Towards AI-supported Health Communication in Plain Language: Evaluating Intralingual Machine Translation of Medical Texts

Silvana Deilen[1], Ekaterina Lapshinova-Koltunski[1], Sergio Hernández Garrido[1], Christiane Maaß[1], Julian Hörner[2], Vanessa Theel[3], Sophie Ziemer[3]
[1]University of Hildesheim, [2]Text und Bild Verlag, [3] SUMM AI
deilen,lapshinovakoltun,hernandezs,maass@uni-hildesheim.de,
j.hoerner@wubv.de, sophie,vanessa@summ-ai.com

Abstract

In this paper, we describe results of a study on evaluation of intralingual machine translation. The study focuses on machine translations of medical texts into Plain German. The automatically simplified texts were compared with manually simplified texts (i.e., simplified by human experts) as well as with the underlying, unsimplified source texts. We analyse the quality of the translations based on different criteria, such as correctness, readability, and syntactic complexity. The study revealed that the machine translations were easier to read than the source texts, but contained a higher number of complex syntactic relations than the human translations. Furthermore, we identified various types of mistakes. These included not only grammatical mistakes but also content-related mistakes that resulted, for example, from mistranslations of grammatical structures, ambiguous words or numbers, omissions of relevant prefixes or negation, and incorrect explanations of technical terms.

Keywords: plain language, medical discourse, accessible health communication, health literacy, machine translation

1. Introduction

In interlingual translation, CAT (computer-aided translation) tools and machine translation systems such as DeepL or Google Translate have significantly changed the translation industry and have become an indispensable component in the translation process, as can be seen in the case of the European Commission: While until a few years ago, all legal texts and official documents were translated by human translators alone, today, the European Commission makes no secret of the fact that their translators now tend to revise and post-edit the texts rather than translate them themselves.

However, it seems that this ”transition to a new era” (Canfora and Ottmann, 2020) has not yet reached the field of intralingual translation. Following Jacobson, intralingual translation is defