on two German text simplification datasets, i.e., *GEOlino* (see Subsection 4.2.7; domain: knowledge acquisition texts; target group: children) and *TextcomplexityDE19* (see Subsection 4.2.6; domain: Wikipedia-based texts; target group: German language learners). With respect to automatic scores, i.e., SARI and BLEU, their system achieves similar results to a pivot model which performs the steps of translation and simplification sequentially (i.e., first translation DE-EN, then simplification EN-EN, and finally translation EN-DE). Hence, their assumption of transferring the knowledge can be confirmed. Furthermore, their model outperforms other unsupervised models which have been originally designed for English TS

(i.e., U-SIMP and U-NMT Surya et al. 2019) wrt. BLEU, SARI and FRE-BLEU evaluated on TextComplexityDE and can also cope with the scores evaluated on GEOlino. It is surprising that ZEST achieves such high scores on the data set with simplifications for children, i.e., GEOlino, although the seen complex-simple pairs have been simplified for non-native speakers, i.e., WikiLarge corpus.

In addition to automatic evaluation, they also include manual evaluation (5 ratings per sentence pair by German native speakers) in their evaluation study. Mallinson et al. (2020) show that their model performs better on both test sets than the pivot model and the unsupervised reference model U-SIMP in terms of human judgments regarding grammaticality, meaning preservation, and simplicity. Unfortunately, the human judgments nor the system outputs are available for further investigation. However, the system could be reproduced with their publicly available code[14] and test set[15]. Except from our work, I am only aware that Ryan et al. (2023) and Fruth et al. (2024) have reused the GEOlino or TextComplexityDE data to evaluate their systems. As previously discussed (see Subsubsection 6.4.2.1), the first uses a shortened version of the datasets and the latter adapted TextComplexityDE for paragraph rather than sentence simplification (see Subsection 6.5.2), resulting in no reasonable comparisons.

6.5.2 GUTS

Recently, Fruth et al. (2024) have proposed the first fully unsupervised paragraph simplification model for German, called "GUTS". Their model is based on a similar English model called "Keep it Simple" (Laban et al., 2021): the idea is to use reinforcement learning for rewarding generated simplifications based on evaluation aspects such as simplicity, meaning preservation, and grammaticality. Simply said, a German GPT-2 model (Minixhofer, 2020) generates a few simplification candidates which are then rewarded regarding the evaluation aspects. Based on the resulting reward score, the GPT-2 model is optimized to generate better simplifications.

For evaluation of GUTS, Fruth et al. (2024) compare it to a pivot model based on a modification of ZEST on an adapted version of TextComplexityDE with a focus on paragraphs instead of sentences. In comparison, their pivot model slightly outperforms GUTS wrt. SARI and FRE. However, the model cannot be compared to other models on the TextComplexityDE data, as the authors have modified the dataset for their own purposes. For further investigations, their model checkpoints, system outputs, and evaluation code are available[16].

6.5.3 BLOOM-zero, BLOOM-sim-10, & BLOOM-random-10

Ryan et al. (2023) also experimented with few-shot and zero-shot learning for multi-lingual sentence simplification, but using the auto-regressive language model BLOOM (with 176 billion parameters) (BigScience Workshop et al., 2023). BLOOM is an open-source decoder-only language model based on the transformer architecture and has been pre-trained on data in 46

14 https://github.com/Jmallins/ZEST [last update: September 19, 2021; last access: July 24, 2024]

15 https://github.com/Jmallins/ZEST-data [last update: May 23, 2021; last access: July 24, 2024]

16 https://github.com/LFruth/unsupervised-german-ts [last update: May 7, 2024; last access: July 24, 2024]

languages, officially *not* including German[17]. The model is available in different sizes, e.g., with 176 billion parameters[18] or 7 billion parameters[19].

Even if BLOOM has not been trained on German data, Ryan et al. (2023) have also prompted it with German TS pairs. Their few-shot prompts contain k complex-simple sentence pairs which are prefixed with either "Original: " or "Simple: ". The last two lines of their prompt (in few-shot and zero-shot setting) are the to-be-simplified complex sentence (accompanied with its prefix "Original: ") and just the prefix "Simple: " for the to-be-generated simplification. As examples in their few-shot setting, they used either k random sentence pairs or k pairs in which source sentences are most similar to the to-be-simplified sentence (where $1 \leq k \leq 20$). They measured the similarity between the original sentences using the cosine distance of their representation in sentence embeddings, i.e., LASER (Schwenk and Douze, 2017). For their zero-shot experiments, they just prompt BLOOM with the to-be-simplified complex sentence.

They tested their approach on a German sentence simplification test set, i.e., GermanNews[20], GEOlino (see Subsection 4.2.7; domain: knowledge acquisition texts; target group: children), and TextcomplexityDE19 (see Subsection 4.2.6; domain: Wikipedia-based texts; target group: German language learners). Same as for their mT5 model, they use their own split and size of these datasets. Hence, their results cannot be compared to the results of the ZEST system (Mallinson et al., 2020) and should be interpreted with caution, since their test sets are smaller than the original test sets.

Comparing their few- and zero-shot approaches with BLOOM wrt. SARI (see Table 6.1), the zero-shot approach achieved the lowest scores on all three datasets. The SARI score tends to increase with the k value when the mean is calculated over the three test sets. However, on average over the three datasets, the SARI scores are the highest when $k = 10$. As expected, the examples of similar pairs were more helpful than random examples.

When comparing the BLOOM approaches with the mT5-MultiSim approach of Ryan et al. (2023) (see Table 6.1), mT5-MultiSim can outperform all zero- and few-shot approaches on the GEOLino and TextComplexityDE test sets but not on the GermanNews set. The results in GermanNews might not be as reliable as the results on the other two datasets due to many misalignments in the dataset (see Subsection 4.3.4).

Furthermore, mT5 has been pre-trained on German data contrary to the BLOOM approaches. Hence, it was expectable that mT5-MultiSim achieve higher scores than BLOOM, but the small distance in the scores between the models is surprising. Unfortunately, no manual evaluation has been conducted on the models regarding German simplification to

17 The 46 languages specified by BigScience Workshop et al. (2023) do not include German. But, when looking closer to the resources of the corpus used for pre-training BLOOM (i.e., ROOTS Laurençon et al. 2022), some resources could contain also German texts, e.g., Wikipedia or OPUS-100 Zhang et al. (2020a).

18 `https://huggingface.co/bigscience/bloom` [last update: July 28, 2023; last access: July 24, 2024]

19 `https://huggingface.co/bigscience/bloom-7b1` [last update: January 2, 2024; last access: July 24, 2024]

20 It is unclear whether the GermanNews corpus refer to the news corpus by Säuberli et al. (2020) or APA-LHA by Spring et al. (2022). When comparing the named size of the corpus, it is more likely that APA-LHA was used than the other news corpus.

verify whether BLOOM is indeed capable of generating German texts without being trained on it.

However, I conclude that multi-lingual task-specific data (i.e., multi-lingual complex-simple pairs), and semantically similar data (i.e., complex-simple pairs in which the complex sentence is similar to the to-be-simplified complex sentence) can improve the quality of text simplification outputs.

	GermanNews	TextComplexityDE	GEOLino
mT5-MultiSim	31.58	41.15	50.75
BLOOM-zero	32.48	32.26	29.59
BLOOM-random-5	34.71	36.68	34.5
BLOOM-random-10	35.58	38.07	35.42
BLOOM-random-20	35.53	38.07	34.62
BLOOM-similarity-5	37.79	38.81	39.5
BLOOM-similarity-10	37.69	38.93	39.70
BLOOM-similarity-20	36.76	38.93	39.44

Table 6.1: Comparison of SARI scores of TS approaches by Ryan et al. (2023). Evaluated on GermanNews, a small version of TextComplexityDE, and a small version GEOlino. Best results are highlighted in bold face.

6.5.4 BLOOM-BiSECT

Ponce et al. (2024) also experiment with BLOOM for sentence simplification, but with the version with a smaller version of 7 billion parameters[21]. Further, in contrast to Ryan et al. (2023), Ponce et al. (2024) focus on structural simplification, i.e., split and rephrase. They report results on the German version of BiSECT (see Subsection 4.1.5; domain: web-based texts; target group: n/a), but do not provide enough information to reproduce (e.g., prompt missing, few-shot or zero-shot?) as it is only a small side result of their work.

Looking more closely at their results, their SARI and BLEU scores are very low (roughly 26) which corresponds to worse lexical simplification and meaning preservation. However, the quality of the syntactic simplification (what has been mainly intended) has not been measured with automatic scores, e.g., SAMSA. Furthermore, the few-shot and the evaluation data of BiSECT contain encoding errors that may result in unreliable scores for BERTScore due to not correctly matching word embeddings. Overall, it is not finally revealed whether BLOOM can syntactically simplify texts in a language not seen during pre-training, but the probability is high that other models perform better on the task of German syntactic simplification than BLOOM.

6.5.5 ChatGPT

ChatGPT is a proprietary auto-regressive large language model built by OpenAI (2024). ChatGPT is based on previous models by OpenAI, i.e., GPT-3.5 (Brown et al., 2020) and InstructGPT (Ouyang et al., 2022), and so-called reinforcement learning from human feedback (RLHF). Simply put, in RLHF, first, human-written prompts are used to generate responses from a model

21 https://huggingface.co/bigscience/bloom-7b1 [last update: January 2, 2024; last access: July 24, 2024]

like GPT-3.5. These prompts and their corresponding system-generated responses are then collected as a dataset. Afterwards, human evaluators provide feedback on the generated responses, usually in the form of rankings or ratings, which indicate the quality and appropriateness of each response. Based on this human feedback, a reward model is trained, which learns to assign a numerical score to each response based on its perceived quality. Finally, the model is fine-tuned using reinforcement learning, with the goal of maximizing the expected reward (i.e., high human approval) for responses generated by the model. This process encourages the model to produce texts that are more likely to receive high ratings from humans. Over time, as the model generates more responses and receives more feedback, it can learn to generalize to new prompts and produce high-quality responses that align with human preferences (Ouyang et al., 2022).

In recent years, ChatGPT has gained much attention in society due to its high capabilities of generating texts. With respect to NLP, Qin et al. have approved in 2023 that the then current ChatGPT version[22] has been very good in general generative tasks, e.g., dialogue or reasoning tasks, but struggles with tasks like summarization due to too long outputs.

In order to analyze the capacity of ChatGPT (OpenAI, 2024) regarding German text simplification, there are a few studies, e.g., Schlippe and Eichinger (2023), or Deilen et al. (2023). I will briefly introduce their approaches in the following, but without further reproduction, we cannot compare these approaches with each other as they are evaluated with different methods and on different data.

Schlippe and Eichinger (2023) have prompted ChatGPT (version not specified) with "Vereinfache den folgenden Satz: " (EN: "Simplify the following sentence: ") using their translated test data (see Subsection 4.2.4; domain: Wikipedia-based texts; target group: n/a). The target group or the simplification purpose has not been specified.
In comparison to their fine-tuned flan-T5-translated model, ChatGPT performed worse on their data wrt. fluency, comprehensibility, grammaticality, and simplification. But, ChatGPT achieved a better result regarding meaning preservation. Unfortunately, they do not provide more insights regarding the system's quality, e.g., by providing more fine-grained or automatic evaluation. Furthermore, their results cannot be reproduced as neither the model version nor data, their data, the generated system outputs, nor the human judgments are made available.

Deilen et al. (2023) have experimented with ChatGPT for German document simplification into German Easy Language by using a holistic and a linguistic approach. In the holistic approach, the translation is enforced all at once, whereas in the linguistic approach the translation is requested step-by-step on text-level, sentence-level, and word-level. They manually evaluated the system-generated simplifications wrt. correctness, readability, and syntactic complexity on public authority texts (domain: web texts; language variety: German Easy Language). They found that 37.5% of the generated simplifications were not correct and only a few syntactic constructions have been simplified. However, following the Hohenheim Comprehensibility Index (HIX) measured with the proprietary tool TextLab[23] the system-generated texts are simpler than the source texts, but do not reach the readability level of German Easy Language. Un-

22 Unfortunately, the version of ChatGPT or the date of the experiments has not been specified.
23 https://www.comlab-ulm.de/textlab-sprachsoftware/ [last access: July 24, 2024]

fortunately, Deilen et al. (2023) have neither reported the version of ChatGPT, nor the prompts that they have used for their experiments. Further, no automatic evaluation scores are available, hence, their results are not comparable to other approaches, nor reproducible.

6.6 (Fine-tuning) Auto-regressive Language Models

In addition to utilizing the learned aspects of a language model, via few-shot learning or checking the out-of-the-box capability of auto-regressive language models (as described in the previous section), there are attempts on fine-tuning and few-shot learning on auto-regressive models (Anschütz et al., 2023) (see Subsection 6.6.1) and combining auto-regressive models with sequence-to-sequence models (Klöser et al., 2024) (see Subsection 6.6.2).

6.6.1 customer-decoder-ats

The approach by Anschütz et al. (2023) expands the approach of Rios et al. (2021): They use their small mBART model as the decoder of their document simplification model. As encoder they fine-tuned German auto-regressive models, also called generative pre-trained transformer models (GPT), e.g., GerPT2 (Minixhofer, 2020), or German_GPT (Schweter, 2020), on roughly 550,000 manually simplified German web texts (language varieties: German Plain and Easy Language; target group: mixed). After combining the GPT-encoder and the mBART decoder, they finally trained the encoder-decoder cross-attention. Therefore, their model has changed many fewer parameters than the mBART model of Rios et al. (2021) and, hence, is more resource-efficient.

Anschütz et al. (2023) first manually evaluated the grammaticality of the GPT-models; they found that German_GPT generates sentences that are more grammatically wrong than the original sentences, whereas GerPT2 contains fewer grammatical errors than the original sentences. Furthermore, they automatically evaluated their approach on the 20minuten dataset (see Subsection 4.1.3; domain: news texts; target group: n/a) and showed that they can significantly outperform the small mBART model by (Rios et al., 2021) with respect to SARI (increased by roughly 9 points, see Table 6.2) and can keep up with their BLEU and ROUGE-L scores (variance of less than 2 points, see Table 6.2). Assuming that both have been evaluated with the same test split and the same implementation of the metrics, the extension of the TS model with a fine-tuned GPT model seems to improve the simplification quality. We can also infer that additional monolingual simplified data (see Subsection 4.7.3) can help to enhance TS systems' outputs.

Further, this approach facilitates text simplification in at least two ways; first, the model uses less computing power than comparable models, and second, fewer parallel complex-simple pairs are required. In more detail, the approach facilitates text simplification experiments with a few parallel texts, e.g., for a specific domain or a low-resource language, and with low computational resources.

	BLEU	SARI	ROUGE-L
mBART-20min	**7.47**	33.29	**21.62**
small mBART-20min	6.29	33.29	19.96
custom-decoder-ats-gerpt2	4.95	42.25	18.52
custom-decoder-ats-german-gpt	4.80	**42.74**	17.93

Table 6.2: Comparison of BLEU, SARI, and ROUGE-L scores of TS approaches by Rios et al. (2021) and Anschütz et al. (2023). Evaluated on 20min. Best results are highlighted in bold face.

6.6.2 GPT-2 & LeoLM

In contrast to the TS approaches with auto-regressive models described above, Klöser et al. (2024) fine-tuned their document simplification models on semi-synthetic complex-simple pairs before testing. Their semi-synthetic data contains the same (and more) simplifications as the training data of Anschütz et al. (2023). However, Klöser et al. (2024) augmented them to parallel data by adding automatic translations of simplified texts as source texts (see Subsection 4.1.8; domain: web texts; language varieties: German Plain and Easy Language; target group: mixed). For their experiments, they selected four German auto-regressive models, i.e., a small and a huge GPT-2-wechsel model (Minixhofer et al., 2022) and a small and a huge LeoLM model (Plüster, 2023) and four decoding algorithms, i.e., greedy, beam search, sampling-based, and contrastive search approach.

In summary, their results show that the larger models outperform the smaller models, LeoLM models outperform the GPT-2-wechsel models, and the beam search performs best on three of the four models wrt. SARI, BLEU, and METEOR. Furthermore, they manually evaluated the meaning preservation (or content similarity) of the best GPT-2 and best LeoLM model. (Klöser et al., 2024) come to the conclusion that the LeoLM preserves much more of the original meaning (2.68, where 0 is worst and 3 is best) than the GPT-2 model (meaning preservation: 1.34).

The reported results of this approach can unfortunately not be well compared with the approach of Anschütz et al. (2023) because both use different auto-regressive models and evaluate on different data. Further, this evaluation has been conducted with the augmented source texts. Hence, more investigation is required to get results on standard German test sets and to make comparisons to other German TS models.

6.7 Proprietary TS Models & Real World Application

In recent years, outside of academia, some companies have also focused on building German text simplification systems. For a more complete overview of German text simplification models, in this section, I also briefly introduce proprietary TS models (see Subsection 6.7.1). Furthermore, I briefly show the usage of TS systems outside the scientific context in the form of real-world applications (see Subsection 6.7.2).

6.7.1 Proprietary TS Models

In the following, I list the names of the companies and their TS models and add information about them which is publicly available (with no claim for completeness). Unfortunately, I can-

not provide more information because no or fewer details are available regarding the used TS model approach or model architecture, as it is part of their company secret.

I am currently aware of the following proprietary TS models:

- *SUMM-AI*[24]: simplification into German Easy Language; computer-aided translation system; post-editing by trained translators wanted,
- *Leichte Sprache.io*[25]: simplification into German Easy Language,
- *Klartext St. Pauli*[26]: simplification into Plain German,
- *capito digital*[27]: simplification into texts with CEFR level A1, A2 or B1,
- *Wortliga Plain*[28]: simplification into Plain German (CEFR level B1 and B2),
- *ChatGPT-Bot "Klar und verständlich"*[29]: simplification into Plain German; promting with ChatGPT.
- *LanguageTool*[30]: simplification into Plain Language (including English, German, French, Dutch, Spanish, and Portuguese)

Furthermore, there are a few approaches regarding specializing the proprietary model ChatGPT for German Plain Language simplification by adding personalized instructions or additional data to the model, e.g., "GPT-Bot Klar und Verständlich (K&V)"[31], "Verständlich"[32], "KI-Lektorat"[33] (Manning, 2024).

To the best of my knowledge, these models have not yet been scientifically evaluated. However, for a non-scientific but extensive comparison of some of these models and their simplification capabilities, I refer to the blog series "KI-Tools für Einfache Sprache" (EN: "AI-Tools for Plain German")[34] by Manning (2023) including feedback by .

6.7.2 Use Cases of Text Simplification in Real World Applications

The previously named proprietary TS models are already used in real-world applications. In the following, I describe some use cases that I am aware of (with no claim for completeness). In all of these examples, the usage of an automatic text simplification system is explicitly mentioned, and the texts are equipped with information that the simplified text is not legally binding.

The SUMM-AI TS model, for example, is used to automatically generate simplified versions of public authority texts, especially press releases, for the city of Hamburg[35] and the city of Aschaffenburg[36]. While the generated simplifications for the city of Hamburg are always post-edited before publication (Stadt Hamburg, 2024), this procedure is not explicitly mentioned for

24 https://summ-ai.com/ [last access: July 24, 2024]
25 https://leichte-sprache.io/ [last access: July 24, 2024]
26 https://einfachesprache.xyz/ [last access: July 24, 2024]
27 https://digital.capito.eu/ [last access: July 24, 2024]
28 https://wortliga.de/textanalyse/ [last access: July 24, 2024]
29 https://chat.openai.com/g/g-eKpTIfwJi-klar-und-verstandlich [last access: July 24, 2024]
30 https://languagetool.org/de/text-umformulieren [last access: July 24, 2024]
31 https://chatgpt.com/g/g-eKpTIfwJi-klar-und-verstandlich-k-v [last access: July 24, 2024]
32 https://chatgpt.com/g/g-HTx94m62K-verstandlich [last access: July 24, 2024]
33 https://chat.openai.com/g/g-X4PVHTKOY-ki-lektorat [last access: July 24, 2024]
34 https://multisprech.org/2023/07/26/ki-tools-fuer-einfache-sprache-1-st-pauli-im-test/ [last access: July 24, 2024]
35 https://www.hamburg.de/barrierefrei/leichte-sprache/[last access: July 24, 2024]

the city of Aschaffenburg (Stadt Aschaffenburg, 2024). For both web pages, the texts are also not proofread by the target group prior to publication, although required in many German Easy Language guidelines (e.g., Bredel and Maaß 2016, Netzwerk Leichte Sprache 2022, or Deutsches Institut für Normung (DIN) 2023).

Similarly, a huge German telecommunications company (called "Telekom")[37] and an established German health magazine (called "Apotheken Umschau")[38] are also automatically simplifying their web pages with the SUMM-AI TS tool and doing additional post-editing (Deilen et al., 2024; Wort & Bild Verlag, 2024). For a small evaluation study of the capabilities of SUMM AI's TS tool regarding health texts, I refer the interested reader to Deilen et al. (2024).

The Klartext St. Pauli TS tool is used to automatically simplify the web page of the German football club FC St. Pauli. A random sample of the web pages is also proofread by the target group (Fußball-Club St. Pauli, 2023).

I am not aware of concrete use cases of the other named proprietary TS systems.

6.8 Summary & Outlook

In summary, in this section, I have presented the current state-of-the-art models for German text simplification, i.e., 11 document simplification systems, 2 paragraph simplification systems, and 13 sentence simplification systems. I introduced five categories of system approaches, i.e., rules-based models, training transformer-based models from scratch, fine-tuning sequence-to-sequence models, prompting auto-regressive models, and fine-tuning auto-regressive models. For a concise overview, the TS systems and their main metadata are summarized in Table 6.3.

Overall, at present, it is difficult to determine the extent of the progress and development of German TS. On the one hand, this status is due to the fast pace and many independent research efforts in the field; often previous research is unknown or not referred to, hence, there are only a few comparisons between German TS systems. On the other hand, German TS is still a niche topic especially with respect to many possible target groups and text domains that too few models and data exist per simplification purpose for a meaningful comparison.

6.8.1 Challenges & Research Gaps

In the following, I will summarize and discuss research gaps which I have identified considering the current state-of-the-art of text simplification models for German, i.e., gaps in current approaches (see Subsubsection 6.8.1.1), mixing domains and target groups (see Subsubsection 6.8.1.2), comparability (see Subsubsection 6.8.1.3) and reproducibility of TS models (see Subsubsection 6.8.1.4), size of the training data (see Subsubsection 6.8.1.5), and target group empowerment.

These challenges again emphasize the importance of answering my research questions regarding the connection between document and sentence simplification (see RQ 6-1), the rele-

36 https://www.aschaffenburg.de/Buerger-in-Aschaffenburg/Buergerservice/Leichte-Sprache/DE_index_3322.html [last access: July 24, 2024]

37 https://www.telekom.com/de/leichte-sprache [last access: July 24, 2024]

38 https://www.apotheken-umschau.de/einfache-sprache/ [last access: July 24, 2024]

System Name	Reference	Type	Level	Domain	Target Group / Language Variety	Training Data	# Simp. Pairs	URL
-	Suter et al. (2016)	rule-based	sent	*n/a*	*n/a*	-	-	-
HDA-ETR	Siegel et al. (2019)	rule-based	sent	*n/a*	German Easy Language	-	-	github.com/hdaSprachtechnologie/easy-to-understand_language
DISSIM	Niklaus et al. (2019a)	rule-based	sent	*n/a*	*n/a*	-	-	github.com/Lambda-3/DiscourseSimplification
Sockeye-Benchmarking	Säuberli et al. (2020)	seq2seq	sent	mixed	German language learners	APA-LHA OR-B1	3,316	-
Sockeye-APA-LHA	Spring et al. (2021) & Ebling et al. (2022)	seq2seq	sent	news	German language learners	APA-LHA OR-A2 & APA-LHA OR-B1	9,456 & 10,268	github.com/ZurichNLP/RANLP2021-German-ATS
mBART-20min	Rios et al. (2021)	fine-tuned seq2seq	doc	news	*n/a*	20Minuten	17,905	github.com/a-rios/longmbart
mBART-APA+capito	Rios et al. (2021)	fine-tuned seq2seq	doc	mixed	German language learners	capito-A1-A2-B1 & APA-LHA OR-A2 & APA-LHA OR-B1	3,424 & 2,250 & 2,302	github.com/a-rios/longmbart
mBART-SimplePatho	Trienes et al. (2022)	fine-tuned seq2seq	par	medical	laypeople	Simple-Patho	3,280	github.com/jantrienes/simple-patho
mBART-capito	Säuberli et al. (2024)	fine-tuned seq2seq	doc	mixed	German language learners	Capito Corpus	*n/a*	-
mBART-GNATS	Schomacker et al. (2023a)	fine-tuned seq2seq	doc	narration	Plain German & German Easy Language	GNATS	28	github.com/tschomacker/aligned-narrative-documents
mT5-MultiSim	Ryan et al. (2023)	fine-tuned seq2seq	sent	mixed	mixed	MultiSim	653,468	github.com/XenonMolecule/MultiSim
flan-T5-translated	Schlippe and Eichinger (2023)	fine-tuned seq2seq	sent	wikipedia	*n/a*	translated ASSET	1,000	-
ZEST	Mallinson et al. (2020)	zero-shot	sent	wikipedia	*n/a*	WikiAuto & WMT19	300*k* & 6.0*mio*	github.com/Jmallins/ZEST
GUTS	Fruth et al. (2024)	unsupervised	par	-	-	-	-	github.com/LFruth/unsupervised-german-ts
BLOOM-zero	Ryan et al. (2023)	prompting	sent	-	-	-	-	github.com/XenonMolecule/MultiSim
BLOOM-sim-10	Ryan et al. (2023)	prompting	sent	news & knowledge acquisition & wikipedia	mixed	GermanNews & TextComplexityDE & GEOlino	10	github.com/XenonMolecule/MultiSim
BLOOM-random 10	Ryan et al. (2023)	prompting	sent	news & knowledge acquisition & wikipedia	mixed	GermanNews & TextComplexityDE & GEOlino	10	github.com/XenonMolecule/MultiSim
BLOOM-BiSECT	Ponce et al. (2024)	prompting	sent	-	-	*n/a*	-	-
ChatGPT-multilingual	Schlippe and Eichinger (2023)	prompting	sent	-	-	-	-	-
-	Deilen et al. (2023)	prompting	doc	-	German Easy Language	-	-	-

Table 6.3: Summary of German TS models (without own work). Each line separates different model approaches. Extended version of Stodden (2024b). All URLs have lastly been accessed at July 24, 2024. Part I (continued on next page).

System Name	Reference	Type	Level	Domain	Target Group / Language Variety	Training Data	# Simp. Pairs	URL
custom-decoder-ats-gerpt2	Anschütz et al. (2023)	AR model + fine-tuned seq2seq	doc	mixed & news	Plain German & German Easy Language & *n/a*	Simplified, monolingual German data & 20Minuten	544,467 & 17,905	github.com/MiriUll/Language-Models-German-Simplification
custom-decoder-ats-german-gpt	Anschütz et al. (2023)	AR model + fine-tuned seq2seq	doc	mixed & news	mixed	Simplified, monolingual German data & 20Minuten	544,467 & 17,905	github.com/MiriUll/Language-Models-German-Simplification
gpt2-wechsel-german	Klöser et al. (2024)	fine-tuned AR model	doc	mixed	Plain German & German Easy Language	Semi-synthetic Simple German Web Corpus	7,130	github.com/MSLars/German-Text-Simplification
gpt2-xl-wechsel-german	Klöser et al. (2024)	fine-tuned AR model	doc	mixed	Plain German & German Easy Language	Semi-synthetic Simple German Web Corpus	7,130	github.com/MSLars/German-Text-Simplification
leo-hessianai-7b	Klöser et al. (2024)	fine-tuned AR model	doc	mixed	Plain German & German Easy Language	Semi-synthetic Simple German Web Corpus	7,130	github.com/MSLars/German-Text-Simplification
leo-hessianai-13b	Klöser et al. (2024)	fine-tuned AR model	doc	mixed	Plain German & German Easy Language	Semi-synthetic Simple German Web Corpus	7,130	github.com/MSLars/German-Text-Simplification

Table 6.4: Summary of German TS models (without own work). Each line separates different model approaches. Extended version of Stodden (2024b). All URLs have lastly been accessed at July 24, 2024. Part II (continued from previous page).

vance of model architecture and quality of training data (see RQ 6-2), and the effect of TS on real-world applications (see RQ 6-3).

6.8.1.1 System Level & Approaches

The document simplification systems are either following a fine-tuning approach on sequence-to-sequence models (4 out of 11), here only mBART, fine-tuned on auto-regressive models (6 out of 11), here GPT versions and LeoLMs, or prompting ChatGPT (1 of 11). However, even if these approaches show good results at the document level, their capability on sentence simplification has not yet been investigated. Furthermore, the capabilities of the same TS approach have not yet been investigated on both levels simultaneously, i.e., document and sentence simplification (further called Model Challenge B).

In comparison, sentence simplification systems are rule-based (3 out of 13), trained from scratch (2 out of 13), fine-tuned mT5 models (2 out of 13), or prompting approaches on auto-regressive models (6 out of 13). However, except for mT5-MultiSim and Sockeye-APA-LHA, the approaches do not include gold parallel training data: Instead, they rely on additional data for transfer learning (translation data and English simplification data) or use just a few samples in few-shot learning. This might be due to missing high-quality training corpora and the current trend of prompting auto-regressive models.

Nevertheless, as shown by Klöser et al. (2024), (high-quality) simplification pairs could also be used to fine-tune and improve auto-regressive models for text simplification. High-quality sentence pairs, e.g., manually aligned pairs, could also improve Sockeye-APA-LHA because currently it is trained only on error-prone automatically aligned sentence pairs.

6.8.1.2 Domains & Target Groups

Many of the TS models are trained on texts of mixed domains (10 out of 26) and texts written for a mix of different target groups (10 out of 26), due to a lack of large datasets focusing especially on one special simplification purpose (further called Model Challenge C). Following Gooding (2022), this mix is generating a "homogeneity effect" where one simplification should fit all people of all target groups at the same time, which is impossible due to the different skills and needs of the readers.

I, therefore, recommend building larger, high-quality datasets, which focus on only one language level or target group, or a few simplification operations which could then be used for more targeted automatic simplification. An alternative to new datasets could be to experiment with data augmentation strategies in which the additional data should focus on the same target group or the same simplification operations, for example, adding monolingual data as in Anschütz et al. (2023) or adding target-to-target pairs to the training data as in Säuberli et al. (2020).

6.8.1.3 Comparability

In most TS studies, the systems are evaluated on a new test dataset disregarding existing evaluation datasets of the same domain or same language variety: For example, ChatGPT-multilingual

could have been evaluated on gold simplifications of TextcomplexityDE (Naderi et al., 2019) (see Subsection 4.2.6) in addition to the translated ASSET version, or the systems by Klöser et al. (2024) could have also been evaluated on gold test data of the web domain, e.g., Simple German Corpus '23 (Toborek et al., 2023) (see Subsection 4.1.4) or DEplain-web (see Part II Subsection 7.4.2).

Following this, no benchmark for German text simplification systems exists because the majority of models are not comparable due to different test data or different evaluation methods (further called Model Challenge D). I was only able to compare TS systems which have been published in the same paper (e.g., mT5-MultiSim vs. prompting on BLOOM) except for custom-decoder-ats (Anschütz et al., 2023) (see Subsection 6.6.2) and GUTS Fruth et al. (2024) (see Subsection 6.5.2) which refer and compare themselves to results of other papers.

Therefore, it remains an open question, which TS model performs best for a domain or for a language variety. Hence, a reproduction study is required for better comparison of the models (considering non-mixing of target groups and domains).

6.8.1.4 Reproducability

Unfortunately, I have also revealed some new issues regarding models' reproduction, and have confirmed previously named problems with respect to the training data and the evaluation process (further called Model Challenge E). I found the following main three issues with the models, i.e.,

1. impossibility of reproduction, e.g., due to missing details, missing code, not-available or restricted-access data, or restricted-access language models (e.g., see Niklaus et al. 2019a, or Schlippe and Eichinger 2023),
2. differences in reproduction and, therefore, less comparison, e.g., due to different data splits (e.g., see Spring et al. 2021 or Ryan et al. 2023), and
3. differences in evaluation scores for reported scores and scores of reproduced models, due to different system outputs or different implementations of metrics (e.g., see Fruth et al. 2024).

For a better reproducibility and better comparison between ATS models, a system description paper should publish as much detail and materials related to the models as possible with respect to copyright and licenses, e.g., publishing

1. the checkpoints of the trained or fine-tuned models and code on how to reuse them, or
2. the code and a description of how to rebuild and re-train the model, including model versions and used prompts.

Additionally, I recommend publishing the system generations (if not restricted by copyright) to enable further analysis of the results (e.g., see Fruth et al. 2024). Due to limited computing resources, system generations cannot always be reproduced, even if the code or the model is provided. I argue that system generations are helpful for better understanding the original work, but they can also be valuable for building better evaluation metrics. A data repository

in which various German simplification test sets and corresponding system outputs are stored would be helpful to tackle this issue.

6.8.1.5 Size & Quality & Post-Editing

The systems also differ distinctly in their size of the documents or sentence pairs used for training or fine-tuning (further called Model Challenge F), while for the prompting approaches only up to 20 sentence pairs have been used, mBART-20min is fine-tuned on roughly 18,000 document pairs (see Table 6.3), Sockeye-APA-LHA trained on roughly 10,000 sentence pairs, and mT5-MultiSim fine-tuned on roughly 653,000 sentence pairs. Following most TS system reports (except Schomacker et al. 2023a), I can conclude that the more data a model is trained or fine-tuned on, the better the simplification quality evaluated (e.g., see Anschütz et al. 2023 or Ryan et al. 2023).

However, even if good results on a leaderboard are achieved, the quality of current TS approaches in research is not ready for their use in production with the target group (Garbacea et al., 2021) (Model Challenge F). In the current state, the system-generated simplifications obligatorily require professional post-editing by trained translators (e.g., see Deilen et al. 2024).[39]

6.8.1.6 Target Group Empowerment

Additionally, more research is required which includes the target group in the process of simplification (further called Model Challenge G). A future direction could be personalized simplification, as already analyzed in Danish TS research (Bingel et al., 2018).

Further, the effects of automatically simplified texts on the target group are underresearched in the scope of automatic German text simplification. For example, how well can automatically simplified texts currently be comprehended by the target group (e.g., see Säuberli et al. 2024)? Would people like to only read the complex, the simplified version of a text, or both in parallel (e.g., see Vollenwyder et al. 2018)? Would people of the target group be more likely to visit web pages with automatically simplified texts (acceptance of automatic simplifications)? Which parts of a text would a target group like to see simplified (target group empowerment)?

6.8.2 Outlook

In this section, I have introduced the last part of the workflow of German automatic text simplification – German TS models. I also discussed the challenges and research gaps on this topic considering information from previous parts of the workflow.

With this section, I close Part I: the overview of the state-of-the-art in German text simplification regarding each component of the simplification workflow including knowledge regarding simplicity (see Chapter 2), building corpora (see Chapter 3), reusing existing corpora (see

39 I do not go into more detail regarding post-editing or usage of TS models outside of academia because this would be out of scope of this PhD thesis. However, I refer the interested reader to Deilen et al. (2024) for a scientific evaluation of SUMM-AI's TS system and recommendations regarding post-editing.

Chapter 4), evaluating (see Chapter 5), and building as well as using models for German text simplification (see current chapter, i.e., Chapter 6).

In the next part, Part II, I present my contributions of this PhD thesis to tackle the previously introduced issues. Then, finally, in Part III, I explain how I have addressed the mentioned challenges and research gaps with respect to all parts of the text simplification workflow.

Part II

Publications

Chapter 7

Publications

7.1 Overview of the Chapter

In this part of my PhD thesis, I include all my published scientific articles that are related to text simplification. The publications are first ordered by their relation to text simplification and then in chronological order (starting with the oldest publication). An overview of all publications linked to their component of the text simplification workflow is presented in Figure 7.1.

Some of the paper refers to more than one TS topic: in this case, I add it to the section with the highest thematic overlap. For each publication, I have also added a short summary and a link to access the full-text version of the paper.[1] For information on supplementary material (e.g., code, poster, or presentation slides) per paper, see Appendix.

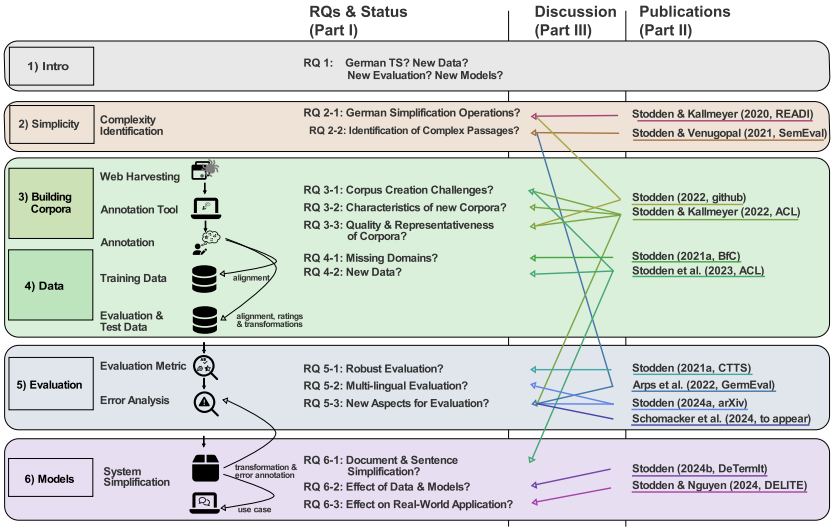


Figure 7.1: Contributions of this thesis including chapters and publications (same as Figure 1.2).

1 It is important to note that all publications are openly licensed and accessible. However, some of my works are published under a slightly more restrictive licence that only allows usage for non-commercial purposes. Consequently, it was not feasible to include all papers in this golden open access version. As we prioritised (open) accessibility over completeness, we have not included the publications here, but have ensured access to all full versions via the provided links and QR codes.

7.2 Complexity & Simplification

In this chapter, I am presenting my publications that mainly address complexity and simplification, i.e.,

- Regina Stodden and Laura Kallmeyer. 2020. A multi-lingual and cross-domain analysis of features for text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties* (*READI*), pages 77–84, Marseille, France. European Language Resources Association (see Subsection 7.2.1), and
- Regina Stodden and Gayatri Venugopal. 2021. RS_GV at SemEval-2021 task 1: Sense relative lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation* (*SemEval-2021*), pages 640–649, Online. Association for Computational Linguistics (see Subsection 7.2.2).

7.2.1 A multi-lingual and cross-domain analysis of features for text simplification.

Reference: Regina Stodden and Laura Kallmeyer. 2020. A multi-lingual and cross-domain analysis of features for text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 77–84, Marseille, France. European Language Resources Association

URL: https://aclanthology.org/2020.readi-1.12/

Summary: This text discusses research on readability and text simplification features that can help to identify sentences in simplified language and transform complex texts into simplified ones. The paper addresses three research questions concerning the differences between complex and simplified texts, the consistency of the simplification process across domains and languages, and the potential use of language-independent features to explain the simplification process.

The study analyzes the relevance of 104 text simplification features (e.g., word and sentence length features, syntactic features, lexical features, and features wrt. proportion of POS tags) across five languages (i.e., Czech, German, English, Spanish, and Italian) and three domains (i.e., web, Wikipedia, and news).

We have identified some features (e.g., readability, parse tree height, or proportion of verbs) that can explain the simplification process and help to distinguish complex vs. simple sentences in certain corpora. But we are also highlighting the need for further exploration of other features like morphological or grammatical features. The study further supports the theory of consistent text simplification across languages, suggesting a higher focus on multi-lingual text simplification approaches.

7.2.2 RS_GV at SemEval 2021 Task 1: Sense Relative Lexical Complexity Prediction

Reference: Regina Stodden and Gayatri Venugopal. 2021. RS_GV at SemEval-2021 task 1: Sense relative lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation* (*SemEval-2021*), pages 640–649, Online. Association for Computational Linguistics

URL: https://doi.org/10.18653/v1/2021.semeval-1.82

Summary: In this work, we present a system called RS_GV to predict the lexical complexity of a word within an English sentence. RS_GV is a neural network that uses a mix of hand-crafted linguistic features, contextualized character embeddings, and a sense-relative normalization technique to predict the complexity of target words.

Our motivation for the sense relative approach is that words can have more than one sense and that depending on the sense the lexical complexity can be higher or lower. For example, the word "*vision*" has five senses (e.g., "ability to see", "supernatural experience", and "foresight") from which the meaning "ability to see" is the easiest to understand. However, the senses per word are not considered in most of the features, e.g., in all frequencies of all senses of the word are summed to get the frequency feature. To consider this imbalance, we have normalized the features based on the number of senses of the target word in WordNet.

Comparing our conducted approaches, normalizing handcrafted features using WordNet senses gives better results than using a min-max normalization. Using contextualized character embeddings improves the model's performance compared to other approaches (e.g., non-contextualized or contextualized word embeddings). In addition, training only on data from one domain is better than training on all domains. Our best approach works well for predicting the complexity of terms in biomedical texts but struggles with the Bible and political texts. It also has issues with predicting complexity values at the end of the scale (i.e., very complex and very simple).

7.3 Building Text Simplification Corpora

In this Chapter, I am presenting my publications that mainly address building text simplification corpora, i.e.,

- Regina Stodden. 2022. Creation of a parallel simplification corpus – Using the annotation tool TS-anno. Annotation guideline, Heinrich Heine University, Düsseldorf, Germany. Also available in German (see Subsection 7.3.1), and
- Regina Stodden and Laura Kallmeyer. 2022. TS-ANNO: An annotation tool to build, annotate and evaluate text simplification corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 145–155, Dublin, Ireland. Association for Computational Linguistics (see Subsection 7.3.2).

7.3.1 Creation of a parallel simplification corpus – Using the annotation tool TS-anno

Reference: Regina Stodden. 2022. Creation of a parallel simplification corpus – Using the annotation tool TS-anno. Annotation guideline, Heinrich Heine University, Düsseldorf, Germany. Also available in German

URL: https://github.com/rstodden/TS_annotation_tool/blob/master/annotation_schema/Annotation-guideline_TS-anno-EN.pdf

Summary: This work is the technical report of the text simplification annotation tool TS-ANNO (see Stodden and Kallmeyer (2022) in Part II Subsection 7.3.2). Therefore, it includes user guidance on how to navigate through the tool, as well as how to technically perform an annotation, e.g., aligning a sentence pair, rating a sentence pair, and labeling passages with simplification operations.

This paper also includes an annotation schema regarding the manual alignment, annotation of simplification operations, and evaluation aspects. For manual alignment, we have specified to annotate 1:1 (e.g., rephrasing or reordering within a sentence), 1:*n* (e.g., splitting a sentence), *m*:1 complex-simple pairs (e.g., merging several sentences) as well as *n*:*m* sentence pairs (e.g., fusion of several sentences). Another specification of our annotation is that word explanations are not aligned with the sentence which includes the complex and explained terms.

Regarding the evaluation aspects, this work provides a list of all aspects, including a statement (also shown in TS-ANNO). Furthermore, each aspect is explicitly described, including annotation recommendations in critical situations, as well as recommendations when a simplification has an effect on more than one evaluation aspect.

The guideline of simplification operations also contains the first typology of simplification operations for German, i.e., overall 56 operations in 9 transformation classes and on 4 levels (i.e., word, phrase, sentence, and paragraph level). Each operation is described in many details to ensure a high overlap in understanding and annotation of the operations.

7.3.2 TS-ANNO: An annotation tool to build, annotate and evaluate text simplification corpora

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

URL: https://doi.org/10.18653/v1/2022.acl-demo.14

Summary: This paper introduces TS-ANNO, an open-source web application designed for the creation and evaluation of parallel corpora in text simplification. It aims to facilitate the manual creation of high-quality TS corpora by supporting web harvesting of parallel documents, manual simplification of complex documents, and manual alignment of parallel documents into *n*:*m* sentence pairs. It allows users to upload plain, paragraph-segmented, or pre-aligned texts, and supports manual simplification if a simple version is not available.

TS-ANNO takes over a few steps in order to facilitate and speed up the manual alignment: All sentences which are identical in the complex and simplified document are automatically aligned and disabled in the frontend. Furthermore, after manual alignment of a document, all unaligned simple and complex sentences are automatically aligned by insertion or omission.

TS-ANNO further offers a range of features for annotation, e.g., rating evaluation aspects, or annotating simplification operations. With respect to the annotation of simplification operations, various levels of granularity are offered for transformation classes and labels. It also features relative and absolute ratings of sentence pairs in order to manually evaluate gold data or system-generated data. Furthermore, TS-ANNO also features an evaluation of the annotations made in the tool, such as inter-annotator agreement, as well as an output of the annotation in various output formats.

Recently, we have extended TS-ANNO to enable the annotation of not only the correct simplification operations but also of errors. An error annotation can be helpful for manually evaluating system-generated simplifications and identifying their weak points.

Overall, TS-ANNO aims to streamline the process of creating and examining corpora for text simplification purposes, while empowering users with the tools and flexibility needed to effectively perform these tasks. In this paper, the usability of the tool has been exemplified by the annotation of parallel German web texts.

7.4 German Simplification Corpora

In this Chapter, I am presenting my publications that mainly address German text simplification corpora, i.e.,

- Regina Stodden. 2021a. Accessibility and comprehensibility of user-generated content: Challenges and chances for easy-to-understand languages. In *Proceedings of the 3rd Swiss Conference on Barrier-free Communication (BfC 2020)*, pages 151–161, Winterthur (online). ZHAW Zürcher Hochschule für Angewandte Wissenschaften (see Subsection 7.4.1), and
- Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16441–16463, Toronto, Canada. Association for Computational Linguistics (see Subsection 7.4.2).

7.4.1 Accessibility and comprehensibility of user-generated content: Challenges and changes for easy-to-read languages

Reference: Regina Stodden. 2021a. Accessibility and comprehensibility of user-generated content: Challenges and chances for easy-to-understand languages. In *Proceedings of the 3rd Swiss Conference on Barrier-free Communication (BfC 2020)*, pages 151–161, Winterthur (online). ZHAW Zürcher Hochschule für Angewandte Wissenschaften

URL: https://digitalcollection.zhaw.ch/handle/11475/22550

Summary: This paper explores the comprehensibility of noisy user-generated content on social network sites for people with low literacy skills. It emphasizes content literacy instead of usability and technical accessibility, concluding that a CEFR (Common European Framework of Reference for Languages) level B2 is required to understand user-generated texts. As a result, simplification of these texts is necessary for individuals with lower reading and writing skills or a lower CEFR level than B2.

The research focuses on four essential characteristics of user-generated texts, i.e., i) sentence and word length, ii) syntactic and lexical complexity, iii) reciprocal comments in real-time, and iv) emotions, humor, and verification. In comparison of English and German user-generated texts and news texts, it is found that user-generated texts are simpler than professionally written news texts in terms of word and sentence length but have a higher complexity in syntax and lexicon.

In conclusion, while user-generated texts are more complex than other text types, they have the potential to become more accessible to individuals with lower literacy skills. This can be achieved through simplification efforts to reduce lexical and syntactical complexity, improve discourse understanding, clarify emotional statuses, and address the identification of fake news issues. Furthermore, incorporating social network sites and user-generated texts into the guidelines for easy-to-understand standards would greatly benefit users with lower reading and writing skills.

7.4.2 DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification

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

URL: https://doi.org/10.18653/v1/2023.acl-long.908

Summary: In this work, we propose a new parallel German corpus for text simplification called DEplain, created to advance sentence and document simplification in German. It consists of two main corpora: DEplain-APA, which contains news articles in plain German, and DEplain-WEB, a dynamic corpus with parallel texts from the Web. Both corpora are available for document and sentence simplification, with both manual and automated alignments to the sentence level.

The manual alignments plus the sentence-wise alignments of the unaligned documents in the web-domain corpus can be used to evaluate different alignment algorithms for different domains. We have adapted existing methods to German and evaluated them with respect to different domains of German TS data and to different alignment types (i.e., 1:1 and *n*:*m* alignments). We concluded that MASSAlign is the most suitable aligner for our use case, as it i) produces n:m alignments and ii) has fairly high scores for 1:1 and n:m alignments.

The resulting manual and automatic alignments have been analyzed based on human ratings and annotations to better understand the quality and simplification processes within the data. Based on our human ratings, we have verified that DEplain also enables the evaluation of various simplification strategies through its variety of simplification and simplification operations, from rephrasing to splitting and merging. Understanding these strategies can also improve the training of text simplification models.

One use case of using DEplain (also demonstrated in the paper) is to train data-driven text simplification models. We have fine-tuned long-mBART on the document TS corpus and fine-tuned the regular mBART on the sentence TS corpus, to further demonstrate the use of DEplain in training and evaluating text simplification models. We have also experimented with mixing DEplain-APA and DEplain-web: Comparing document simplification of long-mBART trained on APA or web, and trained on APA+web, combining the training data helps to produce better simplifications on DEplain-web test, but impairs the scores on DEplain-APA. Hence, fine-tuning only within the domain of the test set seems to be better than adding more data from another domain. Focusing on sentence simplification results, our mBART models trained on simplification into German Plain and Easy Language achieve also good results on simplification for children. More evaluation is required, especially manual evaluation on the system-generated simplifications, to justify these findings.

7.5 Text Simplification Evaluation

In this Chapter, I am presenting my publications that mainly address the evaluation of text simplification, i.e.,

- Regina Stodden. 2021c. When the Scale is Unclear – Analysis of the Interpretation of Rating Scales in Human Evaluation of Text Simplification. In *Proceedings of the First Workshop on Current Trends in Text Simplification* (*CTTS 2021*), pages 84–95, Online. CEUR-WS (see Subsection 7.5.1),
- David Arps, Jan Kels, Florian Krämer, Yunus Renz, Regina Stodden, and Wiebke Petersen. 2022. HHUplexity at text complexity DE challenge 2022. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 27–32, Potsdam, Germany. Association for Computational Linguistics (see Subsection 7.5.2), and
- Regina Stodden. 2024a. EASSE-DE & EASSE-multi: Easier automatic sentence simplification evaluation for German & multiple languages. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability* (*TSAR 2024*), pages 107–116, Miami, Florida, USA. Association for Computational Linguistics (see Subsection 7.5.3).
- Thorben Schomacker, Miriam Anschütz, Regina Stodden, Georg Groh, and Marina Tropmann-Frick. 2024. Overview of the GermEval 2024 shared task on statement segmentation in German easy language (StaGE). In *Proceedings of GermEval 2024 Shared Task on Statement Segmentation in German Easy Language* (*StaGE*), pages 1–14, Vienna, Austria. Association for Computational Lingustics (see Subsection 7.5.4).

7.5.1 When the scale is unclear – analysis of the interpretation of rating scales in human evaluation of text simplification

Reference: Regina Stodden. 2021c. When the Scale is Unclear – Analysis of the Interpretation of Rating Scales in Human Evaluation of Text Simplification. In *Proceedings of the First Workshop on Current Trends in Text Simplification (CTTS 2021)*, pages 84–95, Online. CEUR-WS

URL: http://ceur-ws.org/Vol-2944/paper6.pdf

Summary: This paper examines the interpretation of evaluation scales for human judgments on the quality of automatically simplified texts, particularly regarding meaning preservation and simplicity. The study analyzes five text simplification datasets and compares the ratings of simplification pairs where the original and simplified sentences are identical. The main research questions address the consistency of labels used by human annotators for the simplicity of identical sentence pairs, the meaning preservation of identical sentence pairs, and whether annotators stick to their interpretation of a scale in all ratings. We are using identical (or no-change) pairs as the control variable because, in this case, for all annotators, the effect of the simplification is the same (i.e., no effect). So, we do not need to consider subjective simplicity assessments in our analysis.

The paper indicates that human annotators mostly agree on one label, the highest value, in the judgments of meaning preservation. In contrast, different interpretations of the simplicity scale were found in the dataset with crowd-sourced human ratings on a scale from 0 to 100. Some raters preferred the lowest value (i.e., 0), and others the middle value of the scale (i.e., 50) to indicate the same level of simplicity in identical pairs. However, for two other data sets with expert annotations on a scale from -2 to $+2$, the interpretation of the scale was better understood and more consistent; the annotators preferred the neutral middle score (i.e., 0).

We emphasize in this study that best practices for human evaluation of text simplification are required to reduce misinterpretations of the scales, including defining the scales more clearly, using scales with a neutral element, providing examples, or relying on experts for annotations.

7.5.2 HHUplexity at text complexity DE challenge 2022

Reference: David Arps, Jan Kels, Florian Krämer, Yunus Renz, Regina Stodden, and Wiebke Petersen. 2022. HHUplexity at text complexity DE challenge 2022. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 27–32, Potsdam, Germany. Association for Computational Linguistics

URL: https://aclanthology.org/2022.germeval-1.5

Summary: This paper discusses a submission to the shared task called Text Complexity DE Challenge 2022, in which the authors aimed to predict the complexity of German sentences as measured by the Mean Opinion Score (MOS). We compared the performance of various regression architectures and transformer language models in combination with hand-crafted features. Our best model is a fine-tuned German DistilBert model with a regression head without adding the linguistic features. This model ranked 7th place in the shared task.

Overall, we have calculated 349 features from seven categories, including length, readability assessment, language proficiency, morphological features, and more. The most novel features are based on a fine-tuned 3-class text level classifier which predicts whether a given sentence is written rather for children, youth, or adults. The predicted labels and the softmax scores per label have been used as additional text level features.

Further, we have experimented with fine-tuning various language models directly on the regression task, including English, German, and multilingual versions of BERT and DistilBERT. When combining a transformer language model with and without addition of linguistic features, the features did not improve the results over the fine-tuning baseline. Although many linguistic features had a high correlation with the MOS scores, they were outperformed by a simple fine-tuned transformer language model.

7.5.3 EASSE-DE & EASSE-multi: Easier Automatic Sentence Simplification Evaluation for German & Multiple Languages

Reference: Regina Stodden. 2024a. EASSE-DE & EASSE-multi: Easier automatic sentence simplification evaluation for German & multiple languages. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability* (*TSAR 2024*), pages 107–116, Miami, Florida, USA. Association for Computational Linguistics

URL: https://aclanthology.org/2024.tsar-1.11

Summary: In this paper, we propose EASSE-multi, a framework for easier automatic sentence simplification evaluation for languages other than English. It contains tokenizers and evaluation metrics suitable for multiple languages. The adaptation of EASSE for non-English languages includes adjustments such as a language constant to specify the evaluated language, language-specific evaluation metrics, and additional tokenizers that consider languages other than English. This approach ensures that EASSE-multi is language-independent and more robust for evaluating non-English texts.

The paper demonstrates the usage of EASSE-multi for German TS resulting in EASSE-DE. In comparing the results generated by EASSE and EASSE-DE, we have shown that it is important to consider the language of the text when evaluating. Language-wise settings in EASSE can impact the TS metrics, e.g., different tokenization effects SARI and BLEU or English vs. multilingual BERTScore. Based on the findings, we recommend reporting which settings were used during the evaluation, as they can significantly influence the TS metrics. The scores may be lower when using EASSE-DE compared to EASSE, but we argue that these are more reliable due to their language sensitivity.

In addition, EASSE-multi helps to make sentence simplification evaluation in languages other than English better and easier to compare. We have gathered available German test sets and system outputs in EASSE-DE, which can be used as a benchmark for German TS.

7.5.4 Overview of the GermEval 2024 Shared Task on Statement Segmentation in German Easy Language (StaGE)

Reference: Thorben Schomacker, Miriam Anschütz, Regina Stodden, Georg Groh, and Marina Tropmann-Frick. 2024. Overview of the GermEval 2024 shared task on statement segmentation in German easy language (StaGE). In *Proceedings of GermEval 2024 Shared Task on Statement Segmentation in German Easy Language (StaGE)*, pages 1–14, Vienna, Austria. Association for Computational Lingustics

URL: https://aclanthology.org/2024.germeval-1.1

Summary: The paper introduces a new approach and benchmark regarding the evaluation of syntactic simplification of texts written in German Easy Language. Following most German Easy Language guidelines, a sentence should contain just "few statements", but it is not defined what a statement actually means. In this work, we are presenting an approach to statement identification. For example, each full verb of a sentence is assigned to one statement, and each non-obligatory argument of this verb is considered as an additional statement.[2]

We have annotated in total more than 4,300 sentences of more than 350 Hurraki articles[3] with their number of statements (subtask 1) as well as the spans of the statements (subtask 2). The data (including labels) and the scoring program are openly available.[4] This data is the basis of the GermEval 2024 shared task on statement segmentation in German Easy Language, in which three teams have participated, and of which two have addressed both subtasks.

The first team (called FriGHt) addresses the task by proposing a model based on binary labeling by using a BERT model to identify the head of a statement. Based on dependency trees of the sentences, the children of each identified head are then marked as part of one statement. The second team (called KlarTextCoder) proposes a rule-based model, a feature-based model, a BERT-based approach, and a LLM-based approach. The BERT-based model is the best approach for both tasks and the winning system. The third team has not provided a system description.

However, both teams could easily beat the baselines (i.e., random baseline and all-statements-equal-1 baseline), but their systems still had challenges in identifying the number and segments of the statements. We believe that our dataset as well as the first system approaches regarding automatic identification of statements in German Easy Language can be helpful in future work to detect complex German sentences by focusing on the syntax in addition to previous work regarding lexical complexity.

2 The complete annotation guideline is available here: https://german-easy-to-read.github.io/statements/annotations/.

3 https://hurraki.de/

4 https://github.com/german-easy-to-read/statements

7.6 German Text Simplification Models

In this Chapter, I am presenting my publications that mainly address German text simplification models, i.e.,

- Regina Stodden. 2024b. Reproduction & benchmarking of German text simplification systems. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 1–15, Torino, Italia. ELRA and ICCL (see Subsection 7.6.1), and
- Regina Stodden and Phillip Nguyen. 2024. Can text simplification help to increase the acceptance of E-participation? In *Proceedings of the First Workshop on Language-driven Deliberation Technology (DELITE) @ LREC-COLING 2024*, pages 20–32, Torino, Italia. ELRA and ICCL (see Subsection 7.6.2).

7.6.1 Reproduction & Benchmarking of German Text Simplification Systems

Reference: Regina Stodden. 2024b. Reproduction & benchmarking of German text simplification systems. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 1–15, Torino, Italia. ELRA and ICCL

URL: https://aclanthology.org/2024.determit-1.1

Summary: We reproduced eight TS systems, including rule-based, fine-tuned, and autoregressive models, and found three main issues: reproduction impossibility due to missing details, restricted access to data/models, or copyright issues; variations in reproduction, making comparisons less meaningful; and differences in evaluation scores between reported and reproduced models due to different system outputs or implementations of metrics.

We reviewed existing German TS models and found six that could be reproduced. We compared the scores of the reproduced models with the original papers to assess if they matched. However, we found that the reproduced Sockeye-APA-LHA model was not comparable to the original. The system generations of zero-shot BLOOM, random 10-shot BLOOM, and similarity 10-shot BLOOM seemed slightly different from the original. The mBART-DEplain-APA and mBART-DEplain-APA+web models had minor differences from the original models.

In previous work, due to different test sets and evaluation metric implementations, the models' results are not comparable. To address this, we propose a German sentence simplification benchmark using 11 models (i.e., 8 reproduced and 3 newly built models) across 7 test sets.

The evaluation results indicate that the models perform best in the test set of the corpus they were trained on. However, mBART-DEPlain-APA, mT5-DEplain-APA, and sockeye-DEplain-APA, which are trained on the same data, show differences in performance due to their distinct system architectures. The analysis also reveals that data augmentation strategies affect system generations' quality: e.g., automatically aligned data and data from different domains seem to lower the quality of mBART-DEplain-APA+web generations in the news domain.

The results show that no single system ranks the best in all the test sets with respect to BLEU and SARI. The additional data on which BLOOM and mBART are pre-trained seem to positively affect the system generations or at least the evaluation scores. Some models, like mBART-DEplain-APA+web, achieve good scores on data with domains and target groups they were not trained on, suggesting that they have learned some universal simplification. Overall, the analysis based on SARI vs. BERTScore ranks different models as best, indicating the need for more research on the suitability of evaluation metrics, particularly for test sets with only one reference, to ensure a more reliable interpretation of the German TS benchmark.

Overall, the study emphasizes the importance of transparency for reproducibility and meaningful model comparisons. We recommend publishing details related to the model, checkpoints, code, and training methodologies for better reproducibility and comparison. Furthermore, we suggest releasing system generations for further analyses.

7.6.2 Can Text Simplification Help to Increase the Acceptance of E-Participation?

Reference: Regina Stodden and Phillip Nguyen. 2024. Can text simplification help to increase the acceptance of E-participation? In *Proceedings of the First Workshop on Language-driven Deliberation Technology (DELITE) @ LREC-COLING 2024*, pages 20–32, Torino, Italia. ELRA and ICCL

URL: https://aclanthology.org/2024.delite-1.3

Summary: In this study, we aimed to explore the effect of text simplification in real-world applications, here online deliberative platforms. A near-realistic experimental study with 276 participants was conducted, simulating a participatory budgeting process. We found that the type of text simplification and the role of participants have no direct influence on the intention to use e-participation, although a higher level of participation cannot be achieved through text simplification. The results showed that people with reading and writing difficulties preferred text simplification for proposals in e-participation processes, regardless of whether manual or automatic simplification.

We have observed that proposal texts seemed difficult to understand for non-native speakers and even for native speakers with reading difficulties. People with reading and writing deficits perceived more confusing terms in proposals and found them more difficult to comprehend due to their length than people without these deficits. Further, we analyzed whether writing or reading a simplification has an effect: we found that text simplification did not negatively impact the acceptance of e-participation processes, regardless of whether individuals had written or read proposal texts.

Overall, the results suggest that NLP, especially automatic text simplification, may be beneficial in online deliberation platforms, particularly for people with reading and writing difficulties, and could potentially reduce the language barrier of such processes.

Part III

Discussion & Conclusion

Chapter 8

Discussion & Future Works

8.1 Overview of the Chapter

Within the context of my PhD thesis, I investigated how German text simplification can be supported through machine learning methods. In the final chapter, I will discuss and summarize how my research has helped to facilitate future German text simplification research and how it has helped to narrow or close the research gaps and questions opened in Part I. I will consider each component of the text simplification workflow and review my findings of the contributions made by me and my co-authors in a comprehensive manner (see Figure 8.1). Furthermore, I will suggest potential directions for future research in the field of German text simplification.

The remainder of this chapter is organized following the components of the text simplification workflow (see Figure 8.1) and their corresponding research questions (see Subsection 1.2.1). In more detail, this chapter contains contributions regarding simplicity and simplification (see Section 8.2; component 0 in TS workflow), building text simplification corpora (see Section 8.3; components A to F), German simplification corpora (see Section 8.4; components B, E, and F), evaluation of text simplification (see Section 8.5; component H), and models of German text simplification (see Section 8.6; components G to J).

8.2 Complexity & Simplification

As previously introduced, identification of complexity and applied simplification operations are relevant tasks in the scope of text simplification (see step 0 in text simplification workflow) in order to understand i) what makes a text complex or simple, ii) whether simplification of a text is required, iii) and, if yes, to which extent. Furthermore, knowledge regarding simplification operations can help identify the extent of simplification as well as identify the characteristics of the simplification processes in different languages and language varieties.

In this section, I will discuss the extent to which my research has contributed to improve the identification of sentence and word complexity in German texts. It will also address the role of simplification operations in these tasks.

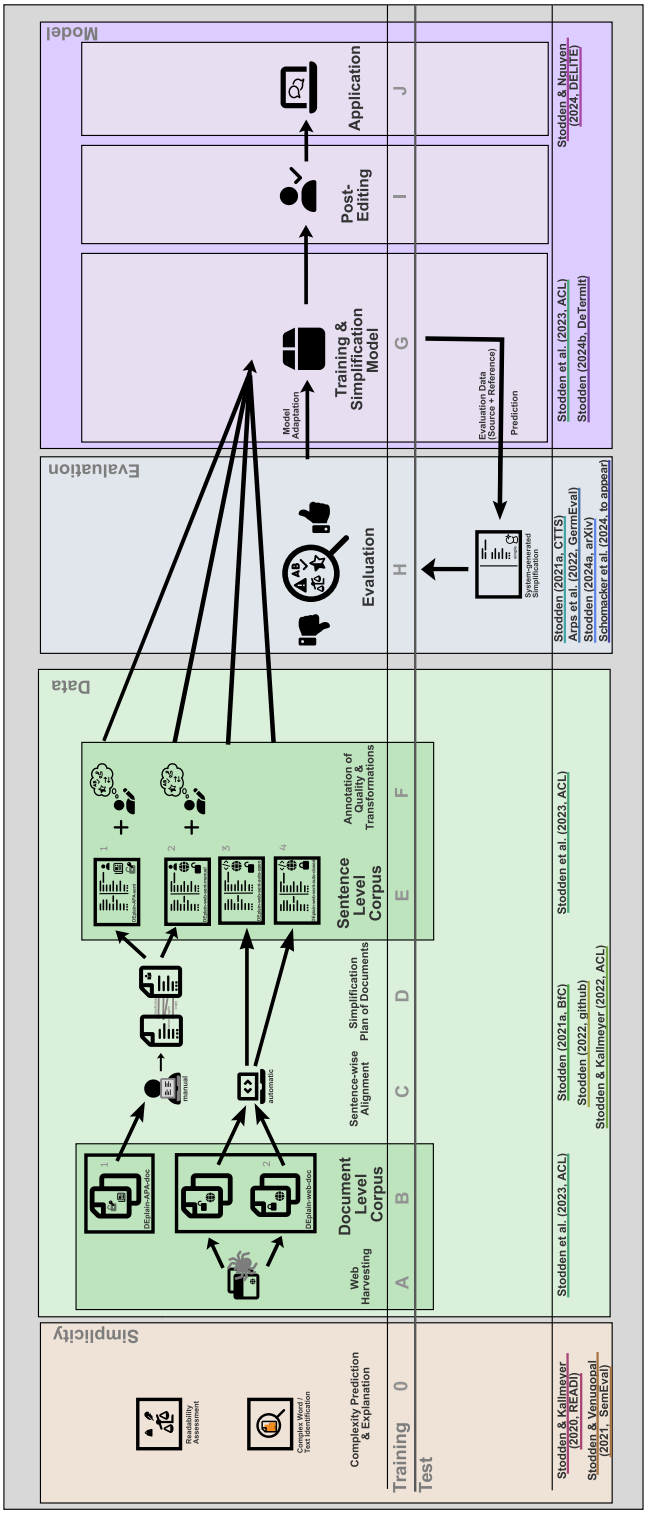


Figure 8.1: Text simplification workflow including contributions of this thesis (extended version of Figure 1.1).

8.2.1 RQ 2-1: German Simplification Operations

In previous research, many guidelines and rules have been proposed for writing simpler German (e.g., Bredel and Maaß 2016; Netzwerk Leichte Sprache 2022; Baumert 2018; Deutsches Institut für Normung (DIN) 2023; Bock and Pappert 2023; Deutsches Institut für Normung (DIN) 2024b), but it has not been clear whether or to what extent they have been actually applied in the manual simplification of German texts. As a consequence, the variety and extent of how simplification operations have been considered in German data-driven text simplification have not been discussed before.

We have addressed RQ 2-1 regarding commonly used simplification operations in German TS by narrowing the gap between manual and automatic German text simplification. In Stodden (2022) (see Subsection 7.3.1), we have reviewed 100 publications regarding simplification operations, that is, 41 documents on recommendations for German Plain and Easy Language and 59 documents regarding features taken into account in automatic text simplification of five languages (i.e., Czech, German, English, Spanish, and Italian).[1] Based on this literature review, including related typologies of other languages (e.g., Italian Brunato et al., 2015, Spanish Bott and Saggion, 2014, Basque Gonzalez-Dios et al., 2018, and French Koptient et al., 2019), we have built the first typology for simplification operations in German ATS (see Section 5 in Stodden 2022, and Table 2 in Stodden and Kallmeyer 2022). In contrast to guidelines regarding how to write simplified languages, the typology includes not only recommendations to avoid complex linguistic phenomena, but also includes strategies on how to rewrite it wrt. simplicity, e.g., lexical substitution with hyponyms or hypernyms, or changing the subject-verb order.

A visual overview of the typology is provided in Figure 8.2. The typology is grouped into the text levels on which the operation will be performed (see y-axis), i.e., word, phrase, sentence or text level, and the priority of the operation (see x-axis), i.e., 1 to 4, where 1 is the highest. The more often an operation has been named in the publications, the higher the priority. Furthermore, the typology includes main operation classes (see operations highlighted in bold face) and subclasses (see operations grouped into boxes) which allow for different levels of granularity.

Based on the analysis and typology, we revealed that some of the simplification operations are general for simplification across languages and domains, e.g., complex word replacement or sentence splitting. But, we have also identified operations that are more specific for one language, e.g., compound segmentation or explanation generation for German (Stodden, 2021b). Furthermore, we found that the named operations are frequently mentioned in German simplification guidelines but are currently not considered in TS research. Reasons for this might be that long, one-token compounds are frequent in German, but not in other languages of current TS research. On the other hand, we found that copying text from the original to the simplified text (e.g., if words or sentences are simple enough and do not require simplification) is often mentioned in TS research, but rarely in German simplification guidelines. Copying might be of higher relevance in automatic in TS as it is relevant for rule-based models as well as for evaluation, e.g., in SARI.

1 Not enough resources on German simplification has been available for a literature review on only German TS.

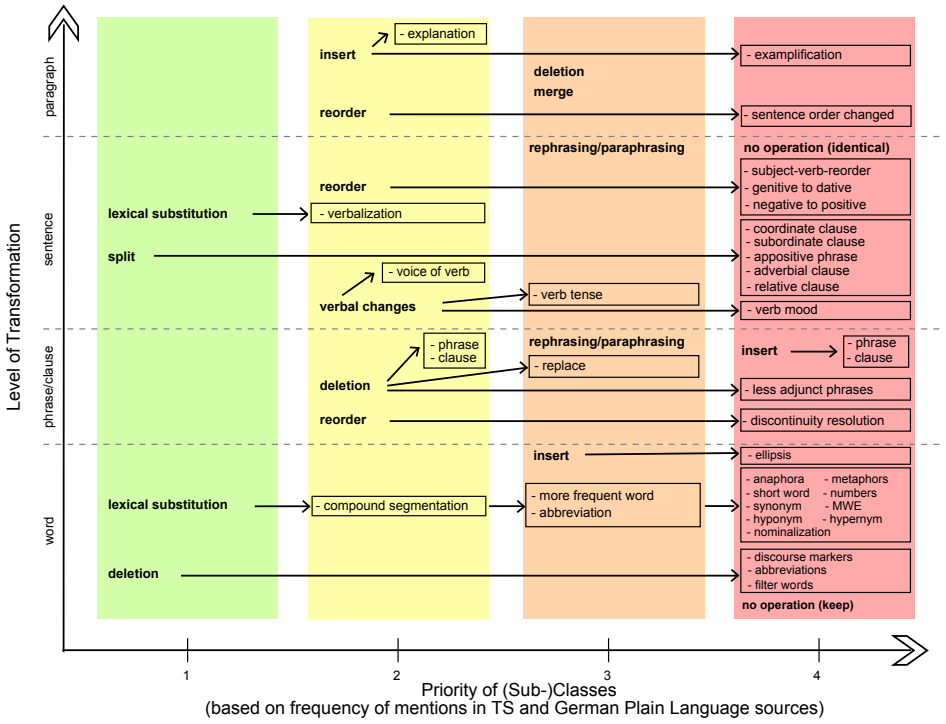


Figure 8.2: Typology of German simplification operations.

As manually verifying the existence of the simplification operations is very time consuming, in Stodden and Kallmeyer (2020) (see Subsection 7.2.1), we have implemented an automatic extraction of linguistic features considering the operations of the typology to get first insights regarding the simplification processes in the TS corpora. We used the feature extraction approach to compare simplification operations across languages (i.e., German, English, Czech, and Italian) and domains (i.e., news, web, and Wikipedia) using gold simplification corpora. Similarly to Vajjala and Meurers (2014), Scarton et al. (2017), and Finnimore et al. (2019), we have found that text simplification seems to be consistent across languages and domains (using our feature set), but features might be weighted differently per language (e.g., compound splitting is more relevant for German than for English).

However, this typology still leaves some room for improvement, e.g., regarding document simplification. Focusing more on discourse-aware simplification within a whole document, other simplification operations could play a more prominent role, e.g. moving sentences from bottom to top following their relevance (e.g., see Lin et al. 2021), or coherence in terms of making sentence contexts and connections between sentences more obvious (e.g., see Vásquez-Rodríguez et al. 2023). Currently, the existing typology could help for more fine-grained evaluation at the sentence level, but in the future, the document level should be put more into focus.

Furthermore, in addition to automatic annotation of the linguistic operations, the typology could also be used to manually annotate manual or automatic simplifications. On the one hand,

this would facilitate a better understanding of the variety of a corpus or the capabilities of a text simplification system. On the other hand, we could verify the quality of our automatic feature extraction compared to manual annotations of the operations. In the next subsection (Subsection 8.2.2), I will introduce how automatic feature extraction can be used to identify the complexity of a word or sentence. In Subsection 8.3.2, I will further discuss the manual application of the typology on our proposed corpus DEplain.

8.2.2 RQ 2-2: Identification and Explanation of Complex Texts

Moreover, research on the automatic identification and explanation of complexity passages in German texts is rare. To overcome this research gap and answer RQ 2-2 about how and to what extent complex passages can be identified by presenting valuable knowledge of our contributions, I present our contributions regarding complexity prediction, i.e., Stodden and Venugopal (2021) and Arps et al. (2022).

Only a few studies exist regarding complex word identification (CWI), that are, five systems that have been built in the scope of the CWI shared task in 2018 (Yimam et al., 2018), as well as only a few systems regarding the prediction of lexical complexity in German, i.e., two system proposals in the MLSP shared task 2024 (Shardlow et al., 2024).

Unfortunately, as there are too few data on the identification of complex German words within the duration of conducting my PhD, I first tackled the prediction of English lexical complexity (with regard to words within a sentence) in Stodden and Venugopal (2021) (see Subsection 7.2.2). Then, I tried to transfer these findings to the prediction of the complexity of German sentences in (Arps et al., 2022). Both approaches are based on the extraction of linguistic features using the implementation of Stodden and Kallmeyer (2020) (see Subsection 7.2.1).

However, the transfer to the German sentence level has not been successful; our best German sentence complexity prediction model is a simple fine-tuned transformer model with a classification head (Arps et al., 2022) (see Subsection 7.5.2). In fact, the addition of linguistic features to this model has reduced its performance. Following this, we can conclude wrt. RQ 2-2 that automatic identification of complex text passages is still challenging due to the high subjectivity of complexity and the lack of quantifiable ways to measure it. Further investigation is required to classify and identify complex passages of a German text. Our approach has been already used by Thome et al. (2024): they combined also fine-tuning of a BERT model with additional features for complexity prediction of German sentences. However, they included person-related features instead of linguistic features and achieved a higher RMSE score on the same test set (i.e., TextComplexityDE) than our approach.

In future work, our findings could be combined with recent work, e.g., the resources and results of Shardlow et al. (2024) or the text leveling models of Klepp (2022a). The improved work could then be integrated into the text simplification workflow, e.g., by identifying complex sentences of a document that require simplification (e.g., see Garbacea et al. 2021) or as an automatic evaluation method wrt. simplicity in extension to or as a replacement of readability metrics.

8.3 Building Text Simplification Corpora

In automatic TS, parallel corpora are precious resources for training and evaluating TS systems (see steps B and E in the text simplification workflow). Ideally, these corpora should contain document pairs, paragraph pairs, or manually aligned sentence pairs consisting of the original text and its corresponding professionally simplified text (see step C in the text simplification workflow). Currently, however, high-quality resources and corpora of this type are rare (Building Challenge A) and often of comparably small size (see Data Challenge E), e.g., Simple German Corpus '13 (Klaper et al., 2013), APA-RST (Hewett, 2023), or ABGB Meister 2023. Reasons for this are, for example, that manual sentence-wise alignment is very time-consuming and there is no digital assistance for manual sentence-wise alignment (Building Challenge B).

Therefore, if a sentence simplification corpus is of a comparable larger size (e.g., more than 1,000 sentence pairs), the pairs are often automatically aligned and contain many misalignments (Building Challenge C). Also, often resources that were not designed for TS in the first place are often used to train TS systems (Stajner, 2021) (see Data Challenge F), i.e., comparable but not parallel resources (e.g., see Wikipedia Corpus Ebling et al. 2022, Lexica Corpus Hewett and Stede 2021, or 20min Rios et al. 2021). However, quality control of the corpora is missing before using them for training or evaluation due, e.g., to a lack of a set of criteria for high-quality corpora (see Building Challenge D). In the remainder of this section, I discuss how my work has addressed these issues.

8.3.1 RQ 3-1: Corpus Creation Challenges

Within the sections of the state of research regarding TS corpora (see Chapter 3 and Chapter 4), I have already introduced the main challenges in the process of creating parallel corpora (see RQ 3-1) that I have also identified in my publications, i.e., Stodden and Kallmeyer (2022) and Stodden et al. (2023). In summary, I have identified the following challenges as most important for the corpus building process that answers the first part of RQ 3-1:

- Building Challenge A: lack of accessible resources for parallel corpora,
- Building Challenge B: time-consuming manual sentence-wise alignment and missing digital assistance for manual sentence-wise alignment,
- Building Challenge C: missing reliability of automatic alignment methods, and
- Building Challenge D: missing quality control of the corpora due to a lack of a set of criteria for high-quality corpora.

In order to tackle the second part of RQ 3-1, we have proposed some methods to overcome these challenges: to facilitate the creation of new corpora considering sentence-wise alignment, in this work, we have proposed a web harvester (see Subsubsection 8.3.1.1), an annotation tool (see Subsubsection 8.3.1.2), and made available manually and automatically aligned sentence pairs as well as automatic alignment methods (see Subsubsection 8.3.1.3).

8.3.1.1 Web Harvester

In Stodden et al. (2023) (see Subsection 7.4.2), we have proposed a web harvester to collect parallel data from the web and tackle Building Challenge A.[2] The tool harvests documents in simplified German and aligns them with complex documents if available. The output of the web harvester is parallel plain texts of complex-simple documents which can be directly used to build a document simplification corpus (see step B in the TS workflow). The harvester code is also capable of being integrated into our TS-ANNO annotation tool (see below).

Compared to other web harvesters introduced in Section 4.1 and Subsection 4.7.3, our web harvester includes more parallel resources than other harvesters (e.g., KED by Jach 2023, Leiko by Jablotschkin and Zinsmeister 2020 or Klepp, 2022b). However, it focuses more on German Plain Language than German Easy Language, while most of the other corpora focus on the latter (e.g., Anschütz et al. 2023 or SGC '23 by Toborek et al. 2023). For a full comparison of the existing web harvesters, including ours, see Appendix.

In future work, our web harvester and those of related work could be merged to a huge web harvester of both German Plain and Easy Language. Furthermore, our web harvester is currently sensitive to changes in the to-be-crawled webpages. Therefore, in another version, the harvester could be extended to rely on archived URLs, such as proposed by Toborek et al. (2023).

8.3.1.2 Annotation Tool

In order to facilitate manual alignment and manual annotation of the parallel documents (e.g., crawled with the web harvester) for sentence simplification corpora (see Building Challenge B), in Stodden and Kallmeyer (2022) (see Subsection 7.3.2), we have proposed an annotation tool called TS-ANNO, which is designed especially for the purpose of building and annotating text simplification corpora.

In general, TS-ANNO supports all parts of the corpus building process, that is, using a web harvester (step A in Figure 8.5), resulting in parallel document simplification corpora (step B), manual simplification (step C.1), manually aligning complex-simple sentence pairs (step C.2), creating simplification plans per document (step D), resulting in sentence simplification corpora (step E), and annotating simplification operations and simplification quality (step F). To the best of my knowledge, TS-ANNO is the only annotation tool that covers the whole TS corpus building process; other tools, e.g., just support alignment (e.g., Tiedemann 2006, Paetzold et al. 2017), simplification (e.g., Caseli et al. 2009), quality annotation (see, e.g., Alva-Manchego et al. 2020a or Gonzalez-Dios et al. 2018), or error annotation (e.g., Heineman et al. 2023).

We have already used TS-ANNO to build, align, and annotate new document and sentence simplification corpora, i.e., DEplain-APA and DEplain-web (Stodden et al., 2023) (see Subsection 8.4.2 for more details). Additionally, we have recently also extended TS-ANNO in order to facilitate error annotation in complex-simple pairs of manually simplified as well as automatically simplified texts for our research on error annotation in German TS system outputs (see Lemgen 2024).

2 The code of the web harvester is available here https://github.com/rstodden/data_collection_german_simplification [last update: June 19, 2023; last access: July 24, 2024].

In further work, I want to extend the simplification step by providing an automatically simplified draft of any TS model; currently, TS-ANNO supports only one multilingual model, i.e., MUSS (Martin et al., 2022).

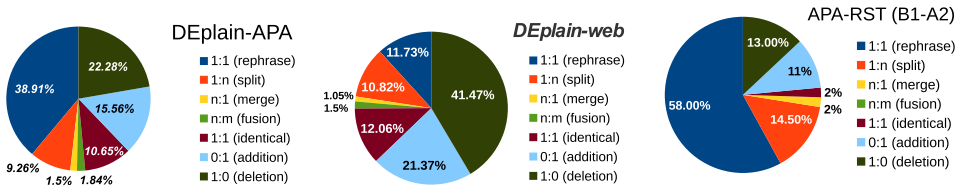
8.3.1.3 Sentence-wise Alignment

As text simplification is often performed on a sentence level, three additional problems of TS corpora arise: no sentence-level alignment in document-level corpora (e.g., see Lexica Corpus Hewett and Stede 2021, Klexikon Aumiller and Gertz 2022, or 20Minuten Rios et al. 2021), alignments do not consider all *n*:*m* alignment types (e.g., see Toborek et al. 2023), or error-prone automatic sentence alignment (e.g., Spring et al. 2021) (Stajner, 2021). That is why sentence-wise alignment (see step C in TS building process) is the biggest challenge when building new representative sentence corpora (see step E); no matter if manually (see Building Challenge B) or automatically (see Building Challenge C) aligned. Furthermore, not only is the alignment process itself a challenge, but also finding a balance between 1:1 and *n*:*m* alignment pairs.

Manual Alignment In order to facilitate manual sentence-wise alignment (see Building Challenge B), we have proposed TS-ANNO which is capable of all *n*:*m* alignment types and crossing alignments (for more information on alignment types, see Stodden 2022). We have already used the tool in practice when aligning the parallel documents of our DEplain corpus, that is based on documents gathered with our web harvester as well as news texts of the Austrian Press Agency (APA). The alignment process has resulted in the sentence-level TS corpus of DEplain (see Subsection 8.4.2), i.e., 1,846 sentence pairs of web texts (called DEplain-web) and 13,122 sentence pairs of news texts (called DEplain-APA).

When analyzing the DEplain corpus and other corpora more closely regarding their alignment types on the sentence level, most corpora contain only a small portion of *n*:*m* alignments (e.g., less than 10% in automatically aligned APA-LHA Spring et al. 2021; see Table 4.8 and Table 4.9) even if *n*:*m* alignments are a frequent result of various simplification operations, e.g., merging, splitting, or sentence fusion. In comparison, in our DEplain-APA corpus (Stodden et al., 2023), we found that roughly 50% of our manually aligned sentence pairs are 1:1 pairs (including ■ rephrases and ■ identical pairs in Figure 8.3a), 10% are a 1:*n* (see ■ split), 1.5% *m*:1 (see ■ merge), 1.8% *m*:*n* (see ■ sentence fusion), and the remaining are split into deletion (roughly 22%, see ■ deletion) and addition (roughly 15%, see ■ addition). However, the APA-RST corpus (see Subsection 4.3.5) which is built on the same data as DEplain-APA and also manually aligned, contains more rephrases and splits, but fewer deletions and identical sentence pairs (see Figure 8.3c). This might be due to different alignment strategies or due to a different data sample and sample size, i.e., 5 documents in APA-RST vs. 483 document pairs in DEplain-APA.

However, for the alignment types in the web domain with various simplification purposes, i.e., DEplain-web, we found a different picture (see Figure 8.3b). In this corpus, most complex sentences have been deleted (41%, see ■ deletion), but 21% of simple sentences added (see ■ addition), 23% 1:1 alignments (see ■ rephrase and ■ identical), 10% 1:*n* (see ■ split) and roughly 1% of each *n*:1 (see ■ merge) and *n*:*m* (see ■ fusion).



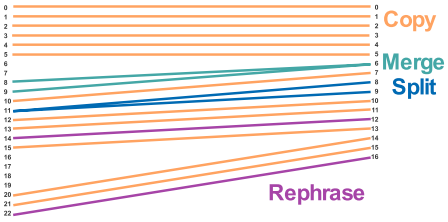
(**a**) DEplain-APA (n=13,122) (**b**) DEplain-web (n=1,846) (**c**) APA-RST-B1-A2 (n=200)

Figure 8.3: Alignment types in different German TS corpora.

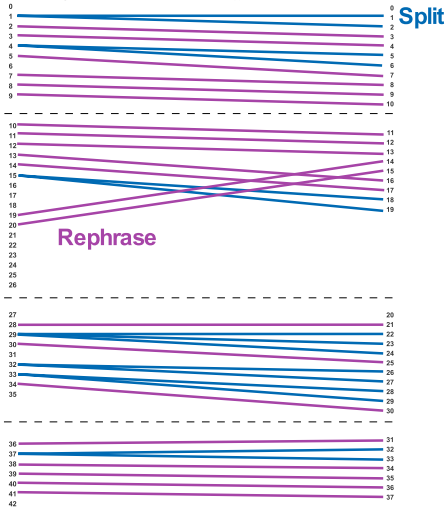
In order to find reasons for the different distribution of alignment types in the three corpora, we had a closer look at the simplification plans of a few documents. Figure 8.4 provides visualization of the manual alignment of three documents of different domains, i.e., language learner texts (see Figure 8.4a), news (see Figure 8.4b), and health texts (see Figure 8.4c). Each document is an example of the typical alignment of the documents of this resource: for example, in texts of Alumniportal, many sentences are copied from the complex document (i.e., written for people with CEFR level B1) to the simplified document (i.e., written for people with CEFR level A2). The most common simplification technique used in APA texts is rephrasing. Another common practice is splitting a long sentence into several shorter ones. Additionally, the texts are reordered during simplification, which results in crossing alignment lines (see complex sentences 19 and 20 in Figure 8.4b). In contrast, the simplified documents of Apotheken Umschau are much shorter than the original documents (hence, they contain many deletions), although they contain many sentences which cannot be aligned to any complex sentence (i.e., additions; see simple sentences 7 to 11 or 31 to 35 in Figure 8.4c). Characteristics of this resource are also a high variation between alignment types, i.e., splitting, sentence fusion, rephrasing, deletion, and addition, as well as crossing alignment lines due to the reordering of the document's content.

We have also analyzed the inter-annotator agreement between the two annotators of the DEplain corpus wrt. different text resources (i.e., domains and target group) to check if or to which extent the annotators follow the alignment guidelines and if they understood it the same way. As summarized in Table 8.1, it seems that manual alignment is still a challenging task. Although both annotators have been provided with the same annotation guidelines, they achieved almost perfect agreement for only one domain (that is, language learner texts from the AlumniPortal) and high moderate agreement for another domain (that is, news texts from DEplain-APA), which contain weakly to moderately simplified text pairs from CEFR level B1 to A2 where the first has a low amount of *n*:*m* alignments. In contrast, the annotators achieved a weak agreement for health-related texts (i.e., texts by Apotheken Umschau or BZFE), which contain many *n*:*m* alignments with long distance cross alignments. Hence, we conclude that the alignment is dependent on the domains and the simplification strength of the texts.

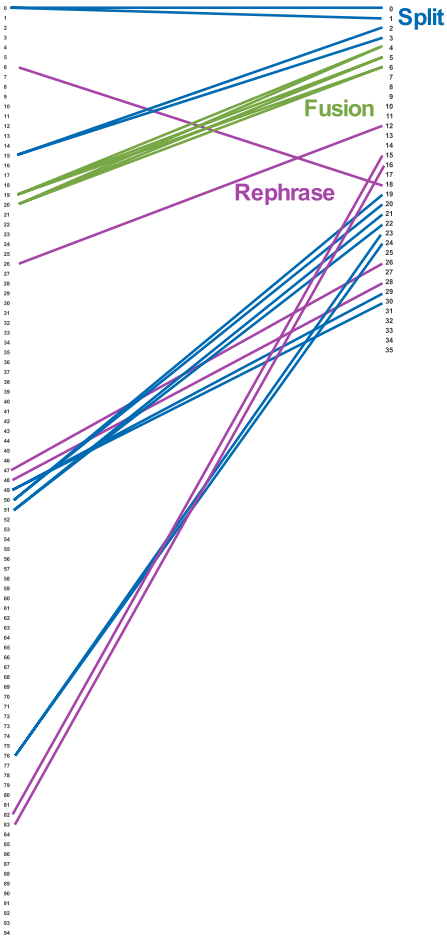
In order to answer RQ 3-1 wrt. manual alignment, the alignment has been significantly faster when using TS-ANNO than when aligning the texts without assistance, i.e., copying sentence pairs from text files. Furthermore, TS-ANNO could save time in the annotation process regarding the alignment of identical pairs, deletions, and additions because they have been au-



(**a**) Alumniportal document (Title: "Mensch ärgere Dich nicht – spiele!"; ID = 103).



(**b**) APA document with four news items separated by a dashed line (Title: "Nachrichten leicht verständlich vom 20191021"; ID = 403).



(**c**) ApothekenUmschau document (Title: "Arthrose der Hand- und Fingergelenke"; ID: 291).

Figure 8.4: Alignments between complex document (left) and simple document (right). The full document texts can be found in the Appendix.

Domain	κ Avg.	Std.	Interpretation	# Sents	# Docs
bible	0.7011	0.31	moderate	6,903	3
health	0.5147	0.28	weak	13,736	6
language learner	0.9149	0.17	almost perfect	18,493	65
narrative	0.6131	0.39	moderate	23,289	3
news	0.7497	0.28	moderate	25,224	10
all	0.8505	0.23	strong	87,645	87

Table 8.1: Inter-Annotator agreement per domain including average, standard deviation, number of sentence combinations (# sents), and number of documents (# docs). Copied from Stodden et al. (2023).

tomatically added to the corpus after finishing the alignment of a document. Additionally, the "find-a-similar-sentence" function has helped to make faster decisions regarding the alignment of complex and simple sentences, especially for crossing and long distance sentence pairs. However, manual alignment is still time-consuming; hence, reliable automatic alignment seems to be of high relevance to create large corpora (with more than 1,000 sentence pairs).

Automatic Alignment In order to also tackle the alignment with automatic methods (see Building Challenge C and Data Challenge G), in Stodden et al. (2023), we have adapted existing alignment methods to German, applied, and evaluated them on different domains of German TS data (i.e., DEplain-APA and DEplain-web). In addition to existing studies on the suitability of automatic alignment methods for German TS, e.g., see Spring et al. (2022), Spring et al. (2023), or Toborek et al. (2023), we have evaluated more alignment methods (e.g., BERTAlign Liu and Zhu 2022 or SentenceTransformer Reimers and Gurevych 2020 with LABSE Feng et al. 2022) and have focused more on 1:1 and *n*:*m* alignment.

In our experiments, we observed that producing *n*:*m* alignments is a challenging task (see results in the lower part in Table 8.2) which we could not solve. We found, on the one hand, that SentenceTransformer using the multilingual LaBSE model (Feng et al., 2022) got very high precision for 1:1 alignments and a fair recall as well (see second row in Table 8.2). On the other hand, MASSAlign performed the best on *n*:*m* results, and also with totally acceptable 1:1 results (see last row in Table 8.2). Hence, we conclude that MASSAlign is the most suitable aligner for our use case, as it produces *n*:*m* alignments and has fairly high scores for 1:1 and *n*:*m* alignments.

	1:1				*n*:*m*			
Name	**P**	**R**	**F_1**	**$F_{0.5}$**	**P**	**R**	**F_1**	**$F_{0.5}$**
LHA	.94	.41	.57	.747	-	-	-	-
Sent-LaBSE	**.961**	.444	.608	**.780**	-	-	-	-
Sent-RoBERTa	.960	.444	.607	.779	-	-	-	-
CATS-C3G	.247	**.553**	.342	.278	-	-	-	-
VecAlign	.271	.404	.323	.290	.260	.465	.333	.285
BERTalign	.743	.465	.572	.664	.387	.561	.458	.412
MASSAlign	.846	.477	**.610**	.733	**.819**	.509	**.628**	**.730**

Table 8.2: Results of the alignment methods with 1:1 (upper part) and *n*:*m* capabilities (lower part) on sentence pairs with 1:1 (n=1750, left part) and *n*:*m* alignments (n=991, right part) wrt. precision (P), recall (R), F1 score (harmonic mean of P&R), and $F_{0.5}$ score (more emphasis on P than R). Copied from Stodden et al. (2023).

As discussed previously, in related work and also in our work, the scores for automatic alignment are in general not satisfactory. Following Stajner (2021), these methods work well for mild simplification, but seem to struggle for strong simplifications where the structure and semantics were highly changed. In order to verify this assumption, we have also analyzed our alignment pairs with respect to the websites. In Stodden et al. (2023), we found that mild simplifications such as simplifications between two close CEFR levels (i.e., Alumniportal, see second row of Table 8.3b) or DEplain-APA (see first row of Table 8.3a) achieve high F1 scores. For all other resources, the score is dramatically low: e.g., for simplification from standard German into German Easy Language (fairy tales or bible, see third and fourth row in Table 8.3b), simplification from old German into German Plain Language (simple books, see first row in Table 8.3b), or simplification including fully reorganisation of a document (BZFE, see last row in Table 8.3b) are challenging for MASSAlign. Hence, the automatic alignment seems to perform well in the domains where the inter-annotator agreement of the manual alignment is also high (see news and language learner texts), and vice versa for low agreement (e.g., see health or bible texts).

Corpus	Domain	P	R	F1
DEplain-APA	news	0.63	0.35	0.45
DEplain-web-public	web	0.82	0.51	0.63

(**a**) Results of DEplain-APA vs. DEplain-web

Domain	Subcorpus	P	R	F1
bible	Bible	0.12	0.03	0.05
health	BZFE	0.09	0.02	0.03
lang. learner	Alumniportal	0.90	0.87	0.88
narrative	NDR Fairy Tales	0.07	0.01	0.02
narrative	Spaß am Lesen Verlag	0.11	0.02	0.04

(**b**) Results per web page of DEplain-web.

Table 8.3: Results of MASSAlign on DEplain-web and DEplain-APA.

In order to answer RQ 3-1 wrt. automatic alignment, it still remains an open question how mild and strong simplifications can be automatically aligned with satisfactory results. Overall, we could not completely overcome the challenges of alignment (i.e., Building Challenge C, and Data Challenge G), but we have offered many manually aligned sentence pairs (including different simplification extents, simplification operations, and alignment types) that can be used to improve and evaluate alignment algorithms. One direction for future work might be to train a neural model on manually aligned sentence pairs of the original and simplified documents and learn whether the pairs are fully aligned, partially aligned, or not aligned (similar to the approach proposed by Jiang et al. 2020). In this case, the manual alignment of our DEplain corpus (Stodden et al., 2023) could be used for the training and evaluation process.

However, disregarding the method of automatic alignment, before using an automatically aligned sentence simplification corpus, we recommend always manually checking some alignment pairs in order to ensure sufficient quality.

8.3.2 RQ 3-2: Characteristics of new Corpora and RQ 3-3: Quality & Representativeness of Corpora

When using a newly built or existing corpus for automatic text simplification, the question arises whether the corpus is of high quality and suitable for the planned purpose (see RQ 3-2) and what features can be applied during the building process in order to ensure high quality (see

RQ 3-3 and Building Challenge D). In order to verify the quality of gold data and facilitate the annotation, in Stodden and Kallmeyer (2022) (see Subsection 7.3.2) and Stodden (2022) (see Subsection 7.3.1), we propose that a manual, quantitative analysis regarding the quality of complex-simple pairs can give more insight in the data than automatically measuring scores regarding readability or sentence length (e.g., see Vajjala and Lučić 2018 or Scarton et al. 2018).

Previously to our work, the quality of the gold data of a corpus has very rarely been assessed, e.g., by annotation of simplification operations (see e.g., Cardon et al. 2023) or by human judgments regarding fluency, meaning preservation, or simplicity (see e.g., Sulem et al. 2018b or Alva-Manchego et al. 2020a; see step F in text simplification workflow). The collection of gold data judgments has mostly been overlooked, or gathered simultaneously with system output judgments. In this case, however, the data has already been used to train a TS model before applying a quality check, which makes more sense the other way round. Currently, no German TS corpus is available that is accompanied by annotations of simplification operations or by human judgments regarding fluency, meaning preservation, or simplicity Also, only a few corpora in other languages contain information about the actual types of simplification (simplification transformations, respective grammaticality, lexical complexity, etc. of the aligned sentences, etc.), e.g., PorSimple Corpus for Brazilian Portuguese Caseli et al. (2009), SimpleSEW corpus for Amancio and Specia (2014), Terence & Teacher corpus for Italian Brunato et al. (2015), or ASSET$_{ANN}$ for English (Cardon et al., 2022). To the best of my knowledge, our DEplain corpus (Stodden et al., 2023) is the first German corpus with annotations on evaluation aspects and simplification operations on the golden simplification pairs.

In Stodden (2022) (see Subsection 7.3.1), we argue, on the one hand, that the three common evaluation aspects (that is, meaning preservation, simplicity, and fluency) are not enough to evaluate the quality of a TS corpus or system-generated simplification (this also addresses Evaluation Challenge A and Evaluation Challenge B). If we would rate the following example (item 1 and item 2) with respect to common aspects, we may get the result of a grammatically correct sentence which might preserve the original meaning but is longer than the original text, and hence the gold simplification might be overall a little bit more complex than the original text. However, we would overlook that the simplification is better readable wrt. coherence because the main message of the simplification (in contrast to the original sentence) is understandable without reading the whole paragraph. This is due to newly acquired information, that is, the exact year and a description of the reports. But, without any context, an annotator cannot verify whether the new information is correct, and, also, a text simplification system could not correctly generate this information based on just the isolated complex sentence.

(1) *Standard German:*
Seitdem gab es jedes Jahr mehr als 500.000 Anzeigen.

'Since then, there have been every year more than 500,000 complaints.'

(2) *German Plain Language:*
Nach 1999 gab es jedes Jahr über 500.000 Anzeigen wegen Verbrechen in Österreich.

'After 1999 there were every year over 500,000 criminal complaints in Austria.'

Therefore, in Stodden (2022) (see Subsection 7.3.1) and Stodden and Kallmeyer (2022) (see Subsection 7.3.2), we have proposed an extensive recommendation for the intrinsic human evaluation of (German) TS that includes new adapted evaluation aspects. For an overview of all aspects including their statements in English, see Table 8.4.[3] These rating aspects are also included in TS-ANNO (Stodden and Kallmeyer, 2022) to facilitate the assessment of the system outputs and the gold data following our annotation schema. However, TS-ANNO is easily adaptable, and hence more or less criteria can be integrated into the analysis.

Aspect	Statement	Scale	Aspect	Statement	Scale
Simplicity (simple)	The simplified sentence is easy to understand.	1 to 5	**Ambiguity (simple)**	The simplified sentence is ambiguous. It can be read in different ways.	1 to 5
Simplicity (original)	The original sentence is easy to understand.	1 to 5	**Ambiguity (original)**	The original sentence is ambiguous. It can be read in different ways.	1 to 5
Grammaticality (simple)	The simplified sentence is fluent, there are no grammatical errors.	-2 to +2	**Lexical Simplicity**	The words of the simplified sentence are easier to understand than the words of the original sentence.	-2 to +2
Grammaticality (original)	The original sentence is fluent, there are no grammatical errors.	-2 to +2	**Structural Simplicity**	The structure of the simplified sentence is easier to understand than the structure of the original sentence.	-2 to +2
Coherence (simple)	The simplified sentence is understandable without reading the whole paragraph.	1 to 5	**Overall Simplicity**	The simplified sentence is easier to understand than the original sentence.	-2 to +2
Coherence (original)	The original sentence is understandable without reading the whole paragraph.	1 to 5	**Meaning Preservation**	The simplified sentence adequately expresses the meaning of the original sentence, perhaps omitting the least important information.	1 to 5
			Information Gain	In the simplified sentence, information is added or gets more explicit than in the original sentence.	-2 to +2

Table 8.4: Rating aspects.

In current human evaluation, only the grammaticality of the system output is assessed (Alva-Manchego et al., 2020b). But, if a source sentence already contains grammatical errors, it could be that these have been retained in the simplification (e.g., when using the BiSECT corpus; Subsection 4.1.5). Hence, in this case, the TS system would have not generated but simply copied the grammatical errors. Therefore, in Stodden (2022), we argue that we should assess the grammaticality of the system output, but also of the source sentence (see *Grammaticality* (*original*) and the references to verify the quality of the gold data (see *Grammaticality* (*simple*) in Table 8.4).

Furthermore, we propose to evaluate simplicity more fine-grained: Hence, our aspect collection includes structural simplicity (similar to Sulem et al. 2018b), overall simplicity (similar to Brunato et al. 2018 or Alva-Manchego et al. 2020b), but also lexical simplicity (new), and simplicity of the original as well as the simplified sentence (new). If the initial complex sentence is already simple, there may not be much room for improvement in terms of simplification. We therefore propose to take into account the simplicity level of the original sentence as a starting point (see and *Simplicity* (*original*) in Table 8.4) as well as the simplicity level of the simplified sentence (either manually or automatically) as an endpoint. Additionally, on the basis of these two values, the distance or extent of simplicity between the source text and the simplified text can be measured. This procedure avoids previously discussed misunderstandings of the scale, i.e., whether a low value of the scale indicates that a simplification is more complex than

3 For evaluation with German speakers, we have also developed a German version. This version is available in the Appendix.

an original sentence or only very slightly simplified (see Stodden 2021c). However, our proposed evaluation schema also includes *Overall Simplicity* to measure the extent following the common evaluation procedure and to be better comparable to previous work. In future work, both extent values, i.e., overall simplicity, and the difference between simplicity (simple) and simplicity (original), can be compared to verify which method is more reliable.

Additionally, we propose to include coherence (new), and ambiguity (new) for both the simplified and the original sentences in the evaluation procedure to also tackle the discourse level. The original sentence in the previous example (see item 1) is more ambiguous and less coherent than the simplification (see item 1) because the German noun "Anzeige" has several possible meanings, e.g., "display", "advertisement", or "complaint" and the date is relatively ("seitdem": "since then") but not absolutely specified ("nach 1999": "since 1999"). Without knowing the previous sentences, neither a human nor a machine could decide which meaning of a "Anzeige" would be correct. Hence, sentence simplification pairs with a high distance between coherence (simple) and coherence (original), or ambiguity (original) and ambiguity (simple) are much more difficult to correctly simplify than pairs with lower ratings regarding these aspects. Furthermore, if a document would contain many pairs with high distances regarding coherence and ambiguity, this would indicate that one should rely more on context-aware or document simplification model approaches than on isolated-sentence approaches.

Finally, we have added meaning preservation (similar to Alva-Manchego et al. 2020b), and information gain to the aspect collection (see Table 8.4) in order to evaluate how much original content is retained and how much new information is added. In the simplification of item 2, new information on the exact year and a complaint specification has been added. When using this example as evaluation data for sentence simplification, it is nearly impossible for a TS system to generate the correct addition without further contextual knowledge. Therefore, we also propose to manually assess the gold data regarding these aspects because these annotations are helpful to ensure a high quality of the training and evaluation data, and also reveal potential issues of the data. For example, gold evaluation data with high scores of information gain could be used to evaluate systems regarding hallucinations, or simplification pairs with improved coherence could be helpful to test context-aware TS systems.

Overall, wrt. RQ 3-2 and 3-3, we argue that our new selection of quality criteria, as well as the simplification operations, should be annotated on both the gold data and the system-generated data. First, the annotation of the gold data would help ensure a high quality of the training and evaluation data and would further give insights regarding what a text simplification can learn from the data and what kinds of simplification are expected in the references. Furthermore, the annotation of the system output should continue, since an automatic evaluation of the simplification is not (yet) reliable. Hence, manual evaluations (especially of the target group) are of the highest importance for the evaluation of a text simplification system (also see Section 8.5).

8.4 German Simplification Corpora

In Chapter 4, I have provided an extensive overview of resources for German text simplification (see step B and E in the TS workflow) and identified several issues with existing corpora, i.e.,

- Data Challenge A: focus on main domains of English TS such as web (see Section 4.1), wiki (see Section 4.2) and news data (see Section 4.3),
- Data Challenge B: they are not available due to copyright issues (e.g., Battisti et al. 2020),
- Data Challenge C: contain too much errors such as encoding problems (e.g., Kim et al. 2021),
- Data Challenge D: the size of previous corpora for German text simplification is very small (e.g., Klaper et al. 2013; Naderi et al. 2019; Mallinson et al. 2020),
- Data Challenge E: corpora are either available on the document or sentence level, but not both (e.g., Naderi et al. 2019; Rios et al. 2021),
- Data Challenge F: contain rather comparable than parallel data (e.g., Aumiller and Gertz 2022; Rios et al. 2021),
- Data Challenge G: are automatically aligned and hence of vague quality (e.g., Spring et al. 2021; Toborek et al. 2023),
- Data Challenge H: are designed with a greater focus on summarization than simplification (e.g., Aumiller and Gertz 2022; Rios et al. 2021),
- Data Challenge I: focus on highly simplified German variant "Leichte Sprache" (e.g., Siegel et al. 2019; Hansen-Schirra et al. 2021), or mixing texts of target groups (e.g., Toborek et al. 2023), and
- Data Challenge J: unclear quality and variety of the simplifications.

Furthermore, I have shown that some domains, such as news or web texts, are already covered in TS research to some extent. However, in this section, I discuss whether other domains are missing (see RQ 4-1 and Subsection 8.4.1) and if new data can help address the current challenges of the TS corpora (see RQ 4-2 and Subsection 8.4.2).

8.4.1 RQ 4-1: Missing Domains

In Chapter 4, I have introduced corpora of many domains, e.g., news, web, Wikipedia, medicine, narration, and law. However, Maaß (2020, p. 176) argues that domains such as texts concerning everyday life, culture, or education are currently not considered in automatic and manual text simplification (Data Challenge A).

In a more detailed analysis of the complexity of texts, i.e., Stodden (2021a) (see Subsection 7.4.1) we found that another relevant domain is missing (which answers RQ 4-1): People tend to inform themselves and discuss with others via digital tools and on the Web, e.g., on social media and user forums. Unlike other TS corpus domains, social media texts are not written by experts, but by the general public. Consequently, they differ more in terms of writing style, syntax, and vocabulary. Despite that, user-generated deliberative texts, such as in forums, discussion platforms, news comment sections, or social networks, are currently not considered in the field of automatic text simplification.

In Stodden (2021a) and Stodden and Nguyen (2024) (see Subsection 7.6.2), we discuss that readers of social media platforms could also benefit from automatic text simplification, as contributions on these platforms are often complex wrt. ungrammatical sentences (Bingel et al., 2018), contain out-of-vocabulary words (Baldwin et al., 2013), are reciprocal (Frieß et al., 2017), and often contain misinformation. Following our previously introduced definitions of complexity and comprehensibility, these aspects make a text difficult to read, e.g., for people with low literacy.

Unfortunately, we were unable to proceed further than the identification of this missing domain because no parallel complex-simple texts of these domains have been available. In future work, we plan to manually simplify user-generated texts in order to also cover this important domain.

8.4.2 RQ 4-2: New Data

Even though we could not build a new German TS corpus for the missing domain of user-generated texts, we have proposed new resources for German TS of other domains, i.e., news and web texts. In this section, we will answer how well our new corpora can address the other challenges previously introduced and how this overall improves current German text simplification (see RQ 4-2).

In more detail, we introduce two new German document simplification corpora, i.e., DEplain-APA-doc and DEplain-web-doc (see step B in Figure 8.5), and four new German sentence simplification corpora, i.e., DEplain-APA-sent, DEplain-web-sent-manual, and DEplain-web-sent-auto (with two different licenses) (see step E in Figure 8.5), which have all been proposed in Stodden et al. (2023) (see Subsection 7.4.2).

In the remainder of this section, I will further introduce the new DEplain corpora (see Subsubsection 8.4.2.1), then discuss how they can improve German text simplification by comparing them to each other and related corpora (see Subsubsection 8.4.2.2), as well as evaluate their characteristics in more detail wrt. the named challenges regarding data (see Subsubsection 8.4.2.3), and finally summarize their value for German TS (see Subsubsection 8.4.2.4).

8.4.2.1 Corpus Presentation – DEplain-APA & DEplain-web

In Table 8.5, I provide an overview of the metadata of the newly proposed subcorpora of DEplain in comparison to other German TS corpora. The main difference between the subcorpora, i.e., DEplain-APA and DEplain-web, is the domain of the texts included: The DEplain-APA corpora are based on news simplifications professionally simplified from CEFR level B1 to A2, while the DEplain-web corpora are based on harvested simplifications from the Web including narrative, language learner, health, bible, and public authority texts.

Although other corpora are often not available due to copyright issues (see Data Challenge B), we are allowed to share the DEplain-APA corpora upon request for academic purposes[4] and the main part of DEplain-web is openly available (see DEplain-web-doc, DEplain-web-sent-manual, and DEplain-web-sent-auto-open in Table 8.5)[5]. However, one part of DEplain-

4 To access the data, please send a request via zenodo: https://zenodo.org/records/8304430 [last update: August 31, 2023; last access: July 24, 2024].

Reference	Name	Target Simple	Domain	Available	# Docs	# Pairs	Aligned	Split
Klaper et al. (2013)	SGC '13	EL	web	on request	256	1,888	manual	*n/a*
Battisti et al. (2020)	SGC '20	CEFR A2	web	*n/a*	36	1,080	manual	train.+val.
Battisti et al. (2020)	SGC '20	CEFR A2	web	*n/a*	378	n/a	automatic	train.+val.
Toborek et al. (2023)	SGC '23	EL + PL	web	available	39	391	manual	train.+val.
Toborek et al. (2023)	SGC '23	EL + PL	web	available	700	5,942	automatic	train.+val.
Kim et al. (2021)	BiSECT	German learner	web	available	*n/a*	186,237	automatic	train.+val.
Siegel et al. (2019)	leichte-sprache-corpus	mixed	web	available	351	*n/a*	*n/a*	val.
Hansen-Schirra et al. (2021)	GEASY	EL	web	*n/a*	93	*n/a*	*n/a*	*n/a*
Jablotschkin et al. (2024)	DE-Lite	mixed	web	not yet	8,000	*n/a*	*n/a*	*n/a*
Spring et al. (2021)	Capito-B1	CEFR B1	web	*n/a*	12	426	manual	train.+val.
Spring et al. (2021)	Capito-A2	CEFR A2	web	*n/a*	8	412	manual	train.+val.
Spring et al. (2021)	Capito-A1	CEFR A1	web	*n/a*	22	416	manual	train.+val.
Spring et al. (2021)	Capito-B1	CEFR B1	web	*n/a*	1,055	54,224	automatic	train.+val.
Spring et al. (2021)	Capito-A2	CEFR A2	web	*n/a*	1,546	136,582	automatic	train.+val.
Spring et al. (2021)	Capito-A1	CEFR A1	web	*n/a*	839	10,952	automatic	train.+val.
Säuberli et al. (2024)	Capito-A2	CEFR A2	web	*n/a*	*n/a*	*n/a*	*n/a*	train.+val.
Naderi et al. (2019)	TextcomplexityDE	German learner	wiki	available	23	265	simplified	val.
Spring et al. (2023)	Wikipedia-Corpus	CEFR A2	wiki	*n/a*	198	1,382	manual	*n/a*
Ebling et al. (2022)	Wikipedia-Corpus	CEFR A2	wiki	*n/a*	106,126	*n/a*	automatic	*n/a*
Schlippe and Eichinger (2023)	Translated ASSET	*n/a*	wiki	*n/a*	*n/a*	1,000	simplified	train.+val.
Hewett and Stede (2021)	Lexica-klexikon	children 6-12	wiki	available	1,090	*n/a*	*n/a*	*n/a*
Hewett and Stede (2021)	Lexica-miniklexikon	children ≤ 6	wiki	available	1,090	*n/a*	*n/a*	*n/a*
Aumiller and Gertz (2022)	Klexikon	children 6-12	wiki	available	2,898	*n/a*	*n/a*	train.+val.
Mallinson et al. (2020)	geolino	children 5-7	wiki	available	20	1,198	simplified	val.
Säuberli et al. (2020)	German News Corpus	CEFR B1	news	*n/a*	*n/a*	3,916	automatic	train.+val.
Spring et al. (2023)	APA-LHA-OR-A2	CEFR A2	news	*n/a*	67	504	manual	n/a
Spring et al. (2023)	APA-LHA-OR-B1	CEFR B1	news	*n/a*	67	518	manual	n/a
Spring et al. (2021)	APA-LHA-OR-A2	CEFR A2	news	on request	2,300	9,456	automatic	train.+val.
Spring et al. (2021)	APA-LHA-OR-B1	CEFR B1	news	on request	2,300	10,268	automatic	train.+val.
Hewett (2023)	APA-RST	CEFR B1	news	available	25	128	manual	*n/a*
Hewett (2023)	APA-RST	CEFR A2	news	available	25	112	manual	*n/a*
Hewett (2023)	APA-RST	CEFR A2	news	available	25	153	manual	*n/a*
Rios et al. (2021)	20Minuten	general	news	available	18,305	*n/a*	*n/a*	train.+val.
Trienes et al. (2022)	simple-patho	laypeople	medical	not yet	850	3,280	simplified	train.+val.
Meister (2023)	ABGB-non-experts	laypeople	politics	available	1	448	manual	val.
Meister (2023)	ABGB-plain	PL	politics	available	1	448	manual	val.
Gutermuth (2020a)	Online Participation	PL	politics	available	1	13	simplified	*n/a*
Gutermuth (2020a)	Online Participation	EL	politics	available	1	13	simplified	*n/a*
Gutermuth (2020a)	Online Participation	EL + PL	politics	available	1	13	simplified	*n/a*
Schomacker et al. (2023a)	MILS+EB+PV+KV	mixed	narrative	available	33	*n/a*	*n/a*	train.+val.
Stodden et al. (2023)	DEplain-APA-doc	CEFR A2	news	on request	483	*n/a*	*n/a*	train.+val.
Stodden et al. (2023)	DEplain-web-doc	EL + PL	web	available	756	*n/a*	*n/a*	train.+val.
Stodden et al. (2023)	DEplain-APA-sent	CEFR A2	news	on request	483	13,122	manual	train.+val.
Stodden et al. (2023)	DEplain-web-sent-manual	EL + PL	web	available	147	1,846	manual	val.
Stodden et al. (2023)	DEplain-web-sent-auto-open	EL + PL	web	available	249	652	automatic	train.
Stodden et al. (2023)	DEplain-web-sent-auto-closed	EL + PL	web	reproducible	360	942	automatic	train.

Table 8.5: Summary of German document, paragraph, and sentence simplification corpora including own work (last part). The lines separate the domains of the corpora. EL = German Easy Language, PL = German Plain Language. All URLs have lastly been accessed at July 24, 2024. Extended version of Table 4.16.

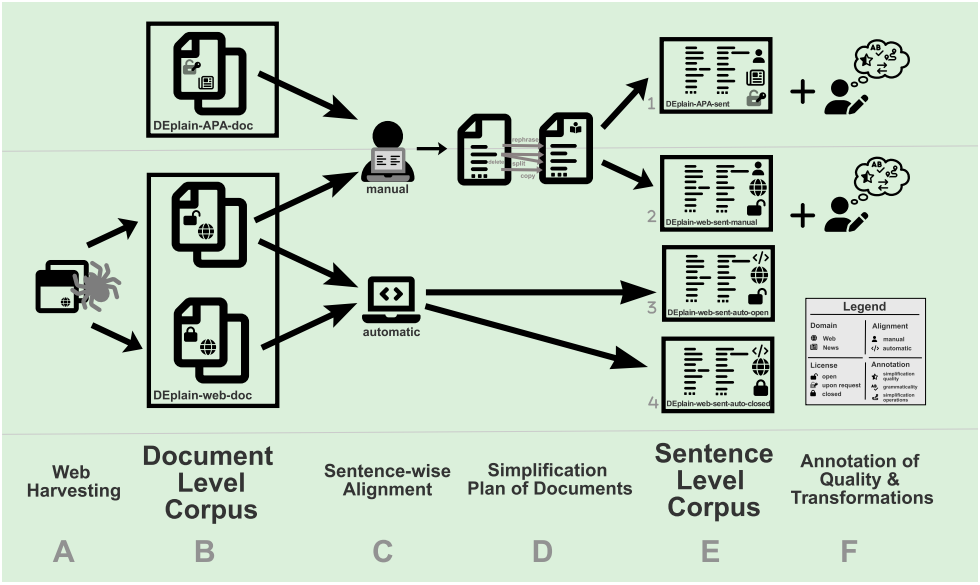


Figure 8.5: Corpus building process including the DEplain corpora.

web (see row 4) is not directly available but can be reproduced using our web harvester and MASSalign for automatic sentence-wise alignment.

Furthermore, each subcorpus is available on the document-level (see column # Document Pairs in Table 8.5 and step B in Figure 8.5), i.e., DEplain-APA-doc and DEplain-web-doc, as well as the sentence-level (see column # Sentence Pairs in Table 8.5 and step E in Figure 8.5), i.e., DEplain-APA-sent (see fourth last row in Table 8.5), DEplain-web-sent-manual (see third last row), DEplain-web-sent-auto-open (see second last row), and DEplain-web-sent-auto-closed (see last row). This characteristic allows us to solve Data Challenge E, i.e., the availability of corpora for document and sentence simplification that is based on the same source. Consequently, TS models can be trained and evaluated on the same resources (i.e., DEplain-APA or DEplain-web corpus), which may result in more accurate comparisons of the capabilities of models regarding sentence and document simplification. The results with the usage of the DEplain corpora would be more accurate than for models trained on other corpora because in our comparison the resources and models would only differ in text levels of the data (the effect of interest) and would be less influenced by different text characteristics, such as domain, writing style, or target group (unintended side effects).

In total, DEplain-APA-doc contains 483 document pairs from which all documents are manually aligned, resulting in 13,122 sentence pairs (see sixth last and fourth last row in Table 8.5). DEplain-web currently contains 756 parallel documents crawled from 11 web pages (see fifth, third, second and first last row in Table 8.5), covering 6 different domains (i.e., fictional texts (literature and fairy tales), bible texts, health-related texts, texts for language learners, texts for accessibility and public authority texts), and two language varieties (i.e., mainly German Plain

5 The DEplain-web data is available here: https://github.com/rstodden/DEPlain [last update: August 31, 2023; last access: July 24, 2024]

and Easy Language). The first three domains are not included in any other German TS corpus (see Data Challenge A).

As previously named in Subsubsection 8.3.1.3), we also address Data Challenge G and Building Challenge B regarding the lack of large sentence simplification corpora with only high-quality manual alignments. The whole DEplain-APA-sent corpus (see step E corpus 1 in Figure 8.5 or fourth last row in Table 8.5) and one subcorpus of the DEplain-web-sent corpus are manually aligned (i.e., DEplain-web-sent-manual; see step E corpus 2 or third to first last rows in Table 8.5), while the other corpora are automatically aligned (DEplain-web-sent-auto; see step E corpora 3 and 4 in Figure 8.5). For the DEplain-web corpora, we recommend using the manually aligned subset to test a TS model and the other subsets as augmented data when training a TS model (see Model Challenge A).

Furthermore, the texts of DEplain-APA address only one target group (i.e., foreign language learners), whereas the texts of DEplain-web address a mix of target groups, i.e., the German Plain Language and Easy Language target group, and foreign language learners. Hence, DEplain-APA reduces Data Challenge I regarding the lack of large corpora focusing on just one target group and domain.

8.4.2.2 Corpus Comparisons

A visual comparison of our DEplain corpora with other German TS corpora is provided in Figure 8.6. Compared to other existing corpora, our DEplain-APA corpus is currently the largest parallel corpus for German sentence simplification with only manual alignments and focusing on a single domain and a single target group, which solves Data Challenge D, F, G, and H. In contrast, DEplain-web is a high-quality test set for sentence simplification, which is manually sentence-wise aligned and contains a nearly balanced mix of sentence pairs of the narrative, language learning, bible, and health domains. In the remainder of this section, we will compare the DEplain corpora to related corpora within the domains, i.e., news domain and web domain.

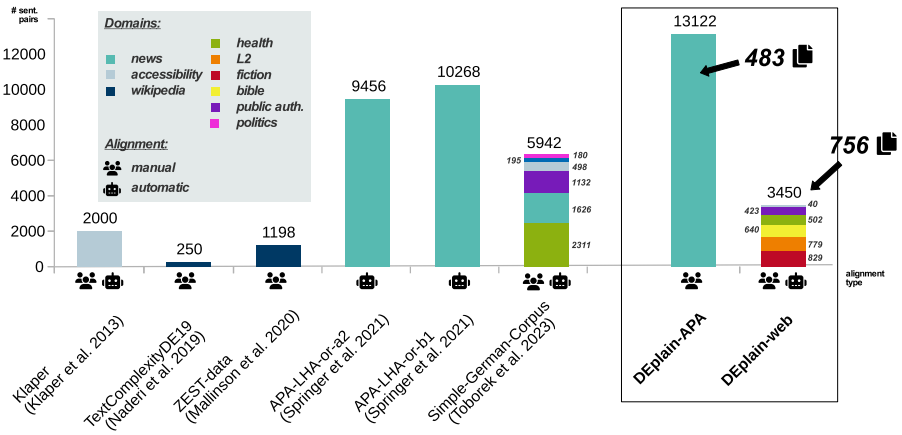


Figure 8.6: Comparison between previous corpora and our DEplain corpora (extension of Figure 4.1).

DEplain-APA vs. APA-LHA & APA-RST In this section, I am comparing our DEplain-APA to other German news corpora which are using the same news article resource, i.e., Austrian Press Agency (APA). In contrast to the APA-LHA corpora (Spring et al., 2021) (see Subsection 4.3.4 and Table 8.5), our DEplain-APA corpus is fully manually aligned, while their corpora are automatically aligned (including training, development, and test sets). Furthermore, our corpus contains milder simplifications than the APA-LHA corpus, although we are using the same resource: We have aligned the text on level B1 (as the source) and level A2 (as the target), whereas they use these two levels as target levels and use the original texts as the source.

Nevertheless, the data of the corpora overlap to some extent, because we and they are using the news articles of the same resource published in the same duration. However, in more detail, our DEplain-APA-doc documents are different from the ones in APA-LHA because the authors interpret one news item of a news article as one document, whereas we interpret the whole news article as one news item. On the one hand, using the whole news article may be more realistic than using only the news item, as that is how the data are published. On the other hand, a news article of APA can contain totally different topics (e.g., a mix of sports and politics), in which it might be difficult to identify the news ending, and hence, it might result in problems during automatic alignment.

As previously introduced (see Subsection 4.3.5 and Table 8.5), the APA-RST corpus (Hewett, 2023) entails manually aligned sentence pairs for foreign language learners with three different combinations of language levels, that is, OR-B1, OR-A2 and B1-A2. Compared with APA-RST, our corpus, both corpora have in common that they contain the full, original news articles (n=5), but their corpora additionally include the news items (n=25). Nevertheless, their chosen documents are not part of our corpus, as the publishing dates of their documents do not overlap with the time frame of our documents. However, their corpus is comparatively small with in total 5 documents, 25 news articles, and roughly 400 sentence pairs. Therefore, we can use their corpus as additional test data on the document and sentence level on the B1-A2 level or to test in-domain transfer regarding other language levels, e.g., using their OR-B1 and OR-A2 data as another test set.

DEplain-web vs. SGC '13 & SGC '20 & SGC '23 Comparing the available German web corpora with each other and DEplain-web, SGC '13 (also called Klaper corpus) (Klaper et al., 2013) (roughly 2,000 sentence pairs, see Subsection 4.1.3 and Table 8.5) is the smallest of all, then SGC '20 (Battisti et al., 2020) (1,080 sentence pairs, see Subsection 4.1.3 and Table 8.5), then DEplain-web (3,450 sentence pairs), and SGC '23 (Toborek et al., 2023) (see Subsection 4.1.4 and Table 8.5) is the largest with 5,942 sentence pairs. In more detail, all sentence pairs of SGC '20 are manually aligned, while only 1,846 sentence pairs (54%) of DEplain-web, and 500 sentence pairs of SGC '23 are manually aligned (8.4%).

All corpora contain a mix of web texts of different domains and webpages, Table 8.6 provides a comparison between the resources and their met data per German web TS corpora. While for SGC '20 it is unknown which sources have been used, similar resources have been used for SGC '23 and DEplain-web, e.g., both corpora include texts of Apotheken Umschau, city of Cologne, and Lebenshilfe Main Taunus. Compared more to the domains than to the sources, DEplain-web appears to be more balanced with respect to the included domains than SGC '23

(see colored bars in Figure 8.6): a third of their texts are health texts, other news texts, and the last third is a mix of several domains. In comparison, DEplain-web contains a similar amount of texts from 5 domains, i.e., fiction, language learner, bible, health, and public authorities.

As the usage of web texts is usually restricted by copyright, the corpora have in common that they make available their web crawler and not the data directly. But for DEplain-web and SGC '23 a small proportion of the corpora is available with open licenses. Further, only for DEplain-web are the full documents, the sentence-wise alignments, and the document plans available.

However, DEplain-web and SGC '23 have both in common that they include data for mixed target groups, i.e., people with learning disabilities, language learners and people with reading difficulties. Therefore, these corpora should be used with caution as they generalize the simplification process of different domains and target groups.

8.4.2.3 Quality Estimation of Corpus

In order to verify whether our DEplain corpora contain errors (see Data Challenge C) and to determine its quality and variety of simplifications (see Data Challenge J), in this section, I propose how to analyze DEplain regarding manual annotation of simplification quality and operations and discuss the results. With this analysis, we have completed the corpus building process (see step F in Figure 8.5), the building process could be repeated or the data could be filtered more fine-grained.

Application of Rating Aspects Bridging the gap between theory and practice, in our TS-ANNO annotation tool (Stodden and Kallmeyer, 2022), the evaluation aspects of our annotation schema (as described in Subsection 8.3.2) can be applied to evaluate the quality, representativeness, and suitability of gold simplifications, as well as the quality of system-generated simplifications. However, in the annotation tool, the annotation schema can be edited and specified for our own purposes, if our aspects might be too much or less fine-grained for other evaluation purposes. In order to check the suitability of the annotation schema in practice, we have partially annotated our DEplain corpora (Stodden et al., 2023) following the guidelines proposed in Stodden (2022). This includes annotation of the evaluation criteria as well as the simplification operations.

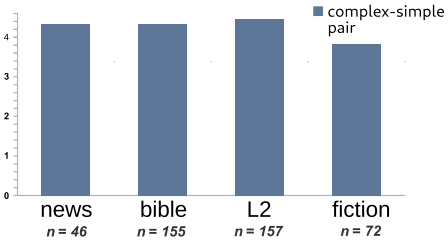
Through manual annotation, we have revealed that our gold data has some issues. Although most simple sentences preserve the meaning of complex sentences well (i.e., roughly 4.4 points on a scale ranging from 1 to 5, where 5 is best), in the fictional texts some information has been deleted (see Figure 8.7a). However, it might be intended by the authors of the fictional texts to shorten the often very long stories, but this is against the assumption of sentence-wise simplification in which only minor information is omitted.

Furthermore, manual annotation has revealed that some gold data in the DEplain-web corpus contain grammatical issues (i.e., data from the bible and the narrative domain), while other data seem to be of higher quality (see Figure 8.7b), e.g., the language learning and news texts are fluent in both the complex and simple versions. In contrast to the BiSECT corpus (see Subsection 4.1.5), the simplifications of all domains in the DEplain corpora contain nearly no gram-

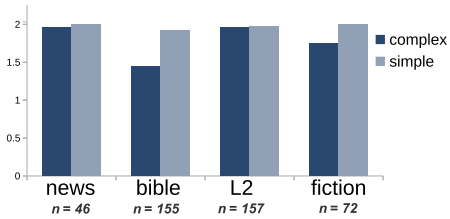
Subcorpus	Website Simple	Website Complex	Simple	Complex	Domain	Description	SGC '13	SGC '23	DEPlain-web
Alumniportal	alumniportal-deutschland.org †	alumniportal-deutschland.org	PL	PL	language learner	Texts related to Germany and German traditions written for language learners.			x
Apotheken Umschau	apotheken-umschau.de/einfache-sprache/ ‡	apotheken-umschau.de	PL	SG	health	Health magazine in which diseases are explained in PL		x	x
BZFE	bzfe.de/einfache-sprache/ †	bzfe.de	PL	SG	health	Information of the German Federal Agency for Food on good nutrition			x
Passanten Verlag	passanten-verlag.de/	projekt-gutenberg.org/	PL	SG/OG	fiction	Books in PL			x
Spaß Am Lesen Verlag	einfachebuecher.de/	projekt-gutenberg.org/	PL	SG/OG	fiction	Books in PL			x
Behinderten-beauftragter	behindertenbeauftragter.de/DE/LS	behindertenbeauftragter.de	EL	SG	accessibility	Official office for disabled people		x	
Bibel	offene-bibel.de/ †‡	offene-bibel.de/	EL	SG	bible	Bible texts in EL			x
brandeins	brandeins.de/themen/rubriken/leichte-sprache	–	EL	SG		Translating excerpts from various topics		x	
Einfach Teilhaben	einfach-teilhaben.de/DE/LS/	einfach-teilhaben.de	EL	SG	accessibility		x		x
Gemeinnützige Werkstätten und Wohnstätten Sindelfingen	gww-netz.de/de-LS/	gww-netz.de	EL		accessibility	Non-profit association in social sector	x		
Heilpädagogische Hilfe Osnabrück	*n/a*	os-hho.de	EL		accessibility	orthopaedagogical support	x		
Lebenshilfe	lebenshilfe-main-taunus.de	lebenshilfe-main-taunus.de	EL	SG	accessibility	Non-profit association for disabled people	x	x	x
MDR Nachrichten	mdr.de/nachrichten-leicht/	-	EL	SG	news	State-funded public broadcasting service		x	
Oberschwäbische Werkstätten	*n/a*	owb.de	EL		accessibility	Non-profit association in social sector	x		
Sozialpolitik	sozialpolitik.com/es	sozialpolitik.com/	EL	SG		Explains social policy in Germany		x	
Stadt Hamburg	hamburg.de/barrierefrei/leichte-sprache	hamburg.de	EL	SG	public authority	Information of and regarding the German city Hamburg			x
Stadt Köln	stadt-koeln.de/leben-in-koeln/soziales/informationen-leichter-sprache	stadt-koeln.de	EL	SG	public authority	Information of and regarding the German city Cologne		x	x
TAZ	taz.de/Politik/Deutschland/Leichte-Sprache/!p5097/	taz.de/	EL	SG	news	German Newspaper (discontinued)		x	

Table 8.6: Resources for German web TS corpora including own contributions (last column). The line separates German Plain (PL) and Easy Language (EL). OG = Old German, SG = Standard German. All URLs have lastly been accessed at July 24, 2024. Extended version of Table 4.2.

matical issues; hence, even if the original texts might contain issues, a TS system might learn to solve them simultaneously.



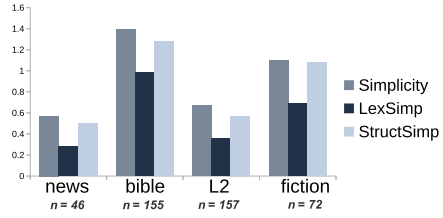
(**a**) Meaning preservation (min: 1; max: 5).



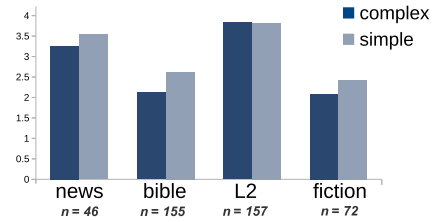
(**b**) Grammaticality (min: -2; max: +2).

Figure 8.7: Ratings of meaning preservation and grammaticality per domain.

We also found differences with respect to the types of simplification (see Figure 8.8a): the subcorpora with bible and narrative texts are generally more strongly simplified (wrt. overall simplicity, structural and lexical simplicity) than the subcorpora with news or language learner texts. Also, all complex-simple pairs contain more structural than lexical changes. Following this, DEplain-APA (as news corpus) and the DEplain-web subcorpus with language learner texts contain rather mild simplifications, whereas the other subcorpora of DEplain-web contain rather strong simplifications.



(**a**) Simplicity per type (min: -2; max: +2).



(**b**) Coherence (min: 1; max: 5).

Figure 8.8: Ratings of simplicity type and coherence per domain.

As shown in Figure 8.8b, the mild simplifications in the language learner and news subcorpus have higher coherence scores and hence are more self-contained than bible or fiction texts. However, on average, in all four subcorpora, the coherence has been increased during manual simplification, which is a hard task for TS systems due to missing context information (e.g., previous or preceding sentences). But, we also found that in 50 of the 430 annotated pairs, the coherence has been decreased during simplification, e.g., by adding conjunctions at the sentence beginning or replacing words with referential expressions. This is again an indicator to include more context into the simplification process, e.g., by adding previous and following sentences or instead simplifying on the paragraph or document level.

Overall, following this fine-grained analysis of the DEplain corpora (see also Subsection 8.2.1 and Stodden et al. 2023), we can state that both corpora are of high quality and include lexical as well as syntactical changes. More in detail, on the one hand, the DEplain-APA corpus is of higher quality than the DEplain-web corpus because of the higher average in

meaning preservation, grammaticality, and coherence ratings. On the other hand, the DEplain-web corpus contains more strong simplifications in terms of lexical and syntactical changes than DEplain-APA, and hence it could be a greater challenge to learn how to automatically simplify texts of DEplain-web than of DEplain-APA.

However, in the DEplain-web corpus, resources from many different domains and written for many different target groups are mixed; hence, optimizing a TS system on this data would rather end in a very general simplification system. By contrast, DEplain-APA only contains news texts aimed at language learners. This makes it a more specific and specialized corpus, enabling different strategies for automatic simplification than DEplain-web. Our findings can be used to choose simplification data based on interests regarding special simplification purposes, e.g., more lexical or syntactical simplification, only high-quality data, or combining all data in order to have a more huge dataset, or evaluating capabilities of a TS system in terms of coherence.

For a TS system, it might be easier to match the gold simplifications of DEplain-APA as they are more mild and hence closer to the original sentence, which is usually easier to predict for a TS system. In comparison, it is hard to achieve good automatic simplification scores on DEplain-web as it contains mostly strong simplifications. Hence, generated mild simplification might not be graded as good on DEplain-APA as on DEplain-web. In future research, we plan to manually simplify the complex sentences, including a mix of mild and strong simplifications, so that for each complex-simple pair more than one gold simplification (or reference) exists for a more robust evaluation. In addition, a manual evaluation of the system-generated sentences is planned regarding the proposed schemata.

Application of Simplification Operations We have further justified these theoretical findings of Stodden (2022) and Stodden (2021b), in Stodden et al. (2023) (see Subsection 7.4.2). Furthermore, in Stodden et al. (2023) we have also used the typology of simplification operations (see Subsection 8.2.1) to annotate simplification operations in our parallel sentence simplification subcorpora of DEplain-APA and DEplain-web. Based on these annotations, we have revealed, for example, that in our test set of the news corpus, the sentences have been more often reordered and more words have been added than in the test set of our web corpus. In contrast, in the web corpus, the simplification operation of rephrasing has been applied more often than in the news corpora. In both corpora, lexical substitution is the most prominent applied simplification operation (see Figure 8.9). Using this information, we can build expectations regarding the to-be-performed text simplification operations of a text simplification model and check whether the model has applied these changes or not.

The annotations provide a foundation for future work; e.g., our parallel corpus and its annotations of simplification operations can be used to build a sequence labeling system for German TS similar to existing English models (e.g., see Alva-Manchego et al. 2017 or Omelianchuk et al. 2021). In contrast to most sequence-to-sequence models, a sequence-labeling model would allow more control and interpretation of the generated simplification by using the applied simplification operations.

In addition, an existing corpus can be extended with the annotation of simplification operations, for example, i) to derive hand-crated rules for a TS system (Koptient et al., 2019), ii) to get more insights into the behaviors of metrics for automatic TS evaluation (e.g., see ASSET_{ann}

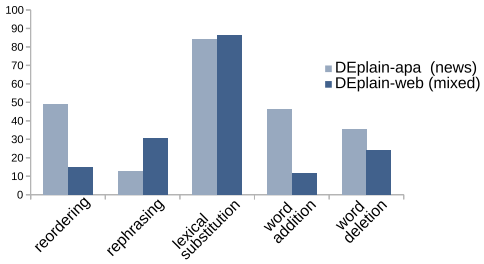


Figure 8.9: Simplification operations per simple-complex pair in DEplain-APA and DEplain-web (in %).

by Cardon et al. 2022), iii) to evaluate system outputs based on applied operations (e.g., see Yamaguchi et al. 2023), or iv) to train evaluation metrics including performed operations (e.g., see Heineman et al. 2023).

8.4.2.4 DEplain as Input for TS Systems

In order to answer RQ 4-2 regarding whether new data can improve German TS, our contribution of the new DEplain corpora has tackled many research gaps or challenges, e.g., the DEplain-APA corpus facilitates training or fine-tuning document and sentence simplification systems on news texts for foreign language learners. Although in recent years, the trend seems to be changing from fine-tuning to few-shot or zero-shot learning (Liu et al., 2023), our corpora can still be relevant for this research direction, because parts of the corpus can also be used as examples for few-shot learning, as well as serve as evaluation dataset (see Section 8.6). Furthermore, with DEplain-web we provide a new test set considering many domains and target groups. This corpus has the advantage that it can be split into smaller portions in order to evaluate against a special phenomenon, e.g., only lexical or syntactical simplification, only simplification into German Easy Language, or only simplification of narrative texts. However, comparative studies are necessary to verify whether the new data can help to increase the capacity of text simplification models. I will provide analysis in this direction in Section 8.6.

In addition, with our manual annotation we have verified the quality of our corpora and pointed out potential problems of the data. Referring back to RQ 3-2, we conclude that new corpora should be provided with insights (e.g., manual annotations) regarding their simplification operations and quality ratings in order to better understand the kind of simplifications of the new corpus and its quality. As shown for the DEplain corpus, the simplification operations, the extent of simplification, and the quality of the original data can highly vary wrt. domain and target group of the texts and, therefore, a data-driven TS system would learn different simplification operations.

To the best of my knowledge, our corpora are the first with manual annotations on the gold complex-simple pairs. Hence, we are unfortunately unable to make a direct comparison of these statistics with other corpora introduced in Chapter 4 due to the lack of annotations on the other corpora. It remains an open question whether other corpora include more unified or even more mixed data.

8.5 Text Simplification Evaluation

In addition to the research area of data for German text simplification, I have also further contributed to the research area of text simplification evaluation (see step H in TS workflow). As previously discussed in the state of the research in Chapter 5, and also shown in my publications (i.e., Stodden 2021c and Stodden 2024a), many challenges exist regarding the evaluation of text simplification, i.e.,

- Evaluation Challenge A: inconsistent scale design and varying scale interpretation in manual evaluation
- Evaluation Challenge B: missing components when intrinsically and extrinsically evaluating document and sentence simplification systems,
- Evaluation Challenge C: automatic metrics are not reliable,
- Evaluation Challenge D: automatic metrics are not interpretable and explainable,
- Evaluation Challenge E: the scores and settings used during automatic evaluation are only evaluated on English TS and are varying across TS studies, and
- Evaluation Challenge F: only less evaluation sets are available and mostly contain only one gold simplification.

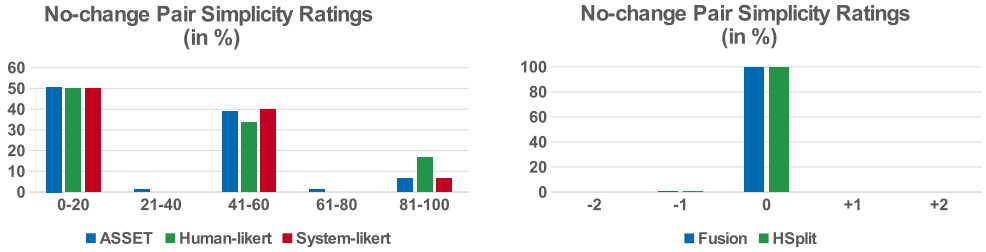
In addition, there is disunity in manual evaluation approaches between the high effort to design and execute high-quality extrinsic evaluation and the more easy, faster, but more dirty intrinsic evaluation approach.

Hence, based on these findings, we can already deny RQ 5-1 as manual and automatic TS evaluation seems to be neither reliable nor robust (see Evaluation Challenge A, B, C, D, E). In order to counteract, the evaluation protocol of text simplification requires an update. In the following, I will discuss new approaches to improve the TS evaluation by answering RQ 5-3. In addition, most evaluation approaches are designed for English and have not been evaluated for other languages. Therefore, I will also provide solutions to counteract this (see RQ 5-2).

8.5.1 RQ 5-1: Robust Evaluation

With respect to RQ 5-1 and Evaluation Challenge A, in Stodden (2021c) (see Subsection 7.5.1), we have found that the current design of manual evaluation studies is not sufficiently clear for annotators who are not trained or are not experts in the field. In more detail, we have compared the ratings regarding simplicity of annotators on five English test sets, three of them with a rating scale ranging from 1 to 100 (without a neutral element) and two with a scale ranging from -2 to +2 (including a neutral element, i.e., 0). In order to exclude biases regarding subjective simplicity estimation, we have evaluated only sentence pairs without any change between complex and simple sentences. Hence, the expected rating should express that the complexity of the simplified sentence has not been changed compared to the original sentence, i.e., rating of the neutral element 0 and 1 in the 1-100 scale. However, as visualized in Figure 8.10a, the simplicity rating scales with scale endpoints at 1 to 100 are ambiguous because participants have not properly understood how to rate complex-simple pairs with the same complexity (1 or 50?).

We also found that expert annotators, who have rated on a scale of -2 to +2 with a distinct neutral element, follow the same scale understanding and annotate complex-simple pairs with the same complexity mainly with the neutral scale element of 0 (see Figure 8.10b).



(**a**) Rating scale from 1 to 100.

(**b**) Rating scale from -2 to +2

Figure 8.10: Simplicity ratings in five English test sets with different rating scales.

Based on these findings, we have updated the scale points for our manual evaluation schema to scales with neutral elements (see Subsection 8.3.2 and Table 8.4). However, best practices on how to manually evaluate text simplification are required as already existing for related tasks, e.g., text summarization (Iskender et al., 2020) or machine translation (Freitag et al., 2021), or in general for NLG (van der Lee et al., 2019). Popović et al. (2022) recommend naming the details of the manual evaluation conducted to make the procedure more transparent for reproduction, as well as for building best practices based on previous research.

Unfortunately, current suitability studies of automatic metrics are based on human assessments following the error-prone evaluation schema (e.g., in Alva-Manchego et al. 2021, Scialom et al. 2021, Maddela et al. 2023, or Zhao et al. 2023). Following this, it is unclear whether the correlations between these assessments and the automatic metrics would hold when using unambiguous ratings. A re-evaluation is necessary to ensure the suitability of automatic metrics for English TS. Also, a reproduction study with ratings for other languages than English is required to ensure the suitability of the metrics also for non-English languages. In future work, our manual annotations on DEplain on evaluation aspects (see Subsubsection 8.4.2.3) could be used to check the reliability of automatic metrics for German (Evaluation Challenge C). With TS-ANNO (see Stodden and Kallmeyer 2022) the number of ratings could easily be increased and also be enriched with ratings of system outputs and not (as currently) only gold data.

8.5.2 RQ 5-2: Multi-lingual Evaluation

As mentioned above, most automatic evaluation metrics are designed and verified only for English (see Evaluation Challenge E). Therefore, it is difficult to determine whether the same metrics can also be applied to other languages than English (see RQ 5-2). For some of the metrics, there are also variants of other languages, e.g. the German variant of FRE (Amstad, 1978), multi-lingual or language-dependent BERT models for BERTScore (Zhang* et al., 2020), or multi-lingual feature extraction for readability assessment (e.g., see Lee and Lee 2023 or Stodden and Kallmeyer 2020). Due to missing alternatives, multi-lingual TS approaches of approaches in a language other than English (e.g., for Spanish Gonzalez-Dios et al. 2022, for French Cardon and Grabar 2020, for Swedish Holmer and Rennes 2023, or for multi-lingual

	Tok.	Lang.	BLEU↑	SARI↑	BS-P↑	FRE↑
TCDE19 (n = 250)	spacy	EN	**28.22**	15.31	0.37	**39.16**
	spacy	DE	27.31	14.99	**0.55**	28.1
	13a	DE	27.49	15.05	**0.55**	28.0
	none	DE	24.43	13.78	**0.55**	28.1
DEplain-APA (n = 1231)	spacy	EN	**29.28**	**16.17**	0.45	**77.64**
	spacy	DE	26.89	15.25	**0.63**	58.75
	13a	DE	27.25	15.35	**0.63**	64.6
	none	DE	23.33	13.75	**0.63**	58.75
DEplain-web (n = 1846)	spacy	EN	21.24	12.09	0.25	**70.33**
	spacy	DE	20.85	11.93	0.42	62.95
	13a	DE	20.89	11.94	0.42	62.95
	none	DE	18.82	10.9	0.42	62.95

Table 8.7: Scores of identity baseline on three German test sets when using different language settings and tokenizers. Copied from Stodden (2024a).

TS Ryan et al. 2023) are currently evaluated with a toolkit designed for English TS evaluation (i.e., EASSE) regardless of the named variants for other languages.

To ensure and facilitate that language-specific or multi-lingual metrics (if applicable) are used in the TS evaluation of languages other than English, in Stodden (2024a) (see Subsection 7.5.3), we have proposed a framework for sentence simplification evaluation called EASSE-multi. EASSE-multi extends EASSE (Alva-Manchego et al., 2019a) (see Subsection 5.2.4) by supporting TS evaluation in multiple languages, e.g., by tokenization in many languages, using multi-lingual BERT for BERTScore, or using language-specific readability metrics. This ensures a more transparent and easier comparison between existing and prospective German TS models as well as TS models of other languages. The results of non-English TS studies, which currently rely on EASSE for evaluation, could be made more robust by evaluating with EASSE-multi than with the original EASSE.

In addition to Tanprasert and Kauchak (2021), who have shown how to easily manipulate FKGL scores by simple post-processing, in Stodden (2024a), we have shown how to easily manipulate BLEU and SARI, when changing the tokenization method. For example, using a German SpaCy model for tokenization lowers the BLEU and SARI scores compared to tokenization with an English SpaCy model or white space tokenization (see Table 8.7). To exclude effects of the test set or the TS model used, we have evaluated the identity baseline (i.e., just copying the original sentence without simplification) on three different test sets (i.e., TextComplexityDE19 (see Subsection 4.2.6), DEplain-APA, and DEplain-web). Furthermore, the first two rows of each block in Table 8.7 also highlight the differences in the BERTScore (see BS-P) and FRE scores when using different language specifications in EASSE-multi.

Based on these results, I can answer RQ 5-2: the evaluation strategies for English TS have to be adapted when evaluating on other languages, e.g., tokenizer or language-dependent metrics. Therefore, I propose to pay attention to the evaluation of comparative TS models that all rely on the same evaluation settings, e.g., regarding tokenizer, case sensitivity, or BERT model for BERTScore. This procedure is essential in order to ensure that the differences in the scores are due to changes in the TS model and not in the evaluation metric. However, even if most of the scores are language independent or can be easily adapted to work for other languages, there may still be problems in using the same scores for different languages due to language idiosyn-

crasies and different simplification operations per language (see Stodden and Kallmeyer 2020). Approaches in the direction of language-wise evaluation of non-English TS could be learnable metrics (per language) as already proposed for English, e.g., LENS, BETS, or MeaningBERT, or complexity assessment models such as participating systems of the complexity prediction shared task (Mohtaj et al., 2022).

8.5.3 RQ 5-3: New Aspects for Evaluation

As discussed previously, in Subsection 8.3.2, the current common manual TS evaluation aspects do not comprise the full spectrum of text simplification evaluation; therefore, we proposed additional evaluation aspects. However, there is also some room for improvement in automatic evaluation metrics for (German) TS, especially with respect to interpretability (see Evaluation Challenge D). In the remainder of this section, I will discuss wrt. RQ 5-3, how I have contributed in the direction of a better automatic evaluation. With my work, I have focused on a few specific aspects of automatic evaluation that could be enhanced, i.e., complexity evaluation (see Subsubsection 8.5.3.1), evaluation of syntactic simplification (see Subsubsection 8.5.3.2), and evaluation using linguistic features (see Subsubsection 8.5.3.3).

8.5.3.1 Complexity Prediction

As previously discussed in Subsection 8.2.2, the prediction of the complexity of original and simplified sentences could be measured with the systems developed in the scope of the shared task of complexity prediction by Mohtaj et al. (2022). For example, our approach (Arps et al., 2022) (see Subsection 7.5.2) could be further evaluated regarding the capacity for complexity prediction of system-generated simplifications.

In future work, I plan to investigate how sentence complexity in the scope of German TS can be measured using approaches of German text readability assessments (e.g., by Weiss and Meurers 2022, Klepp 2022a, or Thome et al. (2024)).

8.5.3.2 Syntactic Simplification

Another option on how to automatically evaluate structural simplicity beyond SAMSA (Sulem et al., 2018b) is based on the number of statements in simplifications generated by the systems. Following the recommendations of Easy German Language (e.g., Deutsches Institut für Normung (DIN) 2023), each simplified sentence should contain just a few statements. In order to verify whether a TS system fulfills this criterion, the number of statements per sentence in the system output could be counted. In contrast to other TS metrics, this method could also be applied to the document level. If more than one statement exists, each statement could be labeled for a better interpretability of the measurement. On the other hand, the statement identification and segmentation could also be used as a first step for sentence splitting as one prominent task of syntactical simplification.

In order to facilitate the development of statement identification and segmentation, we have recently organized a shared task (Schomacker et al., 2024)) (see Subsection 7.5.4), co-located with KONVENS (Konferenz zur Verarbeitung natürlicher Sprache / Conference on Natural

Language Processing) 2024. For this purpose, we have developed annotation guidelines regarding how to count statements in German sentences and applied them to German Wikipedia-like articles written in German Easy language by non-trained translators.[6] In total, we have annotated 4,282 sentences with their number and segments of statements (2,988 sentences for training, 416 for validation, and 878 for testing) where more than 50% of the sentences contain just one statement[7].

Overall, three teams have participated, from which two have contributed to automatically counting and segmenting the statements, and one team has contributed only to the counting subtask. As can be seen in Table 8.8, all systems could beat our baselines (i.e., labeling all sentences with 1, and randomly labeling the sentences) in the subtask for counting the statements. We believe that the results of this shared task can contribute to automatically assessing the structural complexity of a (automatically or manually) generated simplification. In future work, we will further investigate how the resources of this shared task can be used to automatically split sentences with more than one statement into several sentences with one statement each.

Team	MAE
KlarTextCoder	0.35
StaGE FriGHt	0.4
CUET_Big_O	0.4
Baseline-random	0.66
Baseline-all-1	0.82

Table 8.8: Results of StaGE shared task subtask 1. Copied from Schomacker et al. (2024)).

8.5.3.3 Linguistic Features Beyond Simplicity

Furthermore, following the recommendations of Alva-Manchego et al. (2019a), Tanprasert and Kauchak (2021), and Cardon and Bibal (2023), it might be useful to include annotations of simplification operations or linguistic features in the quality analysis of TS systems because they can help to understand how a complex sentence has been transformed into a simple sentence and to some extent give insight into black-box TS models.

Therefore, in Stodden and Kallmeyer (2020) (see Subsection 7.2.1), we have proposed a multi-lingual feature extraction toolkit to get more information about the linguistic changes of system simplifications compared to the original texts. TS-eval-multi is an extension of the reference-less quality estimation tool for English (also called TS-eval) by (Martin et al., 2018). In comparison to a similar feature extraction toolkit called LFTK (Lee and Lee, 2023), the majority of the features in TS-eval-multi are applicable to multiple languages and not only to English. Furthermore, TS-eval-multi focuses more on features for text simplification, whereas LFTK focuses more on features for readability assessment. The TS-eval-multi package is also integrated into the EASSE-multi evaluation package (Stodden, 2024a) (see Subsection 7.5.3) to facilitate easy evaluation of linguistic features and not only of questionable current automatic metrics. In

6 The annotation guidelines are available here: https://german-easy-to-read.github.io/statements/annotations/ [last access: July 24, 2024].

7 The annotated data is available online: https://github.com/german-easy-to-read/statements [last update: July 10, 2024; last access: July 24, 2024]

future work, this linguistic analysis could be extended with features of LFTK, automatic annotation of simplification operations as proposed in Cardon and Bibal (2023) (potentially trained or evaluated on our manually annotated simplification operations; see Section 8.2) or automatic error annotation as proposed in Yamaguchi et al. (2023) (potentially trained or evaluated on our manually annotated errors; see Lemgen 2024).

TS-eval-multi partially includes features which are also relevant for the evaluation of document simplification, e.g., compression ratio, proportion of deleted or added words, or average phrase length per sentence. In future work, it could be investigated whether more discourse-level features, e.g., logical argumentation and idea density (Collins-Thompson, 2014), can help to automatically estimate the coherence of a text and how to integrate them into TS-eval-multi. The degree of repetition of words and ideas across sentences ("referential cohesion" Graesser et al. 2014), number of connectives ("deep cohesion" Graesser et al., 2014) and the amount of standard language ("degree of narrativity" Graesser et al. 2014) could be further helpful for domains different from the usual domains of TS research, e.g., deliberative texts (see Subsection 8.4.1).

8.5.4 Recommendations

In order to answer RQ 5-3, I will summarize the previous discussion and give recommendations regarding manual (see Subsubsection 8.5.4.1) and automatic evaluation for German TS (see Subsubsection 8.5.4.2).

8.5.4.1 Manual Evaluation

For intrinsic manual evaluation aspects I recommend to consider the following aspects: simplicity, grammaticality, coherence, ambiguity, lexical simplicity, structural simplicity, overall simplicity, meaning preservation and information gain (see Subsection 8.3.2 and Evaluation Challenge B). For the first four, I suggest to evaluate the aspects on the original as well as the simplified text in order to see the changes that are made by a TS system or a professional translator. Especially for German simplification, information gain seems to be a more relevant criterion than in other languages because during simplification into German Easy Language implicit content is frequently made more explicit and explanations or examples of complex terms are more often added than for other simplification purposes (see Subsection 8.2.1).

In order to address Evaluation Challenge B, I recommend using Likert scales with a statement as aspect description and an unequal number of points on the scale. I have shown that a scale with negative and positive scale points can help to improve the understanding of annotators regarding how to rate, e.g., more complex and even complex sentence pairs (see Subsection 8.5.1). This evaluation using the (partially) new rating aspects and the suggested scale can easily be conducted using an annotation tool such as our TS-ANNO.

Regarding the group of annotators, I suggest asking native speakers with high reading abilities to rate the aspects regarding grammaticality, coherence, ambiguity, meaning preservation, and information gains because for these aspects it is required to understand both texts well. The remaining aspects, i.e., simplicity, lexical simplicity, structural simplicity, and overall simplicity, should be rated directly by people of the target group because the texts are simplified for

them; hence, they should be readable for them too. On the other hand, extrinsic evaluation such as conducted in Säuberli et al. (2024) (see Subsubsection 5.3.1.2) seems to be an even more reliable strategy in order to estimate the comprehensibility of the simplified texts and is also well suited for the evaluation of documents, but requires also even more time in preparation (e.g., generation of comprehension questions per text).

8.5.4.2 Automatic Evaluation

With respect to automatic evaluation, we have added a few new components to the evaluation protocol in order to answer RQ 5-3. First, I recommend that one still use automatic metrics such as SARI, BLEU, FRE, or BERTScore for the evaluation of German TS systems, although they are not fully robust and reliable. But I suggest using suitable versions of the metrics for German, i.e., either a German or multi-lingual BERT model for BERTScore, the German version of FRE, and preprocess the data with a German tokenizer (see Subsection 8.5.2). All these aspects are considered in our EASSE-multi evaluation framework, which facilitates the evaluation of non-English TS. Furthermore, I recommend specifying the settings that have been used when evaluating a TS system to make the evaluation more comparable to other TS studies (see Evaluation Challenge E).

In order to make the automatic evaluation more interpretable and reliable (see Evaluation Challenge C and D), we have proposed to evaluate the complexity of the simplified sentences with complexity prediction models instead of the evaluation with FRE. Additionally, including the change of linguistic features between the original and the simplified text can help to get more insight into the simplification process, e.g., what has been correctly or wrongly changed. Our TS-eval-multi package assists in gaining this information (see Subsubsection 8.5.3.3). For the purpose of including automatic syntactic evaluation in the evaluation approach of German TS, SAMSA might be an option if a reliable semantic parser is available. In addition, I recommend considering the number of statements within a simplified sentence to evaluate syntactical changes (see Subsubsection 8.5.3.2), which is also applicable to the evaluation of document simplification. Overall, more work on German document simplification and especially its evaluation is required, as the manual and automatic evaluation protocol is currently mostly addressed to sentence simplification.

Finally, I recommend checking the list of available test sets (see Table 8.9) to potentially find suitable data for one's own simplification purposes. Using existing test sets for one's own evaluation is beneficial for a better comparison between existing TS models (see Evaluation Challenge F and Model Challenge D). In future work, it would be especially interesting to see an evaluation on a German test set with more than one reference, i.e., ABGB.

8.6 German Text Simplification Models

Finally, I discuss the contributions of me and my co-authors regarding German text simplification models (see step G in the TS workflow) and their impact (see steps I and J in the TS workflow). In our survey on the state of research regarding German TS models in Chapter 6, we have presented 25 models from which 11 are applied to document simplification (e.g., fine-

Name	Reference	Target Group	Domain	Size	# Ref.	*n*:*m*	complex FRE↓	complex Sent. Len.↑	complex Word Len.↑	simple FRE↑	simple Sent. Len.↓	simple Word Len.↓
ABGB	Meister (2023)	non-experts	law	448	2	40%	42.75	24.85	1.83	44.6	22.39	1.89
APA_LHA-OR-A2	Spring et al. (2021)	lang. learners	news	500	1	6%	44.7	20.2	1.92	69.55	11.27	1.78
APA_LHA-OR-B1	Spring et al. (2021)	lang. learners	news	500	1	8%	43.7	20.48	1.93	62.6	12.82	1.83
BiSECT	Kim et al. (2021)	people w. reading problems	politics	753	1	100%	**8.55**	**30.24**	2.01	35.85	15.72	1.98
DEplain-APA	Stodden et al. (2023)	lang. learners	news	1,231	1	27%	58.75	11.92	1.86	65.8	10.55	1.79
DEplain-web	Stodden et al. (2023)	mixed	web/mixed	1,846	1	57%	62.95	19.13	1.64	**77.9**	10.76	**1.57**
GEOlino	Mallinson et al. (2020)	children	encyclopedia	663	1	40%	61.5	13.31	1.7	66.0	9.94	1.66
SGC '23	Toborek et al. (2023)	mixed	web/mixed	391	1	73%	41.15	13.96	2.0	65.4	**9.31**	1.83
TextComplexityDE	Naderi et al. (2019)	lang. learners	encyclopedia	250	1	83%	28.1	27.75	**2.08**	51.2	14.17	1.9

Table 8.9: Overview of test sets for German sentence simplification which are included in EASSE-DE. Extended version of Table 1 in Stodden (2024a).

tuning pre-trained autoregressive models, or fine-tuning sequence-to-sequence models), 1 to paragraph simplification, and the remaining 13 to sentence simplification (e.g., fine-tuning pre-trained sequence-to-sequence models, or prompting pre-trained models).

However, I also introduced many issues regarding the existing TS models, i.e.:

- Model Challenge A: only a few studies experiment with augmented data (e.g., Säuberli et al. 2020; Schlippe and Eichinger 2023),
- Model Challenge B: lack of comparisons between TS system approaches, e.g., regarding different test sets, domains, or text levels,
- Model Challenge C: mixing of text of different domains & target groups in training and evaluation of TS models,
- Model Challenge D: existing TS models are not comparable to each other due to different test sets or evaluation methods,
- Model Challenge E: existing TS models are not reproducible due to missing details of the approaches,
- Model Challenge F: the TS models are trained on corpora with different sizes and quality, and
- Model Challenge G: the target group is often overlooked when building TS models.

In my publications, I have addressed some of these challenges as I will further explain in this section. Overall, we have introduced five new text simplification models in Stodden et al. (2023) (see Subsection 7.4.2), three for the document level, i.e., mbart-DEplain-doc-APA, mbart-DEplain-doc-web, and mbart-DEplain-doc-APA+web, and two for the sentence level, i.e., mbart-DEplain-sent-APA, and mbart-DEplain-sent-APA+web. We have shown that the same approach (with similar settings), i.e., fine-tuning mBART with a trimmed vocabulary of 30,000 German words, works for both simplification levels (see Model Challenge B). In this work, we have proposed the first sentence simplification system that is fine-tuned only on manually aligned sentence pairs of one domain addressed to one target group, i.e.,

DEplain-sent-APA on news texts for German language learners (see Model Challenge F). Furthermore, in Stodden (2024b) (see Subsection 7.6.1), we have reproduced several German sentence simplification systems and proposed three new models, i.e., sockeye-DEplain-APA, mT5-DEplain-APA, and mT5-SGC. For a comparison of the metadata of previous and our newly proposed models see Table 8.10.

8.6.1 RQ 6-1: Document & Sentence Simplification

In previous research, there has been a strict separation of document and sentence simplification (see RQ 6-1 and Model Challenge B), although for some corpora, parallel texts are available for the document and sentence level, e.g. APA-RST (Hewett, 2023), Simple German Corpus '23 (Toborek et al., 2023), or Newsela for English (Xu et al., 2015). Cripwell et al. (2023b) bridge the gap between both research directions by introducing simplification plans for English TS: Each sentence of a source document gets labeled with a simplification operation (i.e., copy, rephrase, split, or delete) corresponding to the parallel target document. In this scenario, the simplification process is split into two steps, i.e., classification of simplification operation per sentence and sentence simplification with a control token specifying the simplification operation, which improves simplification wrt. SARI and FKGL in comparison to a standard sentence or document sequence-to-sequence TS model.

However, about the same time as Cripwell et al. (2023b), in Stodden and Kallmeyer (2022) and Stodden et al. (2023) we have also proposed simplification plans for documents[8]. In contrast to them, our simplification plans are not based on automatic alignments but on manual alignments (e.g., copy, rephrase, split, or fusion) or left-over alignments (e.g., deletions and additions) including more simplification operations than them (i.e., fusion and addition). We argue that it is possible to bridge the gap between sentence and document simplification with our manually annotated simplification plans (see RQ 6-1 and Model Challenge B).

In future work, the automatic labeling approach of Cripwell et al. (2023b) could be evaluated on our manual simplification plans, and similar experiments using their two-step simplification could be performed using our DEplain-data. The simplification plans could also improve context-aware sentence simplification similar to Sun et al. (2020) since, based on the sorted order of the aligned pairs following their occurrence in the original document, the preceding and following sentences of a complex-simple sentence pair can be easily picked up and included in a TS system.

8.6.2 RQ 6-2: Effect of Data & Models

In this work, we are focusing on sentence and document simplification. Nevertheless, I assume that with some effort our sentence-wise aligned corpora could be adapted to paragraph simple-complex pairs due to the available simplification plans of the documents. Further experiments on the paragraph level could be interesting in tackling the problem of too long inputs in document simplification, but too little context in sentence simplification.

8 Also see step D in the corpus building process in Figure 8.5 and https://github.com/rstodden/DEPlain/tree/main/D__Simplification_Plans [last update: August 31, 2023; last access: July 24, 2024].

System Name	Reference	Type	Level	Domain	Target Group / Language Variety	Training Data	# Simp. Pairs	URL
-	Suter et al. (2016)	rule-based	sent	*n/a*	*n/a*	-	-	-
HDA-ETR	Siegel et al. (2019)	rule-based	sent	*n/a*	German Easy Language	-	-	github.com/hdaSprachtechnologie/easy-to-understand_language
DISSIM	Niklaus et al. (2019a)	rule-based	sent	*n/a*	*n/a*	-	-	github.com/Lambda-3/DiscourseSimplification
Sockeye-Benchmarking	Säuberli et al. (2020)	seq2seq	sent	mixed	German language learners	APA-LHA OR-B1	3,316	-
Sockeye-APA-LHA	Spring et al. (2021) & Ebling et al. (2022)	seq2seq	sent	news	German language learners	APA-LHA OR-A2 & APA-LHA OR-B1	9,456 & 10,268	github.com/ZurichNLP/RANLP2021-German-ATS
mBART-20min	Rios et al. (2021)	fine-tuned seq2seq	doc	news	*n/a*	20Minuten	17,905	github.com/a-rios/longmbart
mBART-APA+capito	Rios et al. (2021)	fine-tuned seq2seq	doc	mixed	German language learners	capito-A1-A2-B1 & APA-LHA OR-A2 & APA-LHA OR-B1	3,424 & 2,250 & 2,302	github.com/a-rios/longmbart
mBART-SimplePatho	Trienes et al. (2022)	fine-tuned seq2seq	par	medical	laypeople	Simple-Patho	3,280	github.com/jantrienes/simple-patho
mBART-capito	Säuberli et al. (2024)	fine-tuned seq2seq	doc	mixed	German language learners	Capito Corpus	*n/a*	-
mBART-GNATS	Schomacker et al. (2023a)	fine-tuned seq2seq	doc	narration	Plain German & German Easy Language	GNATS	28	github.com/tschomacker/aligned-narrative-documents
mT5-MultiSim	Ryan et al. (2023)	fine-tuned seq2seq	sent	mixed	mixed	MultiSim	653,468	github.com/XenonMolecule/MultiSim
flan-T5-translated	Schlippe and Eichinger (2023)	fine-tuned seq2seq	sent	wikipedia	*n/a*	translated ASSET	1,000	-
ZEST	Mallinson et al. (2020)	zero-shot	sent	wikipedia	*n/a*	WikiAuto & WMT19	300*k* & 6.0*mio*	github.com/Jmallins/ZEST
GUTS	Fruth et al. (2024)	unsupervised	par	-	-	-	-	github.com/LFruth/unsupervised-german-ts
BLOOM-zero	Ryan et al. (2023)	prompting	sent	-	-	-	-	github.com/XenonMolecule/MultiSim
BLOOM-sim-10	Ryan et al. (2023)	prompting	sent	news & knowledge acquisition & wikipedia	mixed	GermanNews & TextComplexityDE & GEOlino	10	github.com/XenonMolecule/MultiSim
BLOOM-random 10	Ryan et al. (2023)	prompting	sent	news & knowledge acquisition & wikipedia	mixed	GermanNews & TextComplexityDE & GEOlino	10	github.com/XenonMolecule/MultiSim
BLOOM-BiSECT	Ponce et al. (2024)	prompting	sent	-	-	*n/a*	-	-
ChatGPT-multilingual	Schlippe and Eichinger (2023)	prompting	sent	-	-	-	-	-
-	Deilen et al. (2023)	prompting	doc	-	German Easy Language	-	-	-

Table 8.10: Summary of German TS models including own work. Each line separates different model approaches. Extended version of Stodden (2024b) and Table 6.3. All URLs have lastly been accessed at July 24, 2024. Part I (continued on next page).

System Name	Reference	Type	Level	Domain	Target Group / Language Variety	Training Data	# Simp. Pairs	URL
custom-decoder-ats-gerpt2	Anschütz et al. (2023)	AR model + fine-tuned seq2seq	doc	mixed & news	Plain German & German Easy Language & *n/a*	Simplified, monolingual German data & 20Minuten	544,467 & 17,905	github.com/MiriUll/Language-Models-German-Simplification
custom-decoder-ats-german-gpt	Anschütz et al. (2023)	AR model + fine-tuned seq2seq	doc	mixed & news	mixed	Simplified, monolingual German data & 20Minuten	544,467 & 17,905	github.com/MiriUll/Language-Models-German-Simplification
gpt2-wechsel-german	Klöser et al. (2024)	fine-tuned AR model	doc	mixed	Plain German & German Easy Language	Semi-synthetic Simple German Web Corpus	7,130	github.com/MSLars/German-Text-Simplification
gpt2-xl-wechsel-german	Klöser et al. (2024)	fine-tuned AR model	doc	mixed	Plain German & German Easy Language	Semi-synthetic Simple German Web Corpus	7,130	github.com/MSLars/German-Text-Simplification
leo-hessianai-7b	Klöser et al. (2024)	fine-tuned AR model	doc	mixed	Plain German & German Easy Language	Semi-synthetic Simple German Web Corpus	7,130	github.com/MSLars/German-Text-Simplification
leo-hessianai-13b	Klöser et al. (2024)	fine-tuned AR model	doc	mixed	Plain German & German Easy Language	Semi-synthetic Simple German Web Corpus	7,130	github.com/MSLars/German-Text-Simplification
trimmed_mbart_sents_DEplain-APA	Stodden et al. (2023)	fine-tuned seq2seq	sent	news	German language learners	DEplain-APA	10,660	huggingface.co/DEplain/trimmed_mbart_sents_apa
trimmed_mbart_sents_DEplain-APA_web	Stodden et al. (2023)	fine-tuned seq2seq	sent	mixed	Plain German & German Easy Language	DEplain-APA+web	10,660 & 1,594	huggingface.co/DEplain/trimmed_mbart_sents_apa_web
mT5-DEplain-APA	Stodden (2024b)	fine-tuned seq2seq	sent	news	German language learners	DEplain-APA	10,660	huggingface.co/DEplain/mT5-DEplain-APA
mT5-SGC	Stodden (2024b)	fine-tuned seq2seq	sent	mixed	German language learners	Simple German Corpus '23	4,430	huggingface.co/DEplain/mt5-simple-german-corpus
trimmed_mbart_docs_DEplain-APA	Stodden et al. (2023)	fine-tuned seq2seq	doc	news	German language learners	DEplain-APA	483	huggingface.co/DEplain/
trimmed_mbart_docs_DEplain-APA_web	Stodden et al. (2023)	fine-tuned seq2seq	doc	mixed	Plain German & German Easy Language	DEplain-APA+web	483 & 756	huggingface.co/DEplain/

Table 8.11: Summary of German TS models including own work (last part). Each line separates different model approaches. Extended version of Stodden (2024b) and Table 6.3. All URLs have lastly been accessed at July 24, 2024. Part II (continued from previous page).

Regarding the other two levels, Cripwell et al. (2023b) are also one of the first who trained and evaluated a TS system on both the document and sentence levels. Again, at about the same time, in Stodden et al. (2023), we have also proposed the first experiments on German TS considering the same resources for document and sentence simplification. Like Cripwell et al. (2023b), we have experimented with BART model variants (Lewis et al., 2020) for both text units: using the long-mBART approach of Rios et al. (2021) for document simplification and the general mBART model for sentence simplification.

In the following, I will discuss further the results of our document simplification models (see Subsubsection 8.6.2.1) and sentence simplification models (see Subsubsection 8.6.2.2).

8.6.2.1 Document Simplification

For document simplification, we have proposed three new German document simplification corpora based on the long-mBART approach of Rios et al. (2021), i.e., mBART-doc-DEplain-APA, m-BART-doc-DEplain-web, and mBART-doc-DEplain-APA+web where the latter is trained on the data of the two models named first.

Data Augmentation Since all three models are trained with the same model approach and the same hyperparameter, we can compare the effect of their training data (see RQ 6-2) when evaluating on the same test sets, i.e., DEPlain-web-doc and DEplain-APA-doc.

When evaluating on DEplain-APA-doc, as can be seen in Table 8.12a, the model that is only trained on the news data of APA achieves the best results wrt. SARI, BLEU, and BERTScore. Training on only out-of-domain (i.e., DEplain-web-doc) or a mix of in-domain and out-of-domain data, all scores are decreased. But, when evaluating on DEplain-web-doc, a combination of the web and news training documents improves the results wrt. SARI and BLEU, but decreases BERTScore compared to training on either in-domain or out-of-domain data (see Table 8.12b). We assume that this effect is due to the high variety of simplifications within DEplain-web versus the homogeneous simplifications in DEPlain-APA. Based on these results, adding more data does not automatically increase the result; it also depends on the domain, variety, and quality of the simplification data.

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

(**a**) DEplain-apa test (n=48)

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

(**b**) DEplain-web test (n=147)

Table 8.12: Results on Document Simplification using finetuned long-mBART. *n* corresponds to the length of the training data. Copied from Table 4 in Stodden et al. (2023).

Furthermore, following the inter-annotator agreement during manual alignment (see Table 8.1) and the insights in the annotation based on the examples of the alignment simplification plans (see Figure 8.4), we assume that document simplification is a tougher problem in the web documents than in the news documents. Especially the high amount of cross-sentence simplification operations such as reordering, merging, and splitting in the web document might make

the automatic simplification much stronger. Due to the lack of more than one reference simplification document, even if a TS system would generate a good simplification of a document, it might be assessed with low scores wrt. automatic metrics as the generation does not meet the expected gold simplification.

Model Comparison In order to compare the quality of our document simplification models with other models, i.e., mBART-20min, we have evaluated it also on the 20min corpus that also includes simplifications of news texts. As can be seen in Table 8.13, our models achieve very low scores on all automatic metrics and are greatly outperformed by the model trained on the 20min corpus. In comparison, our model trained on web documents achieves the best SARI and BLEU scores. On the one hand, this is surprising as in the previous section, we have shown that in-domain training performs better than out-of-domain training. But, on the other hand, the 20min data includes very strong rewriting as it is a mix of simplification and summarization, whereas the APA data contains only mild simplification with a low to non-existing degree of summarization. By contrast, the DEplain-web corpus also contains strong simplification with a high degree of deletions, which might be more similar to the 20min corpus than the DEplain-APA corpus, although the texts are written for another text domain.

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

Table 8.13: Results on Document Simplification Testing on 20min with long-mBART. Copied from Table 15 of Stodden et al. (2023).

Following this, sentence simplification on the DEplain-web corpus and the 20min corpus or document simplification using document plans might be easier and more suitable tasks than document simplification. However, document simplification on the mild simplifications of the DEplain-APA document seems to be suitable; in future work, I plan to focus more on solutions for document simplification using this data. For example, it would be interesting whether zero or one-shot experiments could surpass our fine-tuning approach using long-mBART.

8.6.2.2 Sentence Simplification

For German sentence simplification, more models, more evaluation sets, and better evaluation metrics exist than for document simplification. In this section, I discuss the capability of German sentence simplification systems, which can vary regarding the training data, e.g., by adding augmented data, or comparing different model architectures trained and evaluated on the same data (see RQ 6-2). So far, TS models have been rarely compared to TS models proposed in other papers (see Model Challenge D); exceptions are Mallinson et al. (2020), or Anschütz et al. (2023). But if so, as previously discussed in Subsection 6.8.1, it has not been clear whether the same data split (e.g., for APA-LHA Spring et al., 2021) or the same evaluation method has been used (e.g., same SARI or FKGL implementation). Reasons for this are a lack of provided

resources, e.g., generated simplifications, model checkpoints, or descriptions on evaluation implementation (see Subsection 5.4.1 or Subsection 6.8.1).

To enable comparison of the models in the reproduction study (see Model Challenge E), as well as further improve the comparability of German sentence simplification models (see Model Challenge D), we have proposed EASSE-multi (Stodden, 2024a) (see Subsection 7.5.3. EASSE-multi facilitates the use of the same evaluation metrics and settings in non-English TS reports and contains a GitHub repository in which available German simplification test sets and system outputs on these test sets can be easily stored, compared, and uploaded.

As mentioned previously, currently, a comparison between existing TS models is hampered as they are all evaluated with different metric implementations or on different test sets (see Model Challenge D). To enhance the overview and comparison of the capabilities of German sentence simplification models, in Stodden (2024b) (see Subsection 7.6.1), we have proposed a reproduction study on eight German sentence simplification models (including our own two models, i.e., DEplain-sent-APA and DEplain-sent-APA+web) and additionally proposed three new TS systems. In order to compare several German TS systems with respect to training data and model architectures, we have automatically evaluated and compared the reproduced TS models on six German test sets, i.e., APA-LHA-OR-A2, APA-LHA-OR-B1, DEplain-APA, DEplain-web, GEOlino, and TextComplexityDE. Overall, our own TS model mbart-DEplain-sent-APA+web performed best on four of six test sets wrt. BERTScore-Precision, and on three wrt. BLEU. Regarding SARI, mbart-DEplain-sent-APA performed best on two test sets; BLOOM-10-similarity and Sockeye-APA-LHA performed best on each two test sets.

The previously mentioned TS system can be evaluated more fine-grained. In the following, I provide this kind of analysis by addressing Model Challenge A regarding the lack of TS studies with augmented data, addressing Model Challenge F regarding training on corpora with different sizes, testing the models' capabilities regarding another domain, and providing insights based on an error analysis.

Data Augmentation We have trained a sentence simplification model with the same hyperparameters on only DEplain-APA and additionally with the automatically aligned sentence pairs of DEplain-web, resulting in DEplain-APA+web.

Comparing the metrics' scores of mBART-DEplain-APA and mBART-DEplain-APA+web on the sentence level test sets of DEplain-APA (see Table 8.14a), adding additional training data (i.e., the web data) did not make a big difference, except for SARI. As expected, training only on news data achieves a higher SARI score on the news test. However, when tested on the DEplain-web test set (see Table 8.14b), adding web data to the training data has improved all the measured metrics (except FRE and sentence splits). This supports the finding that adding at least some data of the relevant domain (i.e., 1,281 sentence pairs of DEplain-web) leads to a better generalization of the model for this domain. But I assume if we would have added less error-prone manually aligned sentence pairs of DEplain-web the effect would have been more significant.

Same Resource, but Different Size & Quality Although mBART-DEplain-APA, mT5-DEplain-APA, and sockeye-DEplain-APA are all trained on the same training data, the models achieve

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio ↓	Sent. splits ↑
mBART-DEplain-APA	28.49	38.72	0.64	65.30	0.99	1.07
mBART-DEplain-APA+web	28.03	33.81	0.64	65.20	0.98	1.05

(**a**) mBART evaluation on DEplain-APA.

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio ↓	Sent. splits ↑
mBART-DEplain-APA	13.50	33.11	0.40	69.65	0.90	1.30
mBART-DEplain-APA+web	17.99	34.07	0.44	69.05	0.85	1.16

(**b**) mBART evaluation on DEplain-web.

Table 8.14: Comparison of DEplain-mBART models on DEplain test sets.

quite different scores on the DEplain-APA test set wrt. BLEU, SARI, BERTScore-Precision, and FRE. Hence, the different transformer models and not only the data have an effect on the results. As summarized in Table 8.15, the mBART model achieves the best scores wrt. BLEU and BERTScore-Precision whereas the sockeye model achieves the best scores wrt. SARI. In comparison, the mT5 model compresses sentences but splits sentences the least, whereas the sockeye model generates the most simple sentences wrt. FRE and the most sentence splits. Hence, each model seems to have different capabilities, and each can be of more or less advantage depending on the needs of the target group.

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio ↓	Sent. splits ↑
sockeye-DEplain-APA	19.58	44.14	0.53	71.45	0.94	1.09
mBART-DEplain-APA	28.49	38.72	0.64	65.30	0.99	1.07
mT5-DEplain-APA	22.32	39.41	0.61	63.20	0.87	1.04
mBART-DEplain-APA+web	28.03	33.81	0.64	65.20	0.98	1.05

Table 8.15: Evaluation on DEplain-APA.

Out-of-Domain Evaluation However, a more fine-grained evaluation is required to justify these assumptions made based on automatic evaluation scores. For example, Stroh (2024) has manually evaluated the capability of the mBART-DEplain-APA+web model with respect to the simplification of instructional texts. Overall, the model seems to not generalize to these fairly differently written texts as it focuses more on simplification for non-native speakers than on expert-layman simplification as required in this task. In a comparison to GPT-4 (zero-shot and one-shot), mBART-DEplain-APA+web resolves less nominalizations (i.e., no nominalization at all), less rewriting into active voice and less rewriting into present tense. Even if mBART-DEplain-APA+web is not transferable to instructional texts, in Stodden (2024b), we found that it performs quite well on simplification for children (i.e., the GEOlino test set), simplification of Wikipedia texts for language learners (i.e., TCDE19 test set), but comparatively low on other web test sets (i.e., Simple German Web '23).

Error Analysis Furthermore, Lemgen (2024)[9] has manually analyzed some of the systems, e.g., mT5-DEplain-APA, mBART-DEplain-APA, and mBART-DEplain-APA+web, regarding wrongly applied simplification operations. He comes to the conclusion that on DEplain-APA and DEplain-web mT5-DEplain-APA generates the most erroneous simplifications in comparison to mBART-DEplain-APA+web (second-most errors) and mBART-DEplain-APA (most correctly simplified sentences). The most frequent errors of the models are generating nonsense sentences and removing too much information, whereas generating grammatically correct sentences or inserting wrong information seems to be no big issue. However, these results do not correspond to the evaluation metrics (see Table 8.15), e.g., mT5-DEplain-APA achieves the highest SARI score on DEplain-APA but is annotated with the most errors. This again shows that better evaluation metrics are required which could also take into account erroneous and not only correct simplification operations as proposed, e.g., by Devaraj et al. (2022), Ma et al. (2022) or Yamaguchi et al. (2023).

Summary So I can sum up wrt. RQ 6-2 that in our experiments the model architecture as well as the training data have an effect on the quality of text simplification when considering automatic metrics and manually annotated errors. However, a manual evaluation including target groups would be required to fully verify these findings.

8.6.3 RQ 6-3: Effect on Real-World Application

In previous sections of this work, I have focused on text simplification research, i.e., how to build the most suitable text simplification system for different target groups and for texts of different domains. Based on research on text complexity (e.g. Amstad, 1978 or Mohtaj et al., 2022) and text simplification (e.g., Bock and Pappert, 2023 or Ebling et al., 2022), we have shown how to manually and automatically rewrite a text to make it better readable for people with different requirements on a text (see also Chapter 2). In evaluation studies, researchers have shown that manually and automatically simplified texts are better readable so that they increase the comprehensibility of people with learning difficulties (e.g., see Schlippe and Eichinger 2023 or Säuberli et al. 2024).

On the other hand, we only know little about how text simplification systems can be integrated into the daily life of people with reading problems (see step J in TS workflow). For some use cases, some text simplification systems are already in use, e.g., automatically simplifying texts for language learners at CEFR level A1, A2, and B1 by capito digital (capito, 2024), or automatically simplifying public authority announcements into German Easy Language by SUMM-AI (Stadt Aschaffenburg, 2024). However, to the best of my knowledge, no research exists yet on how (automatic) simplifications affect the usage and the intention of people to use websites with complex texts (see RQ 6-3 and Model Challenge G).

Therefore, in Stodden and Nguyen (2024) (see Subsection 7.6.2), we propose an analysis on how text simplification actually affects the readers' usage intention in real-world applications, i.e., do people with reading problems would more or less likely participate in a discussion if

9 This Bachelor thesis has been supervised by Regina Stodden (first supervisor) and Laura Kallmeyer (second supervisor); the idea of the project comes from Regina Stodden.

(automatic) simplifications are provided? Based on our findings regarding simplification of user-generated texts in Stodden (2021a) (see Subsection 7.4.1), we have selected one possible use case for text simplification, i.e., simplification of contributions in the deliberative online platform and analyzed how text simplification affects readers and writers on these platforms. In our analysis, we have focused on the technology acceptance of automatic text simplification on these platforms by conducting a nearly-realistic user study with German language learners and German native speakers.

The findings from our research indicate that text simplification does not directly impact engagement in online participation processes. However, our results also suggest that text simplification does not hinder the intention to use e-participation platforms. In fact, German language learners and people with reading difficulties tend to favor the existence of text simplification on these platforms, regardless of whether manual or automatic simplification. Further, no unintended side effects are expected for participants without reading and writing difficulties when reading proposals in standard language and Plain Language side-by-side.

Following this, we have shown that our work is not only beneficial for the research regarding German text simplification, but its usage also has a practical impact on real-world applications such as deliberative online platforms. In future work, more similar investigations are required to analyze the impact of text simplification in other use cases, e.g., simplification of public authority texts or news, and for other target groups, e.g., people with learning difficulties or laypeople.

Chapter 9

Limitations

Although my work has enhanced German text simplification in many aspects, the work still shows some limitations. For example, this work is narrowed to automatic simplification of written German texts; it does not include detailed research regarding manual simplification, automatic simplification of spoken texts, or automatic simplification of languages other than German. In the following, we provide more details regarding the limitations and, if available, refer to related work in the named fields.

9.1 Automatic Text Simplification

The focus on automatic text simplification rather than manual simplification is a major but intended limitation of this work. However, automatic simplification depends on manual simplification. Therefore, I have also tackled manual simplification to some extent. In more detail, while automatic approaches are trained and evaluated on manually generated simplifications, trained translators can benefit from computer-assisted translation due to, e.g., a reduction in their cognitive load (Hansen-Schirra et al., 2020b).

Overall, our study has offered some insight into manual simplification and manual post-editing of automatically simplified texts, but has primarily focused on automatic simplification, as manual simplification itself is a huge research topic. Therefore, I have addressed only the most important facts of manual simplification, which are necessary to understand the scope of automatic text simplification. In future work, it would be interesting to collaborate more with researchers in translation studies (including simplified languages), professional translators, and the target group to minimize the gap between manual and automatic text simplification in German.

9.2 Simplification of Written Texts

I have also exclusively focused on written texts, even though research exists regarding simplification in multi-modal settings, for example,

i) manual (but not automatic) consecutive or simultaneous interpreting into a simplified language (e.g., see Degenhardt 2020),

ii) manual simplifications used in spoken contexts, e.g., audio guides (e.g., see Scheele 2020),

iii) manually simplified subtitles in films (e.g., see Marmit 2020), or

iv) television news spoken in simplified language and visualized with videos (e.g., see "Tagesschau in Einfacher Sprache" (engl.: Tagesschau in Plain Language").[1])

In future research, it would be interesting to combine the findings from written automatic text simplification to manual and automatic multi-modal simplification, e.g., by automatically generating scripts in simplified language for subtitles or television news.

9.3 Supporting the Target Group via Post Editing

I overall have aimed to build simplification systems that support and empower the target groups of simplified German by addressing their special needs in the simplified version of the text and by providing better readable texts for them. Nevertheless, I have decided to address our TS systems more to professional translators instead of directly to the target group. Motives for limiting the addressee are that I expect system-generated simplifications to be error-prone to some extent wrt. simpleness, correctness, or fluency (Stajner, 2021), on the one hand, and that people in the target groups (e.g., children, people with cognitive disabilities, dementia, or reading deficits, or people learning a new language) are vulnerable against unverified texts (Stajner, 2021), on the other hand.

As some of the texts are of a safety-critical domain (e.g., medical texts), it is important that the information of the original text is not changed in the simplified text. For example, a change in the dose, name, or frequency of taking the medicine could cause injuries to varying degrees. Especially for people of the target group of German Easy Language, a high sensitivity against publishing wrong information should be considered because some of them do not have (yet or not anymore) the full capabilities to critically reflect on content (Park, 2012), and do not have access to alternative texts that correspond to their reading skills to verify the information (that is also the reason why the texts have been simplified).

Hence, for these ethical reasons, I suggest proof-reading of the system-generated texts prior to publication to the target groups. Translators trained in simplification could use the automatically generated simplifications as a first draft of the simplification and manually verify and post-edit them (Stajner, 2021; Deilen et al., 2024). With this procedure, I decrease the potential risk of my work by requesting manual checks of the automatically generated simplifications before publication. Furthermore, this procedure also reduces the risk of misinformation and misunderstandings for the target group.

While it is arguably a smaller objective (in terms of system output quality) to make simplifications ready for post-editing, there are some challenges with the automatically generated simplifications that need to be overcome. A recent study by Carrer et al. (2024) has demonstrated that professional translators currently face some limitations in their productivity when post-editing automatically generated simplifications. In comparison, trained translators simplify a text more quickly and with less effort when starting directly from the original text and

1 https://www.tagesschau.de/multimedia/sendung/tagesschau_in_einfacher_sprache [last access: July 24, 2024]

not the automatically generated text. It could be just a matter of getting used to this new form of simplification. The findings of Carrer et al. (2024) do not imply that automatic simplification is ineffective for trained translators; rather, the findings suggest that the quality of text simplification systems must be further enhanced before their outputs are suitable and helpful for post-editing, e.g., involving more simplification operations and simplifying more strongly.

9.4 Simplification on Document and Sentence Level

Another limitation of this work is the text unit for simplification, i.e., document, paragraph, and sentence units. Professional translators mostly simplify texts document-wise by applying simplification operations such as reordering, comprising, and making the text more coherent. However, my focus is more predominant on sentence simplification, although there are known issues regarding missing context when processing isolated sentences. But international text simplification research also currently mostly focuses on the sentence level due to more limited simplification operations (e.g., splitting, lexical substitution, or verb tense change) and better evaluation frameworks (Alva-Manchego et al., 2020b). In future work, the document plans of DEplain might help to integrate the previous and following sentences into future TS models or could facilitate paragraph simplification.

9.5 Evaluation of Automatic Text Simplification

The impact of our results is also limited due to the evaluation methods we have chosen. We have only automatically, but unfortunately not manually, evaluated the system generations. The major reason is a lack of best practice guidelines for evaluation. In the contributions of my co-authors and myself, we have extended the manual evaluation protocol (see Stodden 2022 and Stodden 2021c), but have not verified it yet in practice. However, the application of our new evaluation schema on the gold data has shown that it can be well used to estimate the quality and variety of a text simplification corpus. However, I strongly believe that in future work our new annotation schema could also be helpful for the evaluation of system-generated simplifications.

Regarding automatic evaluation, we have focused on optimizing currently available metrics, even though many limitations of them are known. Therefore, our benchmark results presented in (Stodden, 2024b) should be interpreted with caution. In future work, new metrics for German TS could be built based on our manual evaluation on the gold data (see Stodden et al. 2023) or our manual error annotation of the system generations (see Lemgen 2024).

Chapter 10

Conclusion

Following the comprehensive analysis and discussion of my contributed work and suggestions regarding future investigations for German text simplification regarding data, models, and evaluation, I now draw an overall conclusion on the goal of evaluating and enhancing the potential of machine learning methods for automatic German sentence and document simplification.

As also visualized in Figure 10.1, I have enhanced many components of automatic German text simplification with respect to

- **predicting and explaining text complexity** by implementing a sentence complexity prediction system and providing a typology of German simplification operations (see step 0 in Figure 10.1),
- **facilitating the building process of (German) TS corpora** by providing an annotation guideline for German TS and an annotation tool for multi-lingual TS (see step A, C, D, and F),
- **German document and sentence simplification corpora** by building and evaluating (new) corpora regarding their quality (see step B, E, and F),
- **German text simplification models** by proposing and reproducing models for a German TS benchmark on several test sets (see step G),
- **German text simplification evaluation** by providing a platform for unified automatic German TS evaluation as well as raising new concerns regarding manual and automatic evaluation of TS (see step H), and
- the relevance of text simplification in real world applications (see step J).

Overall, with respect to my main research question (i.e., *How can the potential of machine learning methods be explored for the simplification of German texts, considering data availability, evaluation suitability and comparing their effectiveness for document and sentence simplification approaches?*), I have shown that automatic German text simplification has a high potential, but is still rather in its early stages. In particular, with regard to the field of automatic German document simplification: despite the existence of a few corpora and models, the absence of a reliable document evaluation method leaves open a number of questions regarding the quality and suitability of the available resources.

In contrast, for sentence simplification, I have demonstrated the capability of current TS models wrt. different target groups and text domains. Nevertheless, more robust evaluation is

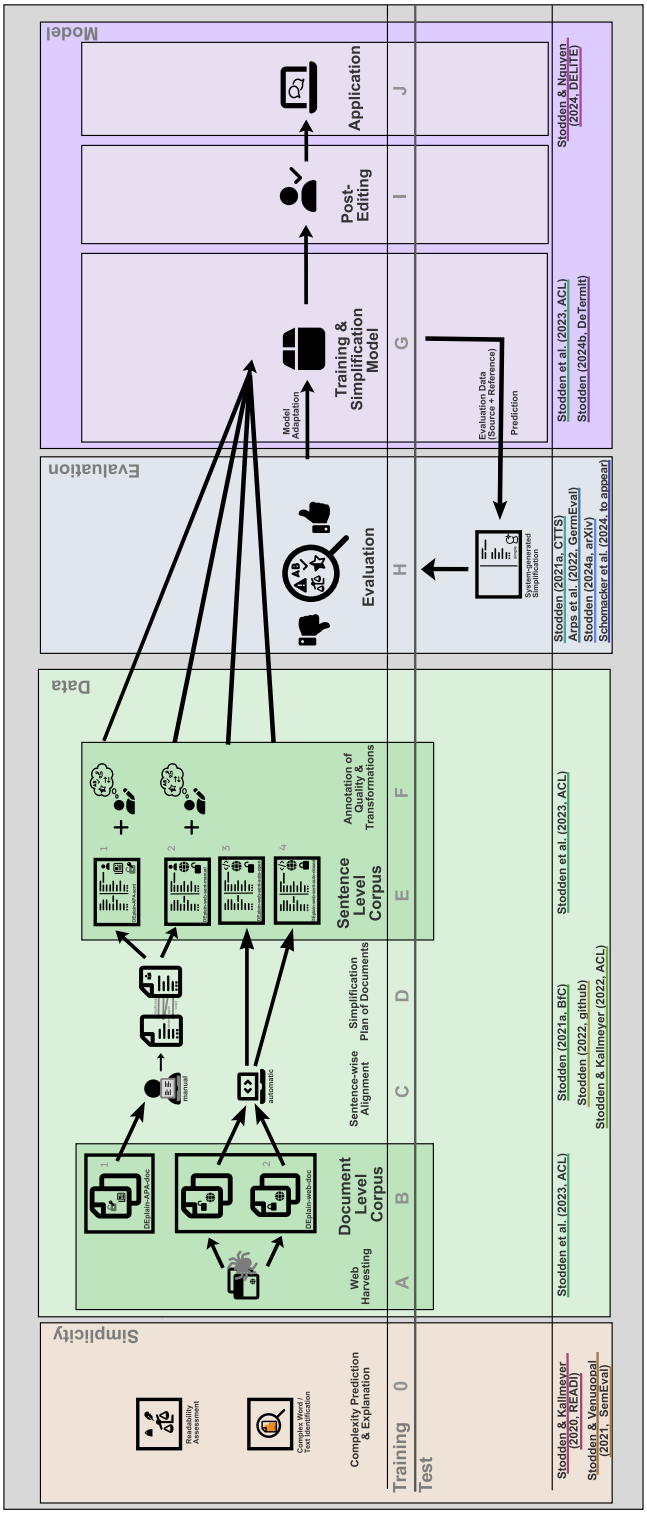


Figure 10.1: Text simplification workflow including contributions of this thesis (same as Figure 8.1).

necessary to ensure the reliability and efficacy of automatic sentence and document simplification prior to its implementation in real-world settings.

I believe that the new typology for German simplification operations, the new annotation tool TS-ANNO, the new DEplain corpora, the new German TS models, the new evaluation aspects, and the EASSE-DE evaluation framework will strongly facilitate German text simplification research in the future due to its ease of access and ease of use.

Chapter 11

Ethics & Impact Statement

In addition to the "Ordnung über die Grundsätze zur Sicherung guter wissenschaftlicher Praxis" (translation: "regulations on the principles for safeguarding good scientific practice") of the Heinrich Heine University Düsseldorf, I also honor the Code of Ethics of the Association for Computational Linguistics (ACL)[1]; in the following I provide concise answers on the Responsible NLP Research check list[2].

Limitations & Risks In previous parts of this work, I have already discussed the limitations and risks of my work (see Chapter 9). In order to assess the risk of my work, I use the Artificial Intelligence Act of the European Union (EU AI Act) European Parliament and Council of the European Union 2024. Our text simplification systems do not correspond to any prohibited AI practices following the EU AI Acts (see Article 5 of the EU AI Act), as their outputs are not expected to be harmful to people. But I follow article 50 of the EU AI Act regarding transparency obligations: if the automatically generated simplifications are used in production, I recommend adding a remark that this text has been generated automatically and might contain errors such as exemplified in Stodden and Nguyen (2024). However, officially, our TS models are exempted from the EU AI Act because the act does not regulate AI models that are developed for scientific purposes (see article 2-6 of the EU AI Act) or models that are published with an open license (see article 2-12 of EU AI Act) which both apply to our TS models.

Scientific Artifacts In my work, I have used several scientific artifacts such as models, code, and data. In order to follow good scientific practice and responsible NLP, on the one hand, I have cited the artifacts, discussed their licenses, and only used them according to their intended usage. On the other hand, I have specified the intended use and the licenses for the artifacts that we have created. Our data do not contain personal data except for the study in Stodden and Nguyen (2024) and the annotators' metadata for the DEplain corpora (Stodden et al., 2023) which have been anonymized. For building our corpora, we have just used professionally simplified texts in which we do not expect any personal data or offensive content.

1 https://www.aclweb.org/portal/content/acl-code-ethics [last update: March 13, 2020; last access: July 30, 2024]

2 https://aclrollingreview.org/responsibleNLPresearch/ [last update: *n/a*; last access: July 30, 2024]

Furthermore, we have provided detailed information regarding the characteristics and statistics of our data, e.g., domains, language, language varieties, and linguistic phenomena (i.e., simplification operations or alignment types).

Computational Experiments In order to train or reproduce German TS models, we have performed computational experiments for which we have specified the hyperparameters in the corresponding publications. We have utilized computational resources in an efficient manner by, e.g., leveraging pre-trained models and fine-tuning for a limited number of epochs. If I have included existing packages in my code, I have cited them and specified the parameter settings used in our experiments.

Human Annotations My work also includes some human annotations, e.g., simplification operations and quality estimation of the DEplain corpus. We have made available the metadata of the annotators and the annotation instructions (i.e., the annotation guideline Stodden 2022 and the annotation tool Stodden and Kallmeyer 2022). As the data and annotation did not contain any sensitive information, it was not necessary to seek approval from an ethics review board.

AI assistants I have used AI assistants during the writing of the thesis text and the contributed articles. I respect the "Leitlinien zur Verwendung generativer Künstlicher Intelligenz in der Lehre" (translation: "Guidelines for the use of generative artificial intelligence in teaching") of the Heinrich Heine University Düsseldorf[3] and the AI Writing Assistance Policy of ACL[4] and have followed them to the best of my ability. The following AI tools or models have been used to improve writing style, inspire the writing process, and to omit spelling and grammar errors, i.e., Grammarly[5], DeepL translator[6], DeepL Writer[7], WriteFull[8]. Additionally, for the same purpose, I have utilized the following generative large models, e.g., ChatGPT[9], Mixtral 8x7B Instruct (Jiang et al., 2024), or LLaMA 3 SauerkrautLM 70B Instruct[10].

In order to extend my literature research, I have used AI assistants to enlarge my list of relevant background or related works, e.g., Semantic Scholar[11], Connected Papers[12], Elicit[13], or Perplexity AI[14]. I have not used any AI assistant for code generation (e.g., Copilot[15]).

3 https://www.hhu.de/fileadmin/redaktion/ZUV/Justitiariat/Amtliche_Bekanntmachungen/AB_14_240425.pdf [last update: April 25, 2024; last access: July 30, 2024]
4 https://2023.aclweb.org/blog/ACL-2023-policy/ [last update: January 10, 2023; last access: July 30, 2024]
5 https://www.grammarly.com/about [last access: July 30, 2024]
6 https://www.deepl.com/de/translator
7 https://www.deepl.com/write [last access: July 30, 2024]
8 https://www.writefull.com/ [last access: July 30, 2024]
9 I have used different versions of ChatGPT between the beginning of 2023 and the middle of 2024. https://chat.openai.com/ [last access: July 30, 2024]
10 https://huggingface.co/VAGOsolutions/Llama-3-SauerkrautLM-70b-Instruct [last access: July 30, 2024]
11 https://www.semanticscholar.org/ [last access: July 30, 2024]
12 https://www.connectedpapers.com/ [last access: July 30, 2024]
13 https://elicit.com/ [last access: July 30, 2024]
14 https://www.perplexity.ai/ [last access: July 30, 2024]
15 https://github.com/features/copilot [last access: July 30, 2024]

References

Alaa Abd-Alrazaq, Jens Schneider, Borbala Mifsud, Tanvir Alam, Mowafa Househ, Mounir Hamdi, and Zubair Shah. 2021. A comprehensive overview of the COVID-19 literature: machine learning–based bibliometric analysis. *Journal of medical Internet research*, 23(3):e23703.

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Fatima Al-Raisi, Weijian Lin, and Abdelwahab Bourai. 2018. A Monolingual Parallel Corpus of Arabic. *Procedia Computer Science*, 142:334–338. Arabic Computational Linguistics.

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated Text Simplification: A Survey. *ACM Computing Surveys*, 54(2).

Oliver Alonzo, Jessica Trussell, Becca Dingman, and Matt Huenerfauth. 2021. Comparison of Methods for Evaluating Complexity of Simplified Texts among Deaf and Hard-of-Hearing Adults at Different Literacy Levels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Fernando Alva-Manchego. 2019. Difference between paper equations and code #8. GitHub issue: `https://github.com/cocoxu/simplification/issues/8`. [Online; Last Change: 2019-03-08; Last Access: 2024-03-05].

Fernando Alva-Manchego, Joachim Bingel, Gustavo Paetzold, Carolina Scarton, and Lucia Specia. 2017. Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing* (*Volume 1: Long Papers*), pages 295–305, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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

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

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2019b. Cross-sentence transformations in text simplification. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 181–184, Florence, Italy. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020b. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Marcelo Amancio and Lucia Specia. 2014. An analysis of crowdsourced text simplifications. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 123–130, Gothenburg, Sweden. Association for Computational Linguistics.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 307–317, Tilburg University, The Netherlands. Association for Computational Linguistics.

Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Phd thesis, University of Zurich, Switzerland.

Mandya Angrosh, Tadashi Nomoto, and Advaith Siddharthan. 2014. Lexico-syntactic text simplification and compression with typed dependencies. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1996–2006, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1147–1158, Toronto, Canada. Association for Computational Linguistics.

APA – Austria Presse Agentur. 2020. APA-TopEasy: Nachrichten leicht verständlich. https://science.apa.at/wp-content/uploads/2021/02/APA_PB_TopEasy.pdf. [Online; Last Change: 2020-02, Last Access: 2024-06-20].

David Arps, Jan Kels, Florian Krämer, Yunus Renz, Regina Stodden, and Wiebke Petersen. 2022. HHUplexity at text complexity DE challenge 2022. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 27–32, Potsdam, Germany. Association for Computational Linguistics.

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

Hiroki Asano, Tomoya Mizumoto, and Kentaro Inui. 2017. Reference-based metrics can be replaced with reference-less metrics in evaluating grammatical error correction systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–348, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Hadi Asghari, Freya Hewett, and Theresa Züger. 2023. On the Prevalence of Leichte Sprache on the German Web. In *Proceedings of the 15th ACM Web Science Conference*, WebSci '23, New York, NY, USA. Association for Computing Machinery.

Accessibility Guidelines Working Group at W3C. 2024. Reading Level (Level AAA) . https://www.w3.org/WAI/WCAG22/Understanding/reading-level.html. [Online; Last Change: *n/a*; Last Access: 2024-06-6].

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1).

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

Auswärtiges Amt. 2020. Deutsch als Fremdsprache weltweit. Datenerhebung 2020. Last Access: 2024-04-12.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan. Asian Federation of Natural Language Processing.

Richard Bamberger and Erich Vanecek. 1984. *Lesen - Verstehen - Lernen - Schreiben. Die Schwierigkeitsstufen von Texten in deutscher Sprache*. Jugend u. Volk Sauerlaender, Wien.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Petra Barancikova and Ondřej Bojar. 2020. COSTRA 1.0: A dataset of complex sentence transformations. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3535–3541, Marseille, France. European Language Resources Association.

Gianni Barlacchi and Sara Tonelli. 2013. ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian: Computational Linguistics and Intelligent Text Processing. In *Computational Linguistics and Intelligent Text Processing*, Berlin & Heidelberg, Germany. Springer.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

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

Andreas Baumert. 2018. *Einfache Sprache : Verständliche Texte schreiben*. Spaß am Lesen, Münster.

David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. MeaningBERT: assessing meaning preservation between sentences. *Frontiers in Artificial Intelligence*, 6.

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

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, and ... Thomas Wolf. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *Preprint*, arXiv:2211.05100.

Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. 2018. Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

BITV. 2011. Barrierefreie-Informationstechnik-Verordnung vom 12. September 2011 (BGBl. I S. 1843) . Zuletzt geändert durch Artikel 1 der Verordnung vom 24. Oktober 2023 (BGBl. 2023 I Nr. 286).

Carl-Hugo Björnsson. 1968. *Läsbarhet*. Liber, Stockholm.

Bettina M. Bock. 2015. Barrierefreie Kommunikation als Voraussetzung und Mittel für die Partizipation benachteiligter Gruppen: Ein (polito-)linguistischer Blick auf Probleme und Potenziale von "Leichter" und "einfacher Sprache". *Linguistik Online*, 73(4).

Bettina M. Bock. 2019. *"Leichte Sprache" – Kein Regelwerk: Sprachwissenschaftliche Ergebnisse und Praxisempfehlungen aus dem LeiSA-Projekt*, volume v.5 of *Kommunikation – Partizipation – Inklusion*. Frank & Timme, Berlin, Germany.

Bettina M. Bock and Sandra Pappert. 2023. *Leichte Sprache, einfache Sprache, verständliche Sprache*. Narr Francke Attempto, Tübingen.

Stefan Bott and Horacio Saggion. 2011. An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 20–26, Portland, Oregon. Association for Computational Linguistics.

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

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Proceedings of Advances in Neural Information Processing Systems* (*NeurIPS*), volume 33, pages 1877–1901. Curran Associates, Inc.

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

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

Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2022. Linguistically-Based Comparison of Different Approaches to Building Corpora for Text Simplification: A Case Study on Italian. *Frontiers in Psychology*, 13.

Bram Bulté, Leen Sevens, and Vincent Vandeghinste. 2018. Automating lexical simplification in Dutch. *Computational Linguistics in the Netherlands Journal*, 8:24–48.

Bundeszentrale für gesundheitliche Aufklärung. 2024. Informationen rund um das Coronavirus . https://www.infektionsschutz.de/coronavirus/. [Online; Last Change: 2023-12-29; Last Access: 2024-06-21].

Peter Bydlinski. 2015. Modernisierung des ABGB . In *Österreichische Juristen Zeitung (ÖJZ)*, 19, pages 869 – 876. Manz, Vienna, Austria.

Leonardo Campillos-Llanos, Ana Rosa Terroba Reinares, Sofía Zakhir Puig, Ana Valverde Mateos, and Adrián Capllonch Carrión. 2022. Building a comparable corpus and a benchmark for Spanish medical text simplification. *Procesamiento del lenguaje natural*, 69:189–196.

Carmen Canfora and Angelika Ottmann. 2020. Risks in neural machine translation. *Translation Spaces*, 9(1):58–77.

capito. 2020. Nachrichten in Leichter Sprache: APA TopEasy News. https://www.capito.eu/projekte/apa-topeasy-news/. [Online; Last Change: *n/a*; Last Access: 2024-06-20].

capito. 2024. Easy Language: What is it and why is it important? https://www.capito.eu/en/easy-language/. [Online; Last Change: *n/a*; Last Access: 2024-04-17].

Rémi Cardon and Adrien Bibal. 2023. On operations in automatic text simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 116–130, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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

Rémi Cardon, Adrien Bibal, Rodrigo Wilkens, David Alfter, Magali Norré, Adeline Müller, Patrick Watrin, and Thomas François. 2023. Annotation linguistique pour l'évaluation de la simplification automatique de textes. In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale*, pages 35–48, Paris, France. ATALA.

Rémi Cardon and Natalia Grabar. 2018. Identification of parallel sentences in comparable monolingual corpora from different registers. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 83–93, Brussels, Belgium. Association for Computational Linguistics.

Rémi Cardon and Natalia Grabar. 2020. French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Luisa Carrer, Andreas Säuberli, Martin Kappus, and Sarah Ebling. 2024. Towards holistic human evaluation of automatic text simplification. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 71–80, Torino, Italia. ELRA and ICCL.

M. John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10, Madison, Wisconsin, USA.

Helena Caseli, Tiago F. Pereira, Lucia Specia, Thiago Alexandre Salgueiro Pardo, Caroline Gasperin, and Sandra Maria Aluísio. 2009. Building a Brazilian Portuguese Parallel Corpus of Original and Simplified Texts. In *Proceedings of the 10th Conference on Intelligent Text Processing and Computational Linguistics*, pages 59–70, Mexico City, Mexico.

Mert Cemri, Tolga Çukur, and Aykut Koç. 2022. Unsupervised Simplification of Legal Texts. *Preprint*, arXiv:2209.00557.

Raman Chandrasekar, Christine Doran, and Srinivas Bangalore. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, page 1041–1044, Copenhagen, Denmark. Association for Computational Linguistics.

Raman Chandrasekar and Bangalore Srinivas. 1997. Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10(3):183–190.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70):1–53.

Seung Youn (Yonnie) Chyung, Katherine Roberts, Ieva Swanson, and Andrea Hankinson. 2017. Evidence-Based Survey Design: The Use of a Midpoint on the Likert Scale. *Performance Improvement*, 56(10):15–23.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925, Brussels, Belgium. Association for Computational Linguistics.

Louis Cohen, Lawrence Manion, and Keith Morrison. 2007. *Research methods in education*, 6. ed. edition. Routledge, New York, USA.

Brady Coleman. 1998. Are clarity and precision compatible aims in legal drafting? *Singapore Journal of Legal Studies*, pages 376–408.

Kevyn Collins-Thompson. 2014. Recent Advances in Automatic Readability Assessment and Text Simplification. *ITL - International Journal of Applied Linguistics*, 165(2):97–135.

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

Michael Cooper and Matthew Shardlow. 2020. CombiNMT: An exploration into neural text simplification models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5588–5594, Marseille, France. European Language Resources Association.

Council of Europe. 2020. *Common European framework of reference for languages: learning, teaching, assessment – Companion volume*. Council of Europe Publishing, Strasbourg, France. [Online; Last Change: *n/a*; Last Access: 2024-07-29].

Council of Europe. 2024. Self-assessment Grids (CEFR). `https://www.coe.int/en/web/portfolio/self-assessment-grid`. [Online; Last Change: *n/a*; Last Access: 2024-07-29].

Liam Cripwell, Joël Legrand, and Claire Gardent. 2022. Controllable sentence simplification via operation classification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2091–2103, Seattle, United States. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023a. Context-aware document simplification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13190–13206, Toronto, Canada. Association for Computational Linguistics.

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

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023c. Simplicity level estimate (SLE): A learned reference-less metric for sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12053–12059, Singapore. Association for Computational Linguistics.

Oscar M. Cumbicus-Pineda, Itziar Gonzalez-Dios, and Aitor Soroa. 2021. Linguistic Capabilities for a Checklist-based evaluation in Automatic Text Simplification. In *Proceedings of the First Workshop on Current Trends in Text Simplification (CTTS 2021)*, pages 70–83. CEUR-WS.

Östen Dahl. 2004. *The Growth and Maintenance of Linguistic Complexity*. John Benjamins.

Antje Dammel and Sebastian Kürschner. 2008. Complexity in nominal plural allomorphy - a contrastive survey of ten Germanic languages. In *Language complexity – Typology, contact, change*, pages 243–262. Benjamins, Amsterdam.

Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. In *Proceedings of CHI'22: Workshop on Generative AI and HCI*, New Orleans, LA, USA. Association for Computing Machinery (ACM).

Domingos de Oliveira. 2016. *Sagen Sie es einfach: Eine Einführung in die einfache Sprache*. Books on Demand, Norderstedt, Germany.

Julia Degenhardt. 2020. Konsekutivdolmetschen in Leichte Sprache. In Katharina Oster Anne-Kathrin Gros, Silke Gutermuth, editor, *Leichte Sprache? Empirische und multimodale Perspektiven*, pages 121–136. Frank & Timme, Berlin, Germany.

Silvana Deilen, Sergio Hernández Garrido, Ekaterina Lapshinova-Koltunski, and Christiane Maaß. 2023. Using ChatGPT as a CAT tool in easy language translation. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 1–10, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Silvana Deilen, Ekaterina Lapshinova-Koltunski, Sergio Hernández Garrido, Christiane Maaß, Julian Hörner, Vanessa Theel, and Sophie Ziemer. 2024. Towards AI-supported health communication in plain language: Evaluating intralingual machine translation of medical texts. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 44–53, Torino, Italia. ELRA and ICCL.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ–IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

Deutsches Institut für Normung (DIN). 2023. Empfehlungen für Deutsche Leichte Sprache (DIN SPEC 33429:2023-04 – Entwurf). Draft Version.

Deutsches Institut für Normung (DIN). 2024a. Plain language - Application for the German language - Part 1: Language-specific provisions (DIN 8581-1:2024-05).

Deutsches Institut für Normung (DIN). 2024b. Plain language - Part 1: Governing principles and guidelines (ISO 24495-1:2023).

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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. Association for Computational Linguistics.

Anna Dmitrieva and Aleksandra Konovalova. 2023. Creating a parallel Finnish-Easy Finnish dataset from news articles. In *Proceedings of the 1st Workshop on Open Community-Driven Machine Translation*, pages 21–26, Tampere, Finland. European Association for Machine Translation.

Anna Dmitrieva and Jörg Tiedemann. 2021. Creating an aligned Russian text simplification dataset from language learner data. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 73–79, Kiyv, Ukraine. Association for Computational Linguistics.

Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. 2020. The sockeye 2 neural machine translation toolkit at AMTA 2020. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 110–115, Virtual. Association for Machine Translation in the Americas.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3393–3402, Florence, Italy. Association for Computational Linguistics.

Sarah Ebling, Alessia Battisti, Marek Kostrzewa, Dominik Pfütze, Annette Rios, Andreas Säuberli, and Nicolas Spring. 2022. Automatic Text Simplification for German. *Frontiers in Communication*, 7.

Anne Eschenbruecher. 2021. What makes a concept complex? measuring conceptual complexity as a precursor for text simplification. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 154–160, Held Online. INCOMA Ltd.

Isabel Espinosa-Zaragoza, José Abreu-Salas, Elena Lloret, Paloma Moreda, and Manuel Palomar. 2023. A review of research-based automatic text simplification tools. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 321–330, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parlament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). [Online; Last Update: 2024-07-12; Last Access: 2024-07-30].

Dan Feblowitz and David Kauchak. 2013. Sentence simplification as tree transduction. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 1–10, Sofia, Bulgaria. Association for Computational Linguistics.

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

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Oliver Ferschke, Johannes Daxenberger, and Iryna Gurevych. 2013. A Survey of NLP Methods and Resources for Analyzing the Collaborative Writing Process in Wikipedia. In Iryna Gurevych and Jungi Kim, editors, *The People's Web Meets NLP: Collaboratively Constructed Language Resources*, pages 121–160. Springer-Verlag, Berlin & Heidelberg, Germany.

Pierre Finnimore, Elisabeth Fritzsch, Daniel King, Alison Sneyd, Aneeq Ur Rehman, Fernando Alva-Manchego, and Andreas Vlachos. 2019. Strong baselines for complex word identification across multiple languages. 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 970–977, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Forschungsstelle Leichte Sprache . 2023. Corona-Virus. `https://www.apotheken-umschau.de/einfache-sprache/krankheiten/corona-virus-723743.html`. [Online; Last Change: 2023-12-29; Last Access: 2024-06-21].

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Dennis Frieß, Katharina Esau, and Christiane Eilders. 2017. How Emotions, Humor and Narratives Interact with Traditional Characteristics of Deliberation Online. In *Proceedings of the 67th ICA Annual Conference (Political Communcation Division)*, volume Panel Paper.

Leon Fruth, Robin Jegan, and Andreas Henrich. 2024. An approach towards unsupervised text simplification on paragraph-level for German texts. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 77–89, Torino, Italia. ELRA and ICCL.

Fußball-Club St. Pauli. 2023. FC St. Pauli stellt KI-Übersetzungstool zur Verfügung. https://www.fcstpauli.com/news/einfache-sprache-fc-st-pauli-stellt-ubersetzungstool-zur-verfugung/ . [Online; Last Change: 2023-06-03; Last Access: 2024-07-04].

Kavita Ganesan. 2018. ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks. *Preprint*, arXiv:1803.01937.

Cristina Garbacea, Mengtian Guo, Samuel Carton, and Qiaozhu Mei. 2021. Explainable prediction of text complexity: The missing preliminaries for text simplification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1086–1097, Online. Association for Computational Linguistics.

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nandakishore Kambhatla. 2022. Text simplification for legal domain: Insights and challenges. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 296–304, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Federico Gaspari. 2006. Look who's translating. impersonations, Chinese whispers and fun with machine translation on the Internet. In *Proceedings of the 11th Annual Conference of the European Association for Machine Translation*, Oslo, Norway. European Association for Machine Translation.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler ..., and Oriol Vinyals. 2024. Gemini: A Family of Highly Capable Multimodal Models. *Preprint*, arXiv:2312.11805.

GEO.de. 2024. GEOlino Magazin. `https://www.geo.de/magazine/geolino-magazin/`. [Online; Last Change: *n/a*; Last Access: 2024-04-22].

Ulrich Germann. 2008. Yawat: Yet Another Word Alignment Tool. In *Proceedings of the ACL-08: HLT Demo Session*, pages 20–23, Columbus, Ohio. Association for Computational Linguistics.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, Beijing, China. Association for Computational Linguistics.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. In *The 22nd Workshop on Biomedical Natural*

Language Processing and BioNLP Shared Tasks, pages 468–477, Toronto, Canada. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. 2018. The Corpus of Basque Simplified Texts (CBST). *Language Resources and Evaluation*, 52(1):217–247.

Itziar Gonzalez-Dios, Iker Gutiérrez-Fandiño, Oscar m. Cumbicus-Pineda, and Aitor Soroa. 2022. IrekiaLFes: a new open benchmark and baseline systems for Spanish automatic text simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 86–97, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

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

Sian Gooding and Ekaterina Kochmar. 2019. Recursive Context-Aware Lexical Simplification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing: EMNLP-IJCNLP '19*, 1 Long Papers, pages 4855–4865. Association for Computational Linguistics.

Sian Gooding, Ekaterina Kochmar, Seid Muhie Yimam, and Chris Biemann. 2021. Word complexity is in the eye of the beholder. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4439–4449, Online. Association for Computational Linguistics.

Sian Gooding and Manuel Tragut. 2022. One size does not fit all: The case for personalised word complexity models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 353–365, Seattle, United States. Association for Computational Linguistics.

Isao Goto, Hideki Tanaka, and Tadashi Kumano. 2015. Japanese news simplification: tak design, data set construction, and analysis of simplified text. In *Proceedings of Machine Translation Summit XV: Papers*, Miami, USA.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – Simple Corpus for Medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Arthur C. Graesser, Danielle S. McNamara, Zhiqang Cai, Mark Conley, Haiying Li, and James Pennebaker. 2014. Coh-Metrix Measures Text Characteristics at Multiple Levels of Language and Discourse. *The Elementary School Journal*, 115(2):210–229.

Anke Grotlüschen, Klaus Buddeberg, Gregor Dutz, Lisanne Heilmann, and Christopher Stammer. 2020. Hauptergebnisse und Einordnung zur LEO-Studie 2018 – Leben mit geringer Literalität. In Anke Grotlüschen and Klaus Buddeberg, editors, *LEO 2018 - Leben mit geringer Literalität*, pages 13–64. wbv Publikation, Bielefeld, Germany.

Anke Grotlüschen, Klaus Buddeberg, Gregor Dutz, Lisanne Heilmann, and Christopher Stammer. 2020. Low literacy in Germany. Results from the second German literacy survey. *European journal for Research on the Education and Learning of Adults*, 11(1):127–143.

Silke Gutermuth. 2020a. *Leichte Sprache für alle?: Eine zielgruppenorientierte Rezeptionsstudie zu Leichter und Einfacher Sprache*, volume 5 of *Easy – Plain – Accessible*. Frank & Timme, Berlin, Germany.

Silke Gutermuth. 2020b. Textkorpus. In *Leichte Sprache für alle?: Eine zielgruppenorientierte Rezeptionsstudie zu Leichter und Einfacher Sprache*, chapter Appendix A, pages 293–308. Frank & Timme, Berlin, Germany.

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.

Silvia Hansen-Schirra, Walter Bisang, Arne Nagels, Silke Gutermuth, Julia Fuchs, Liv Borghardt, Silvana Deilen, Anne-Kathrin Gros, Laura Schiffl, and Johanna Sommer. 2020a. Intralingual Translation into Easy Language - or how to reduce cognitive processing costs. In Silvia Hansen-Schirra and Christiane Maaß, editors, *Easy Language Research: Text and User Perspectives*, volume 2 of *Easy - Plain - Accessible*, pages 197 – 226. Frank & Timme, Berlin, Germany.

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

Silvia Hansen-Schirra, Jean Nitzke, Silke Gutermuth, Christiane Maaß, and Isabel Rink. 2020b. Technologies for the Translation of Specialised Texts into Easy Language. In Silvia Hansen-Schirra and Christiane Maaß, editors, *Easy Language Research: Text and User Perspectives*, volume 2 of *Easy - Plain - Accessible*, pages 99 – 130. Frank & Timme, Berlin, Germany.

Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy. Association for Computational Linguistics.

Renate Hauser, Jannis Vamvas, Sarah Ebling, and Martin Volk. 2022. A multilingual simplified language news corpus. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 25–30, Marseille, France. European Language Resources Association.

Katarina Heimann Mühlenbock. 2008. Readable, legible or plain words - presentation of an easy-to-read Swedish corpus. In *Multilingualism: Proceedings of the 23rd Scandinavian Conference of Linguistics*, volume 8, page 325–327, Uppsala, Sweden. Acta Universitatis Upsaliensis.

Katarina Heimann Mühlenbock. 2013. *I see what you mean – Assessing readability for specific target groups*. Phd thesis, University of Gothenburg, Sweden.

David Heineman, Yao Dou, Mounica Maddela, and Wei Xu. 2023. Dancing between success and failure: Edit-level simplification evaluation using SALSA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3466–3495, Singapore. Association for Computational Linguistics.

Krishna-Sara Helmle. 2017. Leichte Sprache für Institutionen des Arbeitsmarkts – Handreichung zur Einführung Leichter Sprache im Rahmen von Interkulturellen Öffnungsprozessen. `https://klever-iq.de/wp-content/uploads/sites/15/2018/10/IQ_Leichte_Sprache_A4.pdf`. [Online; Last Change: *n/a*, Last Access: 2024-07-29].

Hendrik Heuer, David Fröhlich, Verena Riegler, Michael Radeka, and Julia Gspandl. 2024. Auditing the Text Understandability of German Public Administration Websites. Working Paper 48, Zentrum für Medien-, Kommunikations- und Informationsforschung (ZeMKI), Bremen, Germany.

Freya Hewett. 2023. APA-RST: A text simplification corpus with RST annotations. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse* (*CODI 2023*), pages 173–179, Toronto, Canada. Association for Computational Linguistics.

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.

Freya Hewett and Manfred Stede. 2022. Lexica corpus (v2.0). Zenodo repository: https://doi.org/10.5281/zenodo.6319803. [Online; Last Change: 2022-03-01; Last Access: 2024-07-23].

Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, Surafel Lakew, Prashant Mathur, Anna Currey, and Marcello Federico. 2022. Sockeye 3: Fast Neural Machine Translation with PyTorch. *Preprint*, arXiv:2207.05851.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas* (*Volume 1: Research Track*), pages 200–207, Boston, MA. Association for Machine Translation in the Americas.

Daniel Holmer and Evelina Rennes. 2023. Constructing pseudo-parallel Swedish sentence corpora for automatic text simplification. In *Proceedings of the 24th Nordic Conference on Computational Linguistics* (*NoDaLiDa*), pages 113–123, Tórshavn, Faroe Islands. University of Tartu Library.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Yichen Huang and Ekaterina Kochmar. 2024. REFeREE: A REference-FREE model-based metric for text simplification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (*LREC-COLING 2024*), pages 13740–13753, Torino, Italia. ELRA and ICCL.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning sentences from standard Wikipedia to Simple Wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217, Denver, Colorado. Association for Computational Linguistics.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics* (*GEM*), pages 205–223, Singapore. Association for Computational Linguistics.

Inclusion Europe. 2021. Language versions of easy-to-read standards. https://www.inclusion-europe.eu/easy-to-read-standards-guidelines/. [Online; Last Change: 2021-10-06 ; Last Access: 2024-07-11].

Inclusion Europe. 2024. Easy-to-read explanations: Easy-to-read. https://www.inclusion-europe.eu/easy-to-read-term/#ETR. [Online; Last Change: *n/a*; Last Access: 2021-04-11].

International Plain Language Federation. 2024. Plain language Definitions. https://www.iplfederation.org/plain-language/. [Online; Last Change: *n/a*; Last Access: 2024-04-11].

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2020. Best practices for crowd-based evaluation of German summarization: Comparing crowd, expert and automatic evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 164–175, Online. Association for Computational Linguistics.

Sarah Jablotschkin, Elke Teich, and Heike Zinsmeister. 2024. DE-lite - a new corpus of easy German: Compilation, exploration, analysis. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 106–117, St. Julian's, Malta. Association for Computational Linguistics.

Sarah Jablotschkin and Heike Zinsmeister. 2020. LeiKo: A corpus of easy-to-read German. Poster presentation at the Computational Linguistics Poster Session in the course of the 42nd annual conference of the Deutsche Gesellschaft für Sprachwissenschaft (DGfS) in Hamburg.

Daniel Jach. 2020. Korpus Einfaches Deutsch (KED 1.0). https://daniel-jach.github.io/simple-german/simple-german.html. [Online; Last Change: 2020-11-13, Last Access: 2022-10-20; Not Available Anymore].

Daniel Jach. 2023. Korpus Einfaches Deutsch (KED 2.0). https://daniel-jach.github.io/simple-german/simple-german.html. [Online; Last Change: 2023-11-09; Last Access: 2024-02-08; Not Available Anymore].

Susanne J. Jekat, Esther Germann, Alexa Lintner, and Corinne Soland. 2017. Wahlprogramme in Leichter Sprache: Eine korpuslinguistische Annäherung. In Bettina M. Bock, Ulla Fix, and Daisy Lange, editors, *"Leichte Sprache" im Spiegel theoretischer und angewandter Forschung*, Kommunikation - Partizipation - Inklusion. Frank & Timme Verlag für wissenschaftliche Literatur, Berlin, Germany.

Susanne J. Jekat, Heike E. Jüngst, Klaus Schubert, and Claudia Villiger, editors. 2014. *Sprache barrierefrei gestalten: Perpektiven aus der Angewandten Linguistik*, volume 69 of *TransÜD*. Frank & Timme, Berlin, Germany.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

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

John E. Joseph and Frederick J. Newmeyer. 2012. 'All Languages Are Equally Complex' – The rise and fall of a consensus. *Historiographia Linguistica*, 39(2/3):341 – 368.

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2023. Multilingual simplification of medical texts. In *Proceedings of the*

2023 Conference on Empirical Methods in Natural Language Processing, pages 16662–16692, Singapore. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2nd Edition)*, pearson education international edition, chapter 4.3: N-Grams – Training and Test Sets. Pearson Prentice Hall, Upper Saddle River, New Jersey, USA.

Daniel Jurafsky and James H. Martin. 2024a. Speech and Language Processing (3rd Edition Draft), chapter 10: Transformers and Large Language Models. https://web.stanford.edu/~jurafsky/slp3/10.pdf. [Online; Last Change: 2024-02-03, Last Access: 2024-06-25].

Daniel Jurafsky and James H. Martin. 2024b. Speech and Language Processing (3rd Edition Draft), chapter 11: Fine-tuning and Masked Language Models. https://web.stanford.edu/~jurafsky/slp3/11.pdf. [Online; Last Change: 2024-02-03, Last Access: 2024-06-25].

Daniel Jurafsky and James H. Martin. 2024c. Speech and Language Processing (3rd Edition Draft), chapter 9: RNNs and LSTMs. https://web.stanford.edu/~jurafsky/slp3/9.pdf. [Online; Last Change: 2024-02-03, Last Access: 2024-06-25].

Elisabeth Kaban and Sina Krottmaier. 2023. Das eABGB. *RuZ - Recht und Zugang*, 4(2):164–180.

Akihiro Katsuta and Kazuhide Yamamoto. 2018. Crowdsourced corpus of sentence simplification with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tannon Kew, Marek Kostrzewa, and Sarah Ebling. 2023. 20 minuten: A multi-task news summarisation dataset for German. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 1–13, Neuchatel, Switzerland. Association for Computational Linguistics.

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

Kindermann Verlag. 2024. Über uns – Kindermann Verlag . https://www.kindermannverlag.de/ueber-uns/. [Online; Last Change: *n/a*; Last Access: 2024-06-21].

W. Kintsch, E. Kozminsky, W.J. Streby, G. McKoon, and J.M. Keenan. 1975. Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior*, 14(2):196–214.

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

Klar & Deutlich – Agentur für Einfache Sprache. 2018. Wie erreicht man Klarheit? Mit verständlichen Texten in Einfacher Sprache. http://www.klarunddeutlich.de/html/img/pool/Flyer_Klar___Deutlich.pdf. [Online; Last Change: *n/a*, Last Access: 2020-01-02; Not Available Anymore].

Ruben Klepp. 2022a. Klassifizierung der Textkomplexität von Chatbot-Antworten mittels Transformer-Modellen. Master thesis, Hochschule Darmstadt, Germany.

Ruben Klepp. 2022b. Transformer Text Readability Classification . GitHub repository: https://github.com/krupper/transformer-text-readability-classification. [Online; Last Change: 2022-11-17, Last Access: 2023-04-19].

Sigrid Klerke and Anders Søgaard. 2012. DSim, a Danish parallel corpus for text simplification. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 4015–4018, Istanbul, Turkey. European Language Resources Association (ELRA).

Klexikon contributors. 2023. Klexikon:Presse – Wie umfangreich ist das Klexikon inzwischen? https://klexikon.zum.de/wiki/Klexikon:Presse#Wie_umfangreich_ist_das_Klexikon_inzwischen? [Online; Last Change: 2023-12-11; Last Access: 2024-04-22].

Lars Klöser, Mika Beele, Jan-Niklas Schagen, and Bodo Kraft. 2024. German text simplification: Finetuning large language models with semi-synthetic data. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 63–72, St. Julian's, Malta. Association for Computational Linguistics.

Christian Kneil, Julia Matousek, Franz Spiegelfeld, and Alexandra Roth. 2020. Einfach Verständlich – Erfolgrech Kommunizieren mit zielgruppengerechter Sprache. https://apa.at/whitepaper/einfach-verstaendlich-erfolgreich-kommunizieren-mit-zielgruppengerechter-sprache-juni-2020/. [Online; Last Change: 2020-06, Last Access: 2021-03-01].

Tomonori Kodaira, Tomoyuki Kajiwara, and Mamoru Komachi. 2016. Controlled and balanced dataset for Japanese lexical simplification. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 1–7, Berlin, Germany. Association for Computational Linguistics.

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

Anaïs Koptient and Natalia Grabar. 2020. Fine-grained text simplification in French: steps towards a better grammaticality. In *Proceedings of the 18th International Symposium on Health Information Management Research (ISHIMR)*, Kalmar, Sweden.

Reno Kriz, João Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. 2019. Complexity-weighted loss and diverse reranking for sentence simplification. 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 3137–3147, Minneapolis, Minnesota. Association for Computational Linguistics.

Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova. 2020. Iterative edit-based unsupervised sentence simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7918–7928, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6365–6378, Online. Association for Computational Linguistics.

Daisy Lange. 2018. Comparing "Leichte Sprache", "einfache Sprache" and "Leicht Lesen": A Corpus-Based Descriptive Approach. In *Proceedings of the 1st Swiss Conference on Barrier-free Communication (BfC 2018)*, pages 75–92, Winterthur, Switzerland. ZHAW Zürcher Hochschule für Angewandte Wissenschaften.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2022. The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset. In *Proceedings of Advances in Neural Information Processing Systems* (*NeurIPS*), volume 35, pages 31809–31826. Curran Associates, Inc.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications* (*BEA 2023*), pages 1–19, Toronto, Canada. Association for Computational Linguistics.

Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813, Dublin, Ireland. Association for Computational Linguistics.

Eric Lemgen. 2024. Error analysis of automatically generated text-simplifications in German. Bachelor thesis, Heinrich Heine University Düsseldorf, Germany.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak ..., and Vaishaal Shankar. 2024. DataComp-LM: In search of the next generation of training sets for language models. *Preprint*, arXiv:2406.11794.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):5–55.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhe Lin, Yitao Cai, and Xiaojun Wan. 2021. Towards document-level paraphrase generation with sentence rewriting and reordering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1033–1044, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhe Lin and Xiaojun Wan. 2021. Neural Sentence Simplification with Semantic Dependency Information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35 (15), pages 13371–13379, Online. AAAI Press.

Han Liu, Alexander Gegov, and Mihaela Cocea. 2017. Rule Based Networks: An Efficient and Interpretable Representation of Computational Models. *Journal of Artificial Intelligence and Soft Computing Research*, 7(2):111–123.

Lei Liu and Min Zhu. 2022. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55(9):1–35.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Junru Lu, Jiazheng Li, Byron Wallace, Yulan He, and Gabriele Pergola. 2023. NapSS: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1079–1091, Dubrovnik, Croatia. Association for Computational Linguistics.

Yuan Ma, Sandaru Seneviratne, and Elena Daskalaki. 2022. Improving text simplification with factuality error detection. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 173–178, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Christiane Maaß. 2015a. *Leichte Sprache: Das Regelbuch*, volume 1 of *Barrierefreie Kommunikation*, chapter III – 7. Übersetzen in Leichte Sprache. LiT, Münster, Germany.

Christiane Maaß. 2015b. *Leichte Sprache: Das Regelbuch*, volume 1 of *Barrierefreie Kommunikation*. LiT, Münster, Germany.

Christiane Maaß. 2020. *Easy Language – Plain Language – Easy Language Plus: Balancing Comprehensibility and Acceptability*, volume 3 of *Easy – Plain – Accessible*. Frank & Timme, Berlin, Germany.

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

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Margot Madina, Itziar Gonzalez-Dios, and Melanie Siegel. 2023. Easy-to-Read Language Resources and Tools for Three European Languages. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '23, page 693–699, New York, NY, USA. Association for Computing Machinery.

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

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Sabine Manning. 2023. KI-Tools für Einfache Sprache: (1) Klartext St. Pauli im Test. Blog article: https://multisprech.org/2023/07/26/ki-tools-fuer-einfache-sprache-1-st-pauli-im-test/. [Online; Last Change: 2023-07-26, Last Access: 2024-03-18].

Sabine Manning. 2024. Klar und Verständlich (K&V): Ein KI-Tool für Einfache Sprache. Jetzt testen! Blog article: https://multisprech.org/2024/06/27/klar-und-verstaendlich-ein-ki-tool-fuer-einfache-sprache/. [Online; Last Change: 2024-06-27; Last Access: 2024-07-04].

Raymond A. Mar, Jingyuan Li, Anh T. P. Nguyen, and Cindy P. Ta. 2021. Memory and comprehension of narrative versus expository texts: A meta-analysis. *Psychonomic Bulletin & Review*, 28(3):732–749.

Laura Marmit. 2020. Integrierte Titel in Leichter Sprache für prälinguale Gehörlose. In Katharina Oster Anne-Kathrin Gros, Silke Gutermuth, editor, *Leichte Sprache? Empirische und multimodale Perspektiven*, pages 87–104. Frank & Timme, Berlin, Germany.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Éric de La Clergerie, Antoine Bordes, and Benoît Sagot. 2018. Reference-less quality estimation of text simplification systems. In *Proceedings of the 1st Workshop on Automatic Text Adaptation* (*ATA*), pages 29–38, Tilburg, the Netherlands. Association for Computational Linguistics.

Tania Josephine Martin, José Ignacio Abreu Salas, and Paloma Moreda Pozo. 2023. A Review of Parallel Corpora for Automatic Text Simplification. Key Challenges Moving Forward. In *Proceedings of the 28th International Conference on Applications of Natural Language to Information Systems* (*NLDB*), pages 62–78, Cham, Switzerland. Springer Nature Switzerland.

Takumi Maruyama and Kazuhide Yamamoto. 2018. Simplified corpus with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (*LREC 2018*), Miyazaki, Japan. European Language Resources Association (ELRA).

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

Tony McEnery and Richard Xiao. 2007. *Chapter 2. Parallel and Comparable Corpora: What is Happening?*, pages 18–31. Multilingual Matters, Bristol, Great Britain.

Michael K. McKenna and Robinson Richard D. 1990. Content Literacy: A Definition and Implications. *Journal of Reading*, 34(3):184–186.

Fabian Meister. 2023. ABGB-TextSimplification-Datasets. GitHub repository: https://github.com/MeisterFa/ABGB-TextSimplification-Datasets. [Online; Last Change: 2023-09-24; Last Access: 2023-12-01].

Merriam-Webster.com. 2024. Easy. Thesaurus. https://www.merriam-webster.com/thesaurus/easy. [Online; Last Change: *n/a*; Last Access: 2024-04-11].

Heidi Anne Mesmer, James W. Cunningham, and Elfrieda H. Hiebert. 2012. Toward a Theoretical Model of Text Complexity for the Early Grades: Learning From the Past, Anticipating the Future. *Reading Research Quarterly*, 47(3):235–258.

Matti Miestamo. 2008. Grammatical complexity in cross-linguistic perspective. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity – Typology, contact, change*, page 23–41. John Benjamins Publishing Company, Amsterdam, The Netherlands.

MiniKlexikon contributors. 2024. MiniKlexikon – das Kinderlexikon für Leseanfänger. https://miniklexikon.zum.de/wiki/Hauptseite. [Online; Last Change: 2024-03-28; Last Access: 2024-04-22].

Benjamin Minixhofer. 2020. GerPT2: German large and small versions of GPT2. GitHub repository: https://github.com/bminixhofer/gerpt2. [Online; Last Change: 2020-12-27; Last Access: 2024-07-29].

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Salar Mohtaj, Babak Naderi, Kaspar Ensikat, and Sebastian Möller. 2019. A Dataset for Subjective Assessment of German Text Complexity. In *Proceedings of the Workshop on Dialog for Good - Workshop on Speech and Language Technology Serving Society*, pages 1–6, Stockholm, Sweden.

Salar Mohtaj, Babak Naderi, and Sebastian Möller. 2022. Overview of the GermEval 2022 shared task on text complexity assessment of German text. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 1–9, Potsdam, Germany. Association for Computational Linguistics.

Joss Moorkens and Dave Lewis. 2020. Copyright and the re-use of translation as data . In Minako O'Hagan, editor, *The Routledge Handbook of Translation and Technology*, pages 469–481. Routledge.

Jerome L Myers, Makiko Shinjo, and Susan A Duffy. 1987. Degree of causal relatedness and memory. *Journal of Memory and Language*, 26(4):453–465.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective Assessment of Text Complexity: A Dataset for German Language. *Preprint*, arXiv:1904.07733.

Joel T. Nadler, Rebecca Weston, and Elora C. Voyles. 2015. Stuck in the Middle: The Use and Interpretation of Mid-Points in Items on Questionnaires. *The Journal of General Psychology*, 142(2):71–89. PMID: 25832738.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2115, Austin, Texas. Association for Computational Linguistics.

Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pages 435–445, Baltimore, Maryland. Association for Computational Linguistics.

Shashi Narayan and Claire Gardent. 2016. Unsupervised sentence simplification using deep semantics. In *Proceedings of the 9th International Natural Language Generation conference*, pages 111–120, Edinburgh, UK. Association for Computational Linguistics.

Netzwerk Leichte Sprache. 2022. Die Regeln für Leichte Sprache. https://www.leichte-sprache.org/wp-content/uploads/2017/11/Regeln_Leichte_Sprache.pdf. [Online; Last Update: *n/a*; Last Access: 2024-07-29].

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2019a. DisSim: A discourse-aware syntactic text simplification framework for English and German. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 504–507, Tokyo, Japan. Association for Computational Linguistics.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2019b. Transforming complex sentences into a semantic hierarchy. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3415–3427, Florence, Italy. Association for Computational Linguistics.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2023. Discourse-Aware Text Simplification: From Complex Sentences to Linked Propositions. *Preprint*, arXiv:2308.00425.

Christina Niklaus, André Freitas, and Siegfried Handschuh. 2019c. MinWikiSplit: A sentence splitting corpus with minimal propositions. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 118–123, Tokyo, Japan. Association for Computational Linguistics.

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

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable text simplification with lexical constraint loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266, Florence, Italy. Association for Computational Linguistics.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (*Volume 2: Short Papers*), pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text Simplification by Tagging. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–25, Online. Association for Computational Linguistics.

Ingelore Oomen-Welke. 2015. Leichte Sprache, Einfache Sprache und Deutsch als Zweitsprache. *Didaktik Deutsch*, 20(38):24–32.

OpenAI. 2024. ChatGPT. https://chat.openai.com/. [Online; Last Change: *n/a*; Last Access: March 2024].

Lucía Ormaechea and Nikos Tsourakis. 2023. Extracting sentence simplification pairs from French comparable corpora using a two-step filtering method. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 30–40, Neuchatel, Switzerland. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of Advances in Neural Information Processing Systems* (*NeurIPS*), volume 35, pages 27730–27744. Curran Associates, Inc.

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

Gustavo Paetzold and Lucia Specia. 2016a. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (*SemEval-2016*), pages 560–569, San Diego, California. Association for Computational Linguistics.

Gustavo H. Paetzold and Lucia Specia. 2013. Text simplification as tree transduction. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Gustavo Henrique Paetzold and Lucia Specia. 2016b. Unsupervised Lexical Simplification for Non-native Speakers. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3761–3767, Palo Alto, California. AAAI Press.

Gustavo Henrique Paetzold and Lucia Specia. 2016c. Vicinity-Driven Paragraph and Sentence Alignment for Comparable Corpora. *Preprint*, arXiv:1612.04113.

Gabriele Pallotti. 2015. A simple view of linguistic complexity. *Second Language Research*, 31(1):117–134.

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

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

Sora Park. 2012. Dimensions of Digital Media Literacy and the Relationship with Social Exclusion. *Media International Australia*, 142(1):87–100.

Passanten Verlag. 2024. Passanten Verlag | Einfach Lesen. https://www.passanten-verlag.de/. [Online; Last Change: *n/a*; Last Access: 2024-06-21].

Sarah E. Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Proceedings of Speech and Language Technology in Education (SLaTE 2007)*, pages 69–72, Farmington, PA, USA. International Speech Communication Association (ISCA).

Björn Plüster. 2023. LeoLM: Igniting German- Language LLM Research. https://laion.ai/blog/leo-lm/. [Online; Last Change: 2023-09-28, Last Access: 2024-03-18].

David Ponce, Thierry Etchegoyhen, Jesús Calleja Pérez, and Harritxu Gete. 2024. Split and Rephrase with Large Language Models. *Preprint*, arXiv:2312.11075.

Maja Popović, Sheila Castilho, Rudali Huidrom, and Anya Belz. 2022. Reproducing a manual evaluation of the simplicity of text simplification system outputs. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 80–85, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Namoos Hayat Qasmi, Haris Bin Zia, Awais Athar, and Agha Ali Raza. 2020. SimplifyUR: Unsupervised lexical text simplification for Urdu. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3484–3489, Marseille, France. European Language Resources Association.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Stephen Reder. 2017. Adults' engagement in reading, writing and numeracy practices. In Anke Grotlüschen, David Mallows, Stephen Reder, and John Sabatini, editors, *Adults with Low Proficiency in Literacy or Numeracy*, chapter Chapter 3. Skill Use: Engagement in Reading, Writing and Numeracy Practices. OECD Publishing.

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

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013. Simplify or help? text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, W4A '13, New York, NY, USA. Association for Computing Machinery.

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.

Miguel Rios, Wilker Aziz, and Lucia Specia. 2011. TINE: A metric to assess MT adequacy. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 116–122, Edinburgh, Scotland. Association for Computational Linguistics.

Robert Koch-Institut. 2024a. COVID-19 (Coronavirus SARS-CoV-2). https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/nCoV_node.html. [Online; Last Change: 2024-05-30; Last Access: 2024-06-21].

Robert Koch-Institut. 2024b. Informationen zum Corona-Virus in Leichter Sprache. https://www.rki.de/DE/Service/Leichte-Sprache/LS_Corona-Ratgeber_tab-gesamt.html. [Online; Last Change: *n/a*; Last Access: 2024-06-21].

Michael J Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.

Paul Röttger, Fabio Pernisi, Bertie Vidgen, and Dirk Hovy. 2024. SafetyPrompts: a Systematic Review of Open Datasets for Evaluating and Improving Large Language Model Safety. *Preprint*, arXiv:2404.05399.

Horacio Saggion. 2017. *Automatic text simplification*, volume 32 of *Synthesis lectures on human language technologies*. Morgan & Claypool Publishers, San Rafael, California, USA.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarević. 2015. Making It Simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Transactions on Accessible Computing*, 6(4):1–36.

Andrey Sakhovskiy, Alexandra Izhevskaya, Alena Pestova, Elena Tutubalina, Valentin Malykh, Ivan Smurov, and Ekaterina Artemova. 2021. RuSimpleSentEval-2021 shared task: evaluating sentence simplification for Russian. In *Proceedings of the Annual International Conference "Dialogue" (2021)*, pages 607–617, Online.

Andreas Säuberli and Simon Clematide. 2024. Automatic generation and evaluation of reading comprehension test items with large language models. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024*, pages 22–37, Torino, Italia. ELRA and ICCL.

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.

Andreas Säuberli, Franz Holzknecht, Patrick Haller, Silvana Deilen, Laura Schiffl, Silvia Hansen-Schirra, and Sarah Ebling. 2024. Digital Comprehensibility Assessment of Simplified Texts among Persons with Intellectual Disabilities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Danielle Saunders. 2022. Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation: A Survey. *Journal of Artificial Intelligence Research*, 75.

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. Text simplification from professionally produced corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Carolina Scarton, Alessio Palmero Aprosio, Sara Tonelli, Tamara Martín Wanton, and Lucia Specia. 2017. MUSST: A multilingual syntactic simplification tool. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 25–28, Tapei, Taiwan. Association for Computational Linguistics.

Finja Scheele. 2020. Multimodale Text im barrierefreien Museum: Audioguides in Leichter Sprache. In Katharina Oster Anne-Kathrin Gros, Silke Gutermuth, editor, *Leichte Sprache? Empirische und multimodale Perspektiven*, pages 137–156. Frank & Timme, Berlin, Germany.

Inga Schiffler. 2022. *Das Prüfen auf dem Prüfstand: Die Rolle der Moderatorinnen beim Prüfen von Texten in Leichter Sprache*, 1 edition, volume 8 of *Kommunikation – Partizipation – Inklusion*. Frank & Timme GmbH, Berlin, Germany.

Tim Schlippe and Katharina Eichinger. 2023. Multilingual Text Simplification and Its Performance on Social Sciences Coursebooks. In *Artificial Intelligence in Education Technologies: New Development and Innovative Practices*, pages 119–136, Singapore. Springer Nature Singapore.

Thorben Schomacker, Miriam Anschütz, Regina Stodden, Georg Groh, and Marina Tropmann-Frick. 2024. Overview of the GermEval 2024 shared task on statement segmentation in German easy language (StaGE). In *Proceedings of GermEval 2024 Shared Task on Statement Segmentation in German Easy Language (StaGE)*, pages 1–14, Vienna, Austria. Association for Computational Lingustics.

Thorben Schomacker, Tillmann Dönicke, and Marina Tropmann-Frick. 2023a. Exploring automatic text simplification of German narrative documents. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 139–148, Ingolstadt, Germany. Association for Computational Lingustics.

Thorben Schomacker, Michael Gille, Marina Tropmann-Frick, and Jörg von der Hülls. 2023b. Data and approaches for German text simplification – towards an accessibility-enhanced communication. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 63–68, Ingolstadt, Germany. Association for Computational Lingustics.

Max Schwarzer. 2018. Crowdsourcing Text Simplification with Sentence Fusion. Bachelor thesis, Pomona College, Claremont, California, USA.

Max Schwarzer, Teerapaun Tanprasert, and David Kauchak. 2021. Improving human text simplification with sentence fusion. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 106–114, Mexico City, Mexico. Association for Computational Linguistics.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Stefan Schweter. 2020. German GPT-2 model. Zenodo repository: `https://doi.org/10.5281/zenodo.4275046`; Version: 1.0.0. [Online; Last Change: 2020-11-16; Last Access: 2023-12-01].

Thomas Scialom, Louis Martin, Jacopo Staiano, Éric Villemonte de la Clergerie, and Benoît Sagot. 2021. Rethinking Automatic Evaluation in Sentence Simplification. *Preprint*, arXiv:2104.07560.

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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Matthew Shardlow. 2014. A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications* (*BEA 2024*), pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties* (*READI*), pages 57–62, Marseille, France. European Language Resources Association.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation* (*SemEval-2021*), pages 1–16, Online. Association for Computational Linguistics.

Advaith Siddharthan. 2003. Preserving discourse structure when simplifying text. In *Proceedings of the 9th European Workshop on Natural Language Generation* (*ENLG-2003*) *at EACL 2003*, Budapest, Hungary. Association for Computational Linguistics.

Advaith Siddharthan. 2006. Syntactic Simplification and Text Cohesion. *Research on Language and Computation*, 4(1):77–109.

Advaith Siddharthan. 2011. Text simplification using typed dependencies: A comparision of the robustness of different generation strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 2–11, Nancy, France. Association for Computational Linguistics.

Advaith Siddharthan. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics*, 165(2):259–298.

Advaith Siddharthan and Angrosh Mandya. 2014. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 722–731, Gothenburg, Sweden. Association for Computational Linguistics.

Melanie Siegel, Dorothee Beermann, and Lars Hellan. 2019. Aspects of Linguistic Complexity: A German - Norwegian Approach to the Creation of Resources for Easy-To-Understand Language. In *Proceedings of the 11th International Conference on Quality of Multimedia Experience* (*QoMEX*), pages 1–3, Berlin, Germany. IEEE.

Ronald H. Southerland and Francis Katamba. 1997. Language in social contexts. In William O'Grady, Micheal Dobrovolsky, and Francis Katamba, editors, *Contemporary Linguistics: An Introduction*, 3rd edition, pages 540–590. Longman, London and New York.

Spaß am Lesen Verlag. 2024. Über den Spaß am Lesen Verlag. `https://einfachebuecher.de/UEber-uns/`. [Online; Last Change: *n/a*; Last Access: 2024-06-21].

Lucia Specia. 2010. Translating from Complex to Simplified Sentences. In Thiago Alexandre Salgueiro Pardo, António Branco, Aldebaro Klautau, Renata Vieira, and Vera Lúcia Strube de Lima, editors, *Computational processing of the Portuguese language*, volume 6001 of *Lecture notes in computer science Lecture notes in artificial intelligence*, pages 30–39. Springer, Berlin.

Nicolas Spring, Marek Kostrzewa, David Fröhlich, Annette Rios, Dominik Pfütze, Alessia Battisti, and Sarah Ebling. 2023. Analyzing sentence alignment for automatic simplification of German texts. In Silvana Deilen, Silvia Hansen-Schirra, Sergio Hernández Garrido, Maaß Christiane, and Anke Tardel, editors, *Emerging Fields in Easy Language and Accessible Communication Research*, 14, pages 339–369. Frank&Timme, Berlin, Germany.

Nicolas Spring, Marek Kostrzewa, Annette Rios, and Sarah Ebling. 2022. Ensembling and Score-Based Filtering in Sentence Alignment for Automatic Simplification of German Texts. In *Proceedings of the International Conference on Human-Computer Interaction* (*Volume 7: Universal Access in Human-Computer Interaction. Novel Design Approaches and Technologies*), pages 137–149, Cham. Springer International Publishing.

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

Stadt Aschaffenburg. 2024. Aktuelle Meldungen in Leichter Sprache. `https://www.aschaffenburg.de/Aktuelles/Aktuelle-Meldungen-in-Leichter-Sprache/DE_index_6736.html`. [Online; Last Change: 2024-07-03; Last Access: 2024-07-04].

Stadt Hamburg. 2024. Ein Computer-Programm übersetzt normale Texte in Leichte Sprache. `https://www.hamburg.de/ls-ueberetzung-576450`. [Online; Last Change: *n/a*; Last Access: 2024-07-04].

Sanja Štajner. 2018. How to make troubleshooting simpler? Assessing differences in perceived sentence simplicity by native and non-native speakers. In *Proceedings of the LREC 2018 Workshop "Improving Social Inclusion using NLP: Tools, Methods and Resources"* (*ISI-NLP 2*), pages 13–20, Miyazaki, Japan.

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

Sanja Štajner, Marc Franco-Salvador, Simone Paolo Ponzetto, Paolo Rosso, and Heiner Stuckenschmidt. 2017. Sentence Alignment Methods for Improving Text Simplification Systems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (*Volume 2: Short Papers*), pages 97–102, Vancouver, Canada. Association for Computational Linguistics.

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

Sanja Štajner and Goran Glavaš. 2017. Leveraging event-based semantics for automated text simplification. *Expert Systems with Applications*, 82:383–395.

Sanja Štajner, Maja Popović, and Hanna Béchara. 2016a. Quality estimation for text simplification. In *Proceedings of the Workshop & Shared Task on Quality Assessment for Text Simplification* (*QATS*), pages 15–21, Paris. ELRA-ERDA.

Sanja Štajner, Ruslan Mitkov, and Horacio Saggion. 2014. One step closer to automatic evaluation of text simplification systems. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations* (*PITR*), pages 1–10, Gothenburg, Sweden. Association for Computational Linguistics.

Sanja Štajner and Sergiu Nisioi. 2018. A detailed evaluation of neural sequence-to-sequence models for in-domain and cross-domain text simplification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (*LREC 2018*), Miyazaki, Japan. European Language Resources Association (ELRA).

Sanja Štajner, Maja Popović, Horacio Saggion, Lucia Specia, and Mark Fishel. 2016b. Shared Task on Quality Assessment for Text Simplification. In *Proceedings of the Workshop on Quality Assessment for Text Simplification* (*QATS*), pages 22–37, Portorož, Slovenia. Association for Computational Linguistics.

Sanja Štajner, Kim Cheng Sheang, and Horacio Saggion. 2022. Sentence Simplification Capabilities of Transfer-Based Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36 (11), pages 12172–12180. AAAI Press.

Statistisches Bundesamt. 2023. Bevölkerung: Deutschland, Stichtag – 12411 Fortschreibung des Bevölkerungsstandes. https://www-genesis.destatis.de/genesis//online?operation=table&code=12411-0001&bypass=true&levelindex=1&levelid=1712922974832#abreadcrumb. [Online; Last Change: *n/a*; Last Access: 2024-07-29].

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Regina Stodden. 2021a. Accessibility and comprehensibility of user-generated content: Challenges and chances for easy-to-understand languages. In *Proceedings of the 3rd Swiss Conference on Barrier-free Communication* (*BfC 2020*), pages 151–161, Winterthur (online). ZHAW Zürcher Hochschule für Angewandte Wissenschaften.

Regina Stodden. 2021b. Differences between German and English text simplification. Poster presentation at the Computational Linguistics Poster Session in the course of the 43rd Annual Conference of the German Linguistic Society (DGfS): Poster Session Computational Linguistics.

Regina Stodden. 2021c. When the Scale is Unclear – Analysis of the Interpretation of Rating Scales in Human Evaluation of Text Simplification. In *Proceedings of the First Workshop on Current Trends in Text Simplification* (*CTTS 2021*), pages 84–95, Online. CEUR-WS.

Regina Stodden. 2022. Creation of a parallel simplification corpus – Using the annotation tool TS-anno. Annotation guideline, Heinrich Heine University, Düsseldorf, Germany. Also available in German.

Regina Stodden. 2024a. EASSE-DE & EASSE-multi: Easier automatic sentence simplification evaluation for German & multiple languages. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability* (*TSAR 2024*), pages 107–116, Miami, Florida, USA. Association for Computational Linguistics.

Regina Stodden. 2024b. Reproduction & benchmarking of German text simplification systems. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 1–15, Torino, Italia. ELRA and ICCL.

Regina Stodden and Laura Kallmeyer. 2020. A multi-lingual and cross-domain analysis of features for text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties* (*READI*), pages 77–84, Marseille, France. European Language Resources Association.

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

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

Regina Stodden and Phillip Nguyen. 2024. Can text simplification help to increase the acceptance of E-participation? In *Proceedings of the First Workshop on Language-driven Deliberation Technology (DELITE) @ LREC-COLING 2024*, pages 20–32, Torino, Italia. ELRA and ICCL.

Regina Stodden and Gayatri Venugopal. 2021. RS_GV at SemEval-2021 task 1: Sense relative lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 640–649, Online. Association for Computational Linguistics.

Markus Stroh. 2024. Evaluation of automated instruction text simplification: comparing human performance and LLMs. Master thesis, Heinrich Heine University Düsseldorf, Germany.

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

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

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

Hong Sun and Ming Zhou. 2012. Joint learning of a dual SMT system for paraphrase generation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–42, Jeju Island, Korea. Association for Computational Linguistics.

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

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

Renliang Sun, Zhixian Yang, and Xiaojun Wan. 2023. Exploiting summarization data to help text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 39–51, Dubrovnik, Croatia. Association for Computational Linguistics.

Sai Surya, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. 2019. Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2058–2068, Florence, Italy. Association for Computational Linguistics.

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 (KONVENS 2016)*, volume 16 of *Bochumer Linguistische Arbeitsberichte (BLA)*, pages 279–287, Bochum, Germany.

Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016. Evaluating lexical simplification and vocabulary knowledge for learners of French: Possibilities of using the FLELex resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 230–236, Portorož, Slovenia. European Language Resources Association (ELRA).

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

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

Task Force Corona Leichte Sprache, Anne Leichtfuß, Kirsten Czerner-Nicolas, Simone Fass, Inga Kramer, Julia Bertmann, Natalie Dedreux, Sieglinde Didier, Christian Hehemann, Daniela Pindor, Anna-Lisa Plettenberg, Daniel Rauers, Johanna von Schönfeld, Paul Spitzeck, and Thomas Szymanowicz. 2024. Corona Leichte Sprache – Wissen über Corona in Leichter Sprache. `https://www.rki.de/DE/Service/Leichte-Sprache/LS_Corona-Ratgeber_tab-gesamt.html`. [Online; Last Change: *n/a*; Last Access: 2024-06-21].

Zachary W. Taylor, Maximus H. Chu, and Junyi Jessy Li. 2022. Text simplification of college admissions instructions: A professionally simplified and verified corpus. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6505–6515, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Wolfgang Teubert. 1996. Comparable or Parallel Corpora? *International Journal of Lexicography*, 9(3):238–264.

Boris Thome, Friederike Hertweck, and Stefan Conrad. 2024. Determining Perceived Text Complexity: An Evaluation of German Sentences Through Student Assessments. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 714–721, Atlanta, Georgia, USA. International Educational Data Mining Society.

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

Jörg Tiedemann. 2006. ISA & ICA - two web interfaces for interactive alignment of bitexts alignment of parallel texts. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Vanessa Toborek, Moritz Busch, Malte Boßert, Christian Bauckhage, and Pascal Welke. 2023. A new aligned simple German corpus. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pages 11393–11412, Toronto, Canada. Association for Computational Linguistics.

Amalia Todirascu, Thomas François, Nuria Gala, Cédrick Fairon, Anne-Laure Ligozat, and Delphine Bernhard. 2013. Coherence and Cohesion for the Assessment of Text Readability. In *Proceedings of 10th International Workshop on Natural Language Processing and Cognitive Science* (*NLPCS 2013*), pages 11–19, Marseille, France.

Petro Tolochko and Hajo Boomgaarden. 2019. Determining Political Text Complexity: Conceptualizations, Measurements, and Application. *International Journal of Communication*, 13(0).

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. SIMPITIKI: a Simplification corpus for Italian. In *Proceedings of Third Italian Conference on Computational Linguistics* (*CLiC-it 2016*), Napoli, Italy.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *Preprint*, arXiv:2307.09288.

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

Jan Trienes, Laura Vásquez-Rodríguez, and Tollef Emil Jørgensen. 2024. Text Simplification Datasets. Github repository: `https://github.com/jantrienes/text-simplification-datasets`. [Online; Last Change: 2024-06-5; Last Access: 2024-07-18].

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2014. Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL - International Journal of Applied Linguistics*, 165(2):194–222.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2021. The Role of Text Simplification Operations in Evaluation. In *Proceedings of the First Workshop on Current Trends in Text Simplification (CTTS 2021)*, pages 57–69, Online. CEUR-WS.

Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2023. Document-level text simplification with coherence evaluation. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 85–101, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc.

Sara Vecchiato. 2022. Clear, easy, plain, and simple as keywords for text simplification. *Frontiers in Artificial Intelligence*, 5.

Vikidia contributors. 2024. Vikidia. `https://en.vikidia.org/wiki/Main_Page`. [Online; Last Change: 2024-02-02; Last Access: 2024-04-22].

Beat Vollenwyder, Andrea Schneider, Eva Krueger, Florian Brühlmann, Klaus Opwis, and Elisa D. Mekler. 2018. How to Use Plain and Easy-to-Read Language for a Positive User Experience on Websites. In *Computers Helping People with Special Needs*, pages 514–522, Cham, Switzerland. Springer International Publishing.

Tu Vu, Baotian Hu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Sentence simplification with memory-augmented neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 79–85, New Orleans, Louisiana. Association for Computational Linguistics.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, Brussels, Belgium. Association for Computational Linguistics.

Zarah Weiss, Xiaobin Chen, and Detmar Meurers. 2021. Using broad linguistic complexity modeling for cross-lingual readability assessment. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 38–54, Online. LiU Electronic Press.

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.

Zarah Weiss and Detmar Meurers. 2022. Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference? In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 141–153, Seattle, Washington. Association for Computational Linguistics.

Wikipedia contributors. 2024a. Wikipedia – Content Licensing. `https://en.wikipedia.org/wiki/Wikipedia#Content_licensing`. [Online; Last Change: 2024-04-21; Last Access: 2024-04-22].

Wikipedia contributors. 2024b. Wikipedia – Language editions. `https://en.wikipedia.org/wiki/Wikipedia#Language_editions`. [Online; Last Change: 2024-04-21; Last Access: 2024-04-22].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Sascha Wolfer. 2015. Comprehension and comprehensibility. In Karin Maksymski, Silke Gutermuth, and Silvia Hansen-Schirra, editors, *Translation and comprehensibility*, 1 edition, volume 72 of *TRANSÜD. Arbeiten zur Theorie und Praxis des Übersetzens und Dolmetschens*, pages 33 – 51. Frank & Timme, Berlin, Germany.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK. Association for Computational Linguistics.

World Wide Web Consortium. 2008. Web Content Accessibility Guidelines (WCAG) 2.0. https://www.w3.org/TR/WCAG20/. [Online; Last Change: 2008-12-11 Last Access: 2024-06-20].

Wort & Bild Verlag. 2024. SUMM AI. https://www.apotheken-umschau.de/autor-in/summ-ai-1098033.html. [Online; Last Access: 2024-07-04].

Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. 2023. Elaborative simplification as implicit questions under discussion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5525–5537, Singapore. Association for Computational Linguistics.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Daichi Yamaguchi, Rei Miyata, Sayuka Shimada, and Satoshi Sato. 2023. Gauging the gap between human and machine text simplification through analytical evaluation of simplification strategies and errors. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 359–375, Dubrovnik, Croatia. Association for Computational Linguistics.

Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. 2022. Controllable text simplification with deep reinforcement learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing* (*Volume 2: Short Papers*), pages 398–404, Online only. Association for Computational Linguistics.

Diyi Yang, Ankur Parikh, and Colin Raffel. 2022. Learning with limited text data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 28–31, Dublin, Ireland. Association for Computational Linguistics.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Li Zhang, Huaiyu Zhu, Siddhartha Brahma, and Yunyao Li. 2020b. Small but mighty: New benchmarks for split and rephrase. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 1198–1205, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations* (*ICLR*), Online. OpenReview.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Xuan Zhang, Huizhou Zhao, KeXin Zhang, and Yiyang Zhang. 2020. SEMA: Text simplification evaluation through semantic alignment. In *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 121–128, Suzhou, China. Association for Computational Linguistics.

Xinran Zhao, Esin Durmus, and Dit-Yan Yeung. 2023. Towards reference-free text simplification evaluation with a BERT Siamese network architecture. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13250–13264, Toronto, Canada. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics* (*Coling 2010*), pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.

Plain Text Summary

In Germany, about 20% of people have trouble reading and writing texts. Those particularly affected by reading difficulties include, for example:

- Elderly people
- People who don't speak German as their native language
- People who read slowly
- People with intellectual disabilities

Many German texts are difficult for these people to understand because they are written too complex. This makes it hard for these people to find important information, such as medical texts, government texts, news, or general internet texts. But there is a solution: texts can be rewritten to adapt to the needs of the readers. As a result, they become easier to read for other people as well. Typical changes include, for example:

- shorter sentences,
- fewer subordinate clauses,
- replacing difficult or long words with shorter and easier ones,
- separating compound words,
- clearly stating things implied between the lines.

Usually, professional translators of translation agencies reformulate the texts. However, there are also computer programs that can automatically simplify texts. To a certain extent, these texts are then reformulated automatically using machine learning and artificial intelligence.

The process of automatically rewriting texts is also called "automatic text simplification". Automatic text simplification can be defined as: text simplification is the change of word choice in a text and/or the restructuring a sentence in a text. The change of word choice is often referred to as "lexical simplification". The restructuring of sentence structure is often referred to as "syntactic simplification". The simplification is considered automated if the process is supported by machine learning. The goal of simplification is not to change the original meaning of the difficult text, but to make it easier to understand for a specific target group.

The highest goal of automatic text simplification is: people with reading difficulties should be able to understand texts better. People with reading difficulties should be able to participate more in society and make more decisions about themselves. This goal should be achieved by rewriting difficult-to-understand texts to make them easier to understand. However, mistakes are often made during automatic text simplification. Therefore, there is an important intermediate step towards this goal: automatic models for text simplification should support professional translators. The automatically generated simplification will be presented to the translators as

a first proposal. The translators can then correct and adapt the text. This should make the simplification process faster and easier.

To develop models for text simplification that can simplify texts, you need examples of difficult texts and their easier versions. A collection of many examples is also called a dataset in this context. The model then learns how to simplify texts by going through this dataset. Afterwards, it is tested how well the model can create simplifications by processing another dataset. The automatically generated simplifications on the new dataset are then often compared to a simplification written by a human. If both simplifications are similar, the simplification is considered as good.

In this dissertation, I provide an overview of the current state of research on automatic text simplification of German texts. I explain how datasets are created, how the quality of simplifications is evaluated, and which simplification models already exist. I also point out the difficulties that arise and propose solutions. With this, I hope to advance the research on German text simplification.

One of the biggest challenges in this research area is that there are not enough good datasets to train machine learning models. Existing datasets often have too few examples or are not good enough. I try to reduce this problem in my work. To do this, I first investigated what makes a good simplification. I also analyzed how changes made during simplification can be named and grouped. This grouping helps, among other things, to evaluate the diversity and quality of datasets for text simplification more accurately.

To facilitate the evaluation and creation of new datasets for text simplification, I have designed a platform called TS-ANNO. On this platform, the names of the changes can be assigned to the respective text passages. TS-ANNO also supports the alignment of sentence pairs in texts. A sentence pair consists of a difficult sentence and a simplified sentence with the same meaning. These sentence pairs are necessary to build or train a simplification model for sentences. In contrast to simplifying entire documents, sentence simplification requires sentence pairs, not document pairs.

Using TS-ANNO, I have created new datasets for document simplification and sentence simplification for the German language. I call the dataset "DEplain"; the name consists of "DE" for German and "plain" refers to the simplified language variety called "plain language".

One part of the dataset consists of news texts (called DEplain-APA-doc and DEplain-APA-sent). The news texts were simplified by a translation agency. The simplified news texts are intended to be particularly understandable for people learning German. The other part of the dataset consists of internet texts (called DEplain-web-doc and DEplain-web-sent). Here, the simplified texts are intended to be particularly understandable for people with reading difficulties and intellectual disabilities. The simplification has partially written by trained translators. However, my new datasets are larger than previous datasets and have higher quality. Therefore, my datasets are better suited for automatic text simplification for German than previous datasets.

Another difficulty in researching automated German text simplification is evaluating the quality of the generated texts. The quality can be evaluated either by humans, by filling out questionnaires on simplification, or automatically by computer programs that count and compare certain properties. So far, the questionnaires were not detailed enough to fully evaluate the quality of simplified German texts. In my work, I present new aspects for the evaluation questionnaire. The new aspects relate to, for example:

- How well were long, difficult words replaced by short, simple ones? (Lexical simplification)
- How well was the sentence structure simplified? (Syntactic simplification)
- How well is the sentence understandable when you don't know the previous and following text? (Coherence and cohesion)

- How ambiguous is the content of the text? Can you easily misunderstand the text? (Ambiguity) or
- How many new pieces of information were added to the text? (Information enrichment)

On the other hand, automated text evaluation is often not reliable, transparent, and interpretable. Most methods for evaluation were developed for the English language. It has not been thoroughly tested whether and how well these methods work for other languages (such as German). Furthermore, many research studies use different settings for the automated methods for evaluation. This makes it hard to compare the results.

In my work, I present a program that can standardize the evaluation. The program is called EASSE-DE. EASSE-DE is specifically optimized for evaluating German texts, but it can also be applied to other languages. The evaluation with EASSE-DE is more transparent and interpretable than with previous programs. With EASSE-DE, models for German text simplification can be better compared, as the same settings can be used for evaluation, and the settings are suitable for German texts.

Many models for simplifying German texts only had a description in the publication of the researchers, but often there was no usable or available software or programming code. In my PhD thesis, I have closed this research gap. To do this, I rebuilt models from other researchers and made them publicly accessible. Additionally, I developed eight new models for automatic German text simplification. I trained and evaluated these models on my new dataset DEplain. Compared to models trained on other datasets, my models achieve better results in some cases. One reason for this is the different quality levels of the datasets. My models are also characterized by the following: The simplifications of one model are specifically tailored to a) people learning German and b) only simplify news texts. Other models mix target groups or news texts with other types of texts and thus achieve less precise simplifications. In my analysis of the models, I also show that similar approaches can be used for automatic simplification of documents and sentences.

My work highlights the possibilities and potentials of automatic simplification of German texts. However, the research in this area is still at a relatively early stage of development. My work facilitates and shows the way for new research directions in the area of German text simplification. The reason for this are the following innovations:

- With TS-ANNO, it is now easier to create and evaluate new datasets.
- Researchers can use my new datasets for simplifying German documents, sentences, news texts, and internet texts, evaluate and expand them.
- Researchers can use my new models for simplifying German documents, sentences, news texts, and internet texts, evaluate and expand them.
- The automatic evaluation of German simplifications is improved and standardized through EASSE-DE. Researchers can thus make better comparisons between German models for text simplification.
- Manual evaluation is now possible in more detail. To do this, I expanded the questionnaires that people can use to evaluate automatically generated simplifications. These questionnaires can be used and adapted by researchers.

Appendices

Appendix A

Comparison of Web Harvester for (Parallel) German TS Data

subcorpus	website simple	website complex	simple	complex	domain	description	SGWC '13	SGWC '23	Anschütz et al.	KED	GNATS	LeiKo	Klepp	DEPlain
EinfacheBücher	einfachebuecher.de/	projekt-gutenberg.org/	PL	SG/OG	fiction	Books in plain German					x			x
EinfacheBücherPassanten	passanten-verlag.de/	projekt-gutenberg.org/	PL	SG/OG	fiction	Books in plain German					x			x
ApothekenUmschau	apotheken-umschau.de/einfache-sprache/ ‡	apotheken-umschau.de/	PL	SG	health	Health magazine in which diseases are explained in plain German		x						x
BZFE	bzfe.de/einfache-sprache/ †	bzfe.de	PL	SG	health	Information of the German Federal Agency for Food on good nutrition								x
Alumniportal	alumniportal-deutschland.org/de/ †	alumniportal-deutschland.org	PL	PL	language learner	Texts related to Germany and German traditions written for language learners.								x
Lebenshilfe	lebenshilfe-main-taunus.de/	lebenshilfe-main-taunus.de/	EL	SG	accessibility	Non-profit association for disabled people	x	x	x					x
Bibel	offene-bibel.de/ †‡	offene-bibel.de/	EL	SG	bible	Bible texts in EL								x
NDR-Märchen	ndr.de/fernsehen/barrierefreie_angebote/leichte_sprache/Maerchen-in-Leichter-Sprache,maerchenleichtesprache100.html ‡	projekt-gutenberg.org/	EL	SG/OG	fiction	Fairytales in EL					x		x	x
EinfachTeilhaben	einfach-teilhaben.de/DE/LS/Home/leichtesprache_node.html	einfach-teilhaben.de	EL	SG	accessibility		x							x
StadtHamburg	hamburg.de/hamburg-barrierefrei/leichte-sprache/	hamburg.de	EL	SG	public authority	Information of and regarding the German city Hamburg								x
StadtKöln	stadt-koeln.de/leben-in-koeln/soziales/informationen-leichter-sprache	stadt-koeln.de	EL	SG	public authority	Information of and regarding the German city Cologne		x						x
Hurraki	hurraki.de/wiki/Hauptseite	*n/a*											x	
Einfachstars	einfachstars.info/	*n/a*											x	
nachrichtenleicht	nachrichtenleicht.de/	*n/a*	PL		news					x		x	x	
kurier.at	kurier.at/einfache-sprache				news									
NDR Nachrichten	ndr.de/fernsehen/barrierefreie_angebote/leichte_sprache/Nachrichten-in-Leichter-Sprache,nachrichtenleichtesprache100.html	*n/a*			news	State-funded public broadcasting service						x		
InfoEasy	infoeasy-news.ch/	*n/a*			news									
Behindertenbeauftragter	behindertenbeauftragter.de/DE/AS/startseite/startseite-node.html	behindertenbeauftragter.de	EL	SG	accessibility	Official office for disabled people		x						
brandeins	brandeins.de/themen/rubriken/leichte-sprache	*n/a*	EL	SG		Translating excerpts from various topics		x						
MDR nachrichten	mdr.de/nachrichten-leicht/index.html	*n/a*	EL	SG	news	State-funded public broadcasting service		x					x	
Sozialpolitik	sozialpolitik.com/es	sozialpolitik.com/	EL	SG		Explains social policy in Germany		x						
TAZ	taz.de/Politik/Deutschland/Leichte-Sprache/!p5097/	*n/a*	EL	SG		German Newspaper (discontinued)		x				x		
Gemeinnützige Werkstätten und Wohnstätten GmbH Sindelfingen	gww-netz.de/de-LS/	gww-netz.de			accessibility	Non-profit association in social sector	x							
Heilpädagogische Hilfe Osnabrück	*n/a*	os-hho.de			accessibility	orthopaedagogical support	x							
Oberschwäbische Werkstätten GmbH	*n/a*	owb.de			accessibility	Non-profit association in social sector	x							
Hanisauland	hanisauland.de/index.html	*n/a*			mixed					x				
Klexikon	klexikon.zum.de/	*n/a*			wikipedia					x			x	
Labbe	labbe.de/lesekorb	*n/a*			fiction					x				
Oekoleo	oekoleo.de/	*n/a*								x				
Rechte-einfach	mixed	*n/a*			law					x				
rossipotti	rossipotti.de/	*n/a*			fiction					x				
simplescience	simplyscience.ch/home.html	*n/a*								x				

Table A.1: Webpages per German web TS corpus. EL = German Easy Language, PL = German Plain Language, SG = Standard German, OG = Old German. All URLs have lastly been accessed at July 24, 2024. Part I (continued on next page).

subcorpus	website simple	website complex	simple	complex	domain	description	SGWC '13	SGWC '23	Anschütz et al.	KED	GNATS	LeiKo	Klepp	DEPlain
Kindermann Verlag	kindermannverlag.de/	*n/a*			fiction						x			
SR Nachrichten	sr.de/sr/home/nachrichten/nachrichten_einfach/nachrichten_einfach100.html	*n/a*			news							x		
Arbeit & Gesundheit – Das Portal für Sicherheitsbeauftragte	aug.dguv.de/leichte-sprache/	https://aug.dguv.de/	EL	SG	health								x	
Bibel in LS	evangelium-in-leichter-sprache.de/		EL	x	bible								x	
Einfach Politik: Lexikon	bpb.de/kurz-knapp/lexika/lexikon-in-einfacher-sprache/				politics								x	
Stadt Monheim	monheim.de/footer/leichte-sprache/inhalts-uebersicht		EL	SG	public authority									

Table A.2: Webpages per German web TS corpus. EL = German Easy Language, PL = German Plain Language, SG = Standard German, OG = Old German. All URLs have lastly been acccessed at July 24, 2024. Part II (continued from previous page).

Appendix B

Examples of Simplification Plans

B.1 Document Texts for the Simplification Plan of the Alumniportal

The following text in Table B.1 has been extracted from https://www.alumniportal-deutschland.org/digitales-lernen/deutsche-sprache/deutsch-auf-die-schnelle/online-deutsch-lernen-spielen/ at July 1, 2021. This text is protected by copyright by Kooperation Alumniportal Deutschland.

ID	Complex Text (CEFR level B1)	ID	Simple Text (CEFR level A2)
0	"Die Quelle alles Guten liegt im Spiel."	0	"Die Quelle alles Guten liegt im Spiel."
1	Das wusste schon der deutsche Pädagoge Friedrich Wilhelm August Fröbel (1782 - 1852).	1	Das wusste schon der deutsche Pädagoge Friedrich Wilhelm August Fröbel (1782 - 1852).
2	Aber auch Erwachsene spielen gerne.	2	Aber auch Erwachsene spielen gerne.
3	Das zeigt jedes Jahr die weltweit größte Messe für Computer- und Videospiele "gamescom" in Köln.	3	Das zeigt jedes Jahr die weltweit größte Messe für Computer- und Videospiele "gamescom" in Köln.
4	Der Markt für Computer- und Videospiele hat sich im ersten Halbjahr 2015 positiv entwickelt:	4	Der Markt für Computer- und Videospiele hat sich im ersten Halbjahr 2015 positiv entwickelt:
5	Mit dem Verkauf von Spielen für PC, Konsole, Handheld sowie Smartphones und Tablet Computer wurden 534 Millionen Euro umgesetzt - im Vergleich zum Vorjahr eine Steigerung um 3 Prozent.	5	Mit dem Verkauf von Spielen für PC, Konsole, Handheld sowie Smartphones und Tablet Computer wurden 534 Millionen Euro umgesetzt - im Vergleich zum Vorjahr eine Steigerung um 3 Prozent.
6	Spielekonsolen wie die Playstation und Xbox und immer neue Varianten von Spielen wie "Fifa" oder "World of Warcraft" haben eine stetig wachsende Fangemeinde.	6	Besonders die Spiele-Apps sind nach Informationen der "gamescom" der wichtigste Treiber des App- und Smartphone-Marktes.
7	Die Entwickler freuen über eine Steigerung von 16 Prozent.	7	Drei Viertel des Umsatzes in den App-Stores von Apple und Google wurden in Deutschland mit Spiele-Apps erzielt.
8	Das gilt auch für die zahlreichen Spiele-Apps. Sie sind nach Informationen der "gamescom" der wichtigste Treiber des App- und Smartphone-Marktes.	8	Aber auch die klassischen Gesellschaftsspiele sind sehr beliebt.
9	Sie sind nach Informationen der "gamescom" der wichtigste Treiber des App- und Smartphone-Marktes.	9	
10	Drei Viertel des Umsatzes in den App-Stores von Apple und Google wurden in Deutschland mit Spiele-Apps erzielt.	10	Die jährlichen Verkaufszahlen bewegen sich zwischen 380 bis knapp 400 Millionen Euro.2014
11	Laut dem Verein für Spieleverlage erreichten die klassischen Familienspiele im Jahr 2014 eine Steigerung von 8,6 Prozent gegenüber dem Vorjahr.	11	feierte es seinen 100. Geburtstag:
12	Die jährlichen Verkaufszahlen bewegen sich zwischen 380 bis knapp 400 Millionen Euro.2014	12	Das bekannteste und vielleicht deutscheste aller Gesellschaftsspiele namens "Mensch ärgere Dich nicht" fehlt in keinem deutschen Haushalt.
13	feierte es seinen 100. Geburtstag:	13	90 Millionen Exemplare sind seit den Anfängen verkauft worden.
14	Das bekannteste und vielleicht deutscheste aller Gesellschaftsspiele namens "Mensch ärgere Dich nicht" fehlt in wohl keinem deutschen Haushalt.	14	Das Spiel ist deshalb so bekannt geworden, weil Herr Schmidt eine geniale Idee hatte:
15	90 Millionen Exemplare sind seit den Anfängen verkauft worden.	15	3.000 Exemplare von "Mensch ärgere Dich nicht" schickte er 1914, als gerade der Erste Weltkrieg tobte, als Sachspende an die Lazarette.
16	Der Münchner Händler Joseph Friedrich Schmidt hat es selber aus Karton gebastelt und gemalt, um seine beiden lebhaften Kinder zu beschäftigen.	16	Hier lagen viele Soldaten verwundet, sie sollten wenigsten etwas Ablenkung und Spaß haben.
17	Das Ziel des Spieles ist, als erster seine vier Spielfiguren ins Ziel zu bringen.		
18	Man darf andere Spieler aus ihrer Position "herausschmeißen", wenn sie im Weg stehen.		
19	Diese müssen dann wieder von vorne beginnen und ärgern sich natürlich maßlos.		
20	Das Spiel ist deshalb so bekannt geworden, weil Herr Schmidt eine geniale Idee hatte:		
21	3.000 Exemplare von "Mensch ärgere Dich nicht" schickte er 1914, als gerade der Erste Weltkrieg tobte, als Sachspende an die Lazarette.		
22	Hier lagen viele Soldaten verwundet, sie sollten wenigstens etwas Ablenkung und Spaß haben.		

Table B.1: Parallel original and simplified documents of the Alumniportal.

B.2 Document Texts for the Simplification Plan of the Austrian Press Agency

The following text in Table B.2 has been provided by the Austrian Press Agency, it contains four news items of a news report published at October 21, 2019. This text is protected by copyright by APA - Austria Presse Agentur eG.

ID	Complex Text (CEFR level B1)	ID	Simple Text (CEFR level A2)
0	Heuer ist der Equal Pay Day am 21. Oktober.	0	Frauen verdienen immer noch weniger Geld als Männer.
1	Wien - Der Equal Pay Day ist heuer in Österreich am 21. Oktober.	1	Wien -
2	Equal Pay Day bedeutet Tag der gleichen Bezahlung.	2	Am 21. Oktober ist in Österreich der Equal Pay Day 2019.
3	Dabei geht es um gleiche Bezahlung für Frauen und Männer.	3	Das ist Englisch und bedeutet auf Deutsch Tag der gleichen Bezahlung.
4	Denn Frauen verdienen in Österreich immer noch rund 20 Prozent weniger Geld als Männer.	4	Damit ist gemeint, dass Frauen und Männer gleich viel Geld verdienen sollen.
5	Im Durchschnitt haben Männer am Equal Pay Day schon so viel Geld verdient wie Frauen im ganzen Jahr.	5	Bisher ist das in Österreich nicht so.
6	Das sagt die Statistik.	6	Frauen verdienen im Durchschnitt rund 20 Prozent weniger Geld als Männer.
7	Statistisch gesehen arbeiten Frauen im Vergleich zu Männern ab dem Equal Pay Day gratis bis zum Jahresende.	7	Bis zum Equal Pay Day haben Männer schon soviel verdient wie Frauen im ganzen Jahr.
8	Natürlich arbeiten Frauen nicht wirklich gratis, sie bekommen ihr Geld bis Ende des Jahres.	8	Das ist so, als ob Frauen ab dem Equal Pay Day gratis arbeiten müssen.
9	Aber sie bekommen um so viel weniger Geld als Männer, als ob sie mehr als 2 Monate lang gratis arbeiten würden.	9	Aber in Wirklichkeit arbeiten Frauen und Männer das ganze Jahr lang.
		10	Nur ist der Gehalts-Unterschied so groß, dass Frauen ein Jahr arbeiten müssen, damit sie so viel verdienen wie Männer bis zum Equal Pay Day am 21. Oktober.
10	Ein Toter bei Brand auf Bauernhof in Niederösterreich.		
11	Artstetten -	11	Ein Mann starb bei einem Brand auf einem Bauernhof.
12	Bei einem Brand auf einem Bauernhof in Niederösterreich ist am Montag ein 54 Jahre alter Mann gestorben.	12	Artstetten -
13	Er und seine Ehefrau wurden bewusstlos aufgefunden.	13	Auf einem Bauernhof in Niederösterreich hat es am Montag gebrannt.
14	Für den Mann kam jede Hilfe zu spät.	14	Bei dem Brand entstanden giftige Gase.
15	Seine Ehefrau wurde wiederbelebt und mit dem Hubschrauber in ein Krankenhaus geflogen.	15	Diese Gase strömten in ein Schlafzimmer, wo ein Ehepaar schlief.
16	Sie könnte eine Rauchgas-Vergiftung haben.	16	Die Frau und der Mann atmeten die giftigen Gase ein und wurden bewusstlos.
17	Auch ihr Mann atmete wahrscheinlich zu viele giftige Rauchgase ein.	17	Der Mann starb dadurch.
18	Der Brand begann in einem Raum mit zerkleinertem Holz.	18	Seine Frau konnte wiederbelebt werden.
19	Dabei entstanden giftige Gase.	19	Sie wurde mit einem Hubschrauber in ein Krankenhaus geflogen.
20	Die giftigen Gase gelangen in das Wohnhaus des Ehepaares.		
21	Erklärung: Rauchgas-Vergiftung.		
22	Wenn es brennt entsteht viel Rauch.		
23	Der Rauch enthält viele giftige Stoffe.		
24	Diese können die Lunge schädigen.		
25	Atmen Menschen zu viel von diesem Rauch ein, bekommen sie eine Rauchgas-Vergiftung.		
26	Eine Rauchgas-Vergiftung ist oft tödlich.		
27	Rekordflug von New York nach Sydney.	20	Der längste Passagier-Flug dauerte über 19 Stunden.
28	Canberra -	21	Canberra
29	Australiens Fluglinie Qantas hat den längsten Passagierflug der Welt absolviert.	22	- Qantas ist die Fluglinie von Australien.
30	Eine fabrikneue Boeing 787 Dreamliner flog nonstop von New York in den USA nach Sydney in Australien.	23	Sie hat jetzt einen Rekord aufgestellt.
31	Nonstop bedeutet direkt, also ohne Zwischenlandung.	24	Ein Flugzeug von der Qantas machte den längsten Passagier-Flug von der Welt.
32	Für die 16.200 Kilometer lange Strecke brauchte das Flugzeug 19 Stunden und 16 Minuten.	25	Das Flugzeug flog von New York in den USA nach Sydney in Australien.
33	Der Rekordflug war ein Testflug mit nur wenigen Passagieren.	26	New York und Sydney sind mehr als 16.000 Kilometer von einander entfernt.
34	Damit wollte die Qantas herausfinden, wie die Menschen an Bord einen so langen Flug vertragen.	27	Das Flugzeug brauchte dafür 19 Stunden und 16 Minuten.
35	Die Qantas will bis Jahresende noch weiter Testflüge durchführen.	28	Der Flug war aber kein normaler Flug, sondern ein Test-Flug.
		29	Im Flugzeug waren viel weniger Menschen als sonst.
		30	Man wollte herausfinden, wie die Menschen so einen langen Flug überstehen.
36	Mehr als 500 neue philippinische Inseln entdeckt.	31	Die Philippinen haben mehr Inseln als gedacht.
37	Manila - Das asiatische Land Philippinen besteht aus sehr vielen Inseln.	32	Manila - Die Philippinen sind ein Land in Asien.
38	Bisher dachte man, es wären 7.107 Inseln.	33	Sie bestehen aus sehr vielen Inseln.
39	Nun hat man herausgefunden, dass es um rund 500 Inseln mehr sind.	34	Bisher dachte man, dass es rund 7.100 Inseln sind.
40	Die Philippinen bestehen damit aus 7.641 Inseln.	35	Jetzt ist man draufgekommen, dass es rund 500 Inseln mehr sind.
41	Die neuen Inseln wurden mit einer neuen Methode entdeckt. Dafür wurde ein verbessertes Radar verwendet.	36	Die Philippinen bestehen also aus rund 7.600 Inseln.
42	Dafür wurde ein verbessertes Radar verwendet.	37	Die 500 Inseln hat man mit einem neuen und viel besseren Radar-Gerät entdeckt.

Table B.2: Parallel original and simplified document of the Austrian Press Agency. The lines separates the news items.

B.3 Document Texts for the Simplification Plan of the Apotheken Umschau

The following text in Table B.3 has been extracted from https://www.apotheken-umschau.de/krankheiten-symptome/gelenks-und-knochenerkrankungen/arthrose-der-hand-und-fingergelenke-733759.html at April 4, 2021. This text is protected by copyright by Wort & Bild Verlag Konradshöhe GmbH & Co. KG.

ID	Complex Document (Standard German)	ID	Simple Document (Plain German Language)
0	Bei einer Arthrose kommt es zum schrittweisen Gelenkverschleiß.	0	Arthrose ist eine Gelenk-Erkrankung.
1	Ausgangspunkt ist ein Defekt im schützenden Knorpel des Gelenks, der dann zu Gelenkschmerzen, Schwellungen, Funktionseinschränkungen und der Zerstörung der Gelenkkontur führen kann.	1	Bei einer Arthrose nutzt sich der Knorpel im Gelenk ab.
2	Eine Arthrose der Finger verläuft manchmal aber auch schmerzfrei ohne ernsthafte Beeinträchtigung.	2	Ein Gelenk besteht aus zwei Knochen.
3	Unterschieden werden zwei Formen: Primäre Arthrose:	3	Die Knochen bewegen sich aneinander entlang.
4	Hier ist die Ursache unbekannt.	4	Zwischen den Knochen sind Knorpel und Gelenkflüssigkeit.
5	Sekundäre Arthrose:	5	Sie dienen dem Schutz der Knochen:
6	Sie entsteht durch Verletzungen oder Krankheiten, beispielsweise Gicht, rheumatoide Arthritis oder Osteoporose (Knochenschwund).	6	So reiben die Knochen nicht aneinander.
7	Eine ständige Überlastung der Gelenke kann eine Arthrose fördern, Bewegungsmangel aber auch..	7	Bei einer Arthrose wird der Knorpel zwischen den Knochen immer dünner.
8	Die Arthrose der Fingergelenke (Fingerpolyarthrose) kommt bei Frauen in und nach den Wechseljahren bis zu zehnmal häufiger vor als bei Männern.	8	Deshalb reiben die Knochen aneinander.
9	Möglicherweise spielen Veränderungen im Hormonhaushalt eine Rolle.	9	Das betroffene Gelenk nutzt sich ab.
10	Auch die Erbanlagen scheinen einen Einfluss auf das Erkrankungsrisiko zu haben.	10	Diese Abnutzung des Gelenks kann zu Schmerzen führen.
11	Sind nahe Verwandte wie Mutter oder Großmutter betroffen, erhöht sich das eigene Risiko.	11	Oft lässt sich das Gelenk dann nicht mehr so gut bewegen.
12	Nach dem 50. Lebensjahr nimmt der Verschleiß bei großen und kleinen Gelenke generell zu.	12	Bei einer fortgeschrittenen Arthrose ist der Knorpel stark abgerieben und sehr dünn.
13	Trotzdem handelt es sich nicht um eine unausweichliche "Alterserscheinung".	13	An manchen Stellen zwischen den beiden Knochen gibt es gar keinen Knorpel mehr.
14	Polyarthrose der Finger: Fingerendgelenke (= häufigere Heberden-Arthrose), Fingermittelgelenke (= seltenere Bouchard-Arthrose), Daumensattelgelenk (= Rhizarthrose).	14	Dort reiben die Knochen aufeinander.
15	Arthrose des Hangelenks: im körpernahen Handgelenk zwischen Speiche und Kahnbein / Mondbein, oder: im Gelenk zwischen Speiche und Elle.	15	Arthrose kann in den Fingern vorkommen: in den Endgelenken, in den Mittelgelenken, im Daumensattelgelenk.
16	Eine Arthrose entwickelt sich allmählich.	16	Arthrose kann aber auch im Handgelenk vorkommen: zwischen Speiche und Kahnbein, zwischen Speiche und Mondbein, zwischen Speiche und Elle.
17	Oft macht sie anfangs keine Symptome und bleibt lange unbemerkt (= stumme Arthrose).	17	Die Ursachen einer Arthrose sind nicht immer bekannt.
18	Ob und wann es zu Problemen kommt, ist individuell verschieden.	18	Mögliche Ursachen können sein: falsche Belastungen eines Gelenks an der Hand oder am Finger, Verletzungen an dem betroffenen Gelenk, andere Krankheiten wie Gicht oder Rheuma.
19	Das Ausmaß der Gelenkveränderungen – die zum Beispiel auf dem Röntgenbild sichtbar sind – lässt nicht immer Rückschlüsse auf das Ausmaß der Beschwerden zu.	19	Frauen haben häufiger Arthrose in den Fingergelenken als Männer.
20	Geringe Veränderungen können starke Schmerzen verursachen und umgekehrt.	20	Diese kommt vor allem während und nach den Wechseljahren.
21	Die Arthrose der Fingergelenke beginnt meistens schleichend.	21	Das liegt vermutlich an den Hormonen.
22	Vor allem morgens fühlen sich die Finger steif an, neigen zu Schwellungen.	22	Die Hormone verändern sich nämlich während der Wechseljahre.
23	Eine Faust zu bilden, fällt schwer.	23	Haben die eigene Mutter oder Großmutter Arthrose in den Fingergelenken?
24	Allmählich schmerzen die Finger auch bei Bewegungen, später in Ruhe.	24	Dann ist das Risiko für Arthrose erhöht.
25	Die Beweglichkeit der Finger nimmt ab.	25	Eine Arthrose kann verschiedene Anzeichen haben:
26	Phasenweise können die Gelenke anschwellen, gerötet und überwärmt sein (= aktivierte Arthrose).	26	Tun bestimmte Bewegungen mit den Fingern oder dem Handgelenk weh?, Sind die betroffenen Gelenke manchmal angeschwollen?
27	Sind die Fingerendgelenke betroffen, bilden sich eventuell zystische Verdickungen, sogenannte Mukoid-Zysten.	27	Können die betroffenen Gelenke nicht richtig bewegt werden?
28	Aus ihnen kann sich gallertartige Flüssigkeit entleeren.	28	Haben andere Personen in der Familie Arthrose?
29	Im späteren Stadium können knöcherne Verdickungen rechts und links der Gelenke entstehen und Achsabweichungen auftreten.	29	Waren oder sind die betroffenen Gelenke starken Belastungen ausgesetzt?
30	Eine Arthrose am Daumensattelgelenk (Sattelgelenksarthrose) verursacht meistens Schmerzen bei zahlreichen Alltagsarbeiten – zum Beispiel dem Öffnen von Flaschen, dem Drehen von Schraubverschlüssen, dem Heben schwerer Töpfe.	30	Wurden die betroffenen Gelenke schon einmal verletzt?
31	Denn das Gelenk ist besonders beweglich und bei sämtlichen Bewegungsmustern beteiligt.	31	Eines oder mehrere dieser Anzeichen treffen zu?
32	Eine Arthrose des Handgelenkes ist oft die Folge von Knochenbrüchen oder tritt beim klassischen Rheuma auf.	32	Dann kann eine Arthrose die Ursache sein.
33	Im fortgeschrittenen Stadium kommt es zu Schwellung, Schmerzen und Bewegungseinschränkungen des Handgelenks beim Beugen, Strecken und bei Umwendbewegungen.	33	Manchmal bemerken Patienten mögliche Anzeichen einer Arthrose erst spät.
34	Wie Scharniere verbinden Gelenke unsere Knochen miteinander und ermöglichen Bewegungen.	34	Die Anzeichen sind nämlich oft bei jeder Person unterschiedlich.
35	Eine Kapsel aus Bindegewebe umschließt die beiden Knochenenden.	35	Sie glauben: Ich habe vielleicht Arthrose?
36	Die Innenseite der Gelenkkapsel ist ausgekleidet mit der Gelenkinnenhaut.		
37	Sie produziert Gelenkflüssigkeit (Synovia), die in den schmalen Spalt zwischen den beiden Knochenenden hineinfließt.		
38	Die Gelenkflüssigkeit "schmiert" die Bewegung – so ähnlich wie ein paar Tropfen Öl ein mechanisches Scharnier beweglich halten.		

Table B.3: Parallel original and simplified document of the Apotheken Umschau. Part I (continued on next page).

ID	Complex Document (Standard German)	ID	Simple Document (Plain German Language)
39	Die Gelenkflüssigkeit transportiert außerdem Nährstoffe zu der Knorpelschicht, die die beiden Knochenenden wie eine Schutzschicht überzieht.		
40	Der Gelenkknorpel besitzt keine eigenen Blutgefäße, die ihn ernähren könnten.		
41	Er ist deshalb auf die Nährstoffversorgung über die Gelenkflüssigkeit angewiesen.		
42	Auch seine Stoffwechsel-Abbauprodukte entsorgt der Knorpel auf diesem Weg.		
43	Der Transportmechanismus funktioniert jedoch nur dann perfekt, wenn das Gelenk regelmäßig bewegt wird – ohne dass es dabei zur Überlastung kommt.		
44	Der Knorpel vermindert die Reibung im Gelenk und verteilt den Druck gleichmäßig auf den Knochenenden.		
45	Kommt es zu Schäden im Knorpel, raut er auf, wird rissig und dünner.		
46	Er kann seine Funktion als Schutzschicht nicht mehr richtig erfüllen, Stöße und Druck nicht mehr gleichmäßig auf den ganzen Knochen verteilen.		
47	An manchen Stellen müssen Knorpel und darunter liegender Knochen nun extreme Belastungen aushalten.		
48	Dieser Zustand verursacht zunächst noch keine Schmerzen.		
49	Der benachbarte Knochen reagiert auf die ungünstige neue Situation, indem er stellenweise dichter und massiver wird.		
50	An den Rändern bilden sich kleine Knochenanbauten, die den übermäßigen Druck aufnehmen sollen.		
51	Meistens funktioniert diese "Hilfskonstruktion" nicht mehr so gut wie das ursprünglich geformte Gelenk.		
52	Es kommt zu Abrieb und damit zur Reizung der Gelenkinnenhaut.		
53	Sie produziert mehr Gewebeflüssigkeit als im Normalfall, ein Gelenkerguss kann entstehen.		
54	Auch die Zusammensetzung der Flüssigkeit ändert sich:		
55	Sie enthält Entzündungsstoffe und Abwehrzellen.		
56	Eine Entzündung und Schmerzen können die Folge sein.		
57	Das Gelenk fühlt sich geschwollen an, ist warm und rot.		
58	Oft bessert sich die Entzündung nach einiger Zeit wieder.		
59	Meistens folgen auf schmerzarme Intervalle aber neue Schmerzepisoden.		
60	Die wiederholten Entzündungsschübe schädigen den Knorpel weiter.		
61	Betroffene schonen die kranken Gelenke gezwungenermaßen – was wiederum zur Folge hat, dass die Gelenkflüssigkeit den Knorpel schlechter mit Nährstoffen beliefert.		
62	Der Knorpel wird zusätzlich geschwächt.		
63	Die Knorpelschicht kann stellenweise sogar komplett abgerieben werden, so dass der Knochen völlig ungeschützt frei liegt.		
64	Langfristig werden unter Umständen weitere Gelenkstrukturen wie Bänder und Sehnen in Mitleidenschaft gezogen, es kann zu Fehlstellungen kommen.		
65	Die Gelenkbeweglichkeit nimmt ab.		
66	Der Arzt erkundigt sich, welche Beschwerden auftreten, ob die Finger zum Beispiel bei bestimmten Bewegungen schmerzen oder ob die Gelenke manchmal angeschwollen, gerötet und überwärmt sind.		
67	Außerdem ist von Interesse, ob Verwandte ebenfalls an Arthrose leiden – ein möglicher Hinweis auf eine familiäre Veranlagung zur Krankheit.		
68	Der Mediziner wird in der Regel fragen, ob die Finger besonderen Belastungen ausgesetzt waren oder sind, beispielsweise im Beruf, oder ob die Finger oder das Handgelenk in der Vergangenheit bei Unfällen verletzt wurden.		
69	Dann wird der Arzt die Gelenke genau untersuchen und überprüfen, ob ihre Beweglichkeit eingeschränkt ist.		
70	Meistens sind Röntgenaufnahmen der Hände erforderlich.		
71	Darauf können Arthrose-typische Gelenkveränderungen erkennbar sein.		
72	Der Gelenkspalt ist oft verschmälert, der gelenknahe Knochen verdichtet.		
73	An den Gelenkrändern finden sich nicht selten knöcherne Anbauten (Osteophyten).		
74	Gelegentlich kommen zusätzliche bildgebende Verfahren wie Computertomografie (CT) oder Magnet-Resonanz-Tomografie (MRT) mit Kontrastmittel zum Einsatz.		
75	Durch Blutuntersuchungen oder eine Untersuchung von Gelenkflüssigkeit lassen sich Stoffwechselerkrankungen oder Kristallopathien wie Gicht als Ursachen ausschließen.		
76	Die Ziele der Behandlung lauten: Schmerzen lindern, die Beweglichkeit verbessern, den Gelenkverschleiß bremsen.		
77	Die Ursachen einer Hand- und Fingergelenksarthrose sind in der Regel nicht bekannt und können deshalb auch nicht beseitigt werden.		
78	Die Symptome lassen sich jedoch bekämpfen.		
79	Außerdem ist es ratsam, alles zu meiden, was den Gelenken zusätzlich schadet.		
80	Treten erste Anzeichen der Arthrose auf, rät der Arzt dazu, Überanstrengungen und Fehlbelastungen im Alltagsleben, Beruf und Sport zu reduzieren.		
81	Mit Hilfe von Ergotherapeuten lernen Patienten, welche Handgriffe im Alltag besonders "auf die Gelenke gehen" – und welche Tricks die Finger entlasten.		
82	Es gibt etliche einfache Hilfsmittel, zum Beispiel Griffverstärkungen von Stiften oder Besteck.		
83	Sie helfen, die Fingergelenke zu schonen.		
84	Die Anlage einer Schiene kann bei einer Arthrose des Daumensattelgelenks Schmerzen reduzieren.		
85	Welche Form der Ernährung bei einer Fingerpolyarthrose besonders günstig ist, kann nicht abschließend beurteilt werden.		
86	Generell raten Experten bei Gelenkproblemen, lieber nicht zu fleischreich zu essen, sich ausgewogen zu ernähren und viel Obst und Gemüse auf den Speiseplan zu nehmen.		
87	Ein positiver Effekt von knorpelschützenden Substanzen (zum Bespiel Glucosamin) wird diskutiert.		
88	Bei akuten Schmerzen können schmerzlindernde und entzündungshemmende Medikamente sinnvoll sein, zum Beispiel nicht steroidale Antirheumatika (NSAR).		
89	Welche Präparate geeignet sind und wie sie angewendet werden, sollte mit dem Arzt besprochen werden.		
90	Die Verabreichung knorpelprotektiver Medikamente direkt in die betroffene Gelenke ist nur in frühen Arthrosestadien sinnvoll.		
91	In fortgeschrittenen Fällen kann durch Kortisoninjektionen manchmal der akute Entzündungsschub gelindert werden.		
92	Eine Operation sollte nur dann in Erwägung gezogen werden, wenn andere Therapien nicht mehr ausreichend helfen.		
93	Zu möglichen Risiken und Erfolgschancen der Eingriffe lassen sich Patienten am besten ausführlich vom Arzt beraten.		
94	Welche Behandlung am besten geeignet ist, richtet sich unter anderem danach, welche Gelenke betroffen sind:		

Table B.4: Parallel original and simplified document of the Apotheken Umschau. Part II (continued from previous page).

Appendix C

German Evaluation Aspects & Statements

Aspect	Description
Grammatikalität	Der vereinfachte Satz klingt flüssig, er enthält keine Grammatikfehler.
Grammatikalität (Ausgangssatz)	Der Ausgangssatz klingt flüssig, er enthält keine Grammatikfehler.
Sinnerhaltung	Der vereinfachte Satz gibt auf angemessene Weise die Bedeutung des Ausgangssatzes wieder. Weniger wichtige Informationen werden dabei möglicherweise weggelassen.
Informationsgewinn	In dem vereinfachten Satz sind neue Informationen enthalten oder deutlicher ausgedrückt. Diese Informationen sind in dem Ausgangssatz nicht ausdrücklich enthalten.
Einfachheit (Allgemein)	Der vereinfachte Satz ist leichter zu verstehen als der Ausgangssatz.
Einfachheit (Satzstruktur)	Der Satzbau und die Satzstruktur im vereinfachten Satz sind leichter zu verstehen als der Satzbau und die Satzstruktur im Ausgangssatz.
Einfachheit (Wortwahl)	Die Wörter im vereinfachten Satz sind leichter zu verstehen als die Wörter im Ausgangssatz.
Einfachheit (Leichter Satz)	Der vereinfachte Satz ist leicht zu verstehen.
Einfachheit (Ausgangssatz)	Der Ausgangssatz ist leicht zu verstehen.
Kontextunabhängigkeit (Leichter Satz)	Der vereinfachte Satz kann verstanden werden ohne den ganzen Absatz zu lesen.
Kontextunabhängigkeit (Ausgangssatz)	Der Ausgangssatz kann verstanden werden ohne den ganzen Absatz zu lesen.
Mehrdeutigkeit (Leichter Satz)	Der vereinfachte Satz ist mehrdeutig. Er kann auf verschiedene Weisen verstanden werden.
Mehrdeutigkeit (Ausgangssatz)	Der Ausgangssatz ist mehrdeutig. Er kann auf verschiedene Weisen verstanden werden.

Table C.1: German evaluation criteria and their corresponding statements.

EASY – PLAIN – ACCESSIBLE

Vol. 1 Isabel Rink: Rechtskommunikation und Barrierefreiheit. Zur Übersetzung juristischer Informations- und Interaktionstexte in Leichte Sprache. 472 pages. ISBN 978-3-7329-0593-5

Vol. 2 Silvia Hansen-Schirra/Christiane Maaß (eds.): Easy Language Research: Text and User Perspectives. 288 pages. ISBN 978-3-7329-0688-8

Vol. 3 Christiane Maaß: Easy Language – Plain Language – Easy Language Plus. Balancing Comprehensibility and Acceptability. 304 pages. ISBN 978-3-7329-0691-8

Vol. 4 Elisa Perego: Accessible Communication: A Cross-country Journey. 200 pages. ISBN 978-3-7329-0654-3

Vol. 5 Silke Gutermuth: Leichte Sprache für alle? Eine zielgruppen-orientierte Rezeptionsstudie zu Leichter und Einfacher Sprache. 312 pages. ISBN 978-3-7329-0587-4

Vol. 6 Anne-Kathrin Gros/Silke Gutermuth/Katharina Oster (Hg.): Leichte Sprache – Empirische und multimodale Perspektiven. 160 pages. ISBN 978-3-7329-0708-3

Vol. 7 Maher Tyfour: Sprachmacht auf engstem Raum: Die Inszenierung der Stadt in den Hörfilmen der Münchner Tatort-Filmserie. Eine korpusgeleitete Studie zur Audiodeskription. 246 pages. ISBN 978-3-7329-0699-4

Vol. 8 Camilla Lindholm/Ulla Vanhatalo (eds.): Handbook of Easy Languages in Europe. 660 pages. ISBN 978-3-7329-0771-7

Vol. 9 Silvia Hansen-Schirra/Katja Abels/Sarah Signer/Christiane Maaß: The Dictionary of Accessible Communication. 212 pages. ISBN 978-3-7329-0729-8

Vol. 10 Katrin Lang: Auffindbarkeit, Wahrnehmbarkeit, Akzeptabilität. Webseiten von Behörden in Leichter Sprache vor dem Hintergrund der rechtlichen Lage. 488 pages ISBN 978-3-7329-0804-2

Frank & Timme

EASY – PLAIN – ACCESSIBLE

Vol. 11 Silvana Deilen: Optische Gliederung von Komposita in Leichter Sprache. Blickbewegungsstudien zum Einfluss visueller, morphologischer und semantischer Faktoren auf die Verarbeitung deutscher Substantivkomposita. 782 pages. ISBN 978-3-7329-0834-9

Vol. 12 Elena Husel: Leichte Sprache in der Bundesverwaltung. Was? Wer? Wie? 250 pages. ISBN 978-3-7329-0849-3

Vol. 13 Sarah Ahrens/Rebecca Schulz/Janina Kröger/Sergio Hernández Garrido/Loraine Keller/Isabel Rink (eds.): Accessibility – Health Literacy – Health Information. Interdisciplinary Approaches to an Emerging Field of Communication. 234 pages. ISBN 978-3-7329-0895-0

Vol. 14 Silvana Deilen/Silvia Hansen-Schirra/Sergio Hernández Garrido/Christiane Maaß/Anke Tardel (eds.): Emerging Fields in Easy Language and Accessible Communication Research. 484 pages. ISBN 978-3-7329-0922-3

Vol. 15 Christiane Maaß/Isabel Rink (eds.): Handbook of Accessible Communication. 750 pages. ISBN 978-3-7329-0840-0

Vol. 16 Giulia Pedrini: Medical communication between *Plain Language* and *Einfache Sprache*. A corpus analysis of layperson summaries of clinical trials in English, German, and Italian. 528 pages. ISBN 978-3-7329-1085-4

Vol. 17 Sarah Ahrens: Einfache Sprache in der Gesundheitskommunikation. Patientinnenaufklärung für Frauen mit Deutsch als Zweitsprache. 342 pages. ISBN 978-3-7329-1132-5

Vol. 18 Anke Radinger: Researching Subtitling Processes. Methodological considerations for the investigation of AI-assisted subtitling workflows. 506 pages. ISBN 978-3-7329-1029-8

Vol. 19 Laura Marie Maaß: Erwartungen, Einstellungen, Erfahrungen. Zur Interaktion zwischen hörenden Gebärdensprachdolmetschenden und ihrer tauben Kundschaft. 642 pages. ISBN 978-3-7329-1161-5

Frank & Timme

EASY – PLAIN – ACCESSIBLE

Vol. 20 Silvia Hansen-Schirra/Chris Maaß (Hg.): Text- und nutzerseitige Studien zu Leichter Sprache. 334 pages. ISBN 978-3-7329-0701-4

Vol. 21 Regina Stodden: Automatic German Text Simplification: Data, Evaluation, and Models. 292 pages. ISBN 978-3-7329-1216-2

Frank & Timme

Texts written in simplified language are essential for accessible communication. However, manually writing accessible texts or simplifying difficult-to-read texts into accessible language is very time-consuming. Machine learning models such as large language models (LLMs) can assist with this intra-lingual translation task by providing drafts for professional translators.
This volume takes a computational linguist's perspective on automatic text simplification (ATS), providing a broad introduction to the field. The book begins by introducing the challenges and opportunities. It also explains how to build simplification datasets, use them to train machine learning models, and evaluate them. The book also provides state-of-the-art overviews of German text simplification models, datasets and evaluation metrics for document and sentence simplification. While the focus is primarily on German plain language ("Einfache Sprache"), the book also provides some insights into German easy-to-read language ("Leichte Sprache") and other languages.

Regina Stodden is a computational linguist who obtained her PhD at Heinrich Heine University Düsseldorf. During her PhD, she mainly worked on creating open-source datasets, evaluation frameworks, and models for automatic text simplification in German. Later, her research interests have shifted within the area of natural language processing from automatic text simplification and language proficiency assessment to the broader area of NLP for social good, including practical solutions for non-research partners.

ISBN 978-3-7329-1216-2

www.frank-timme.de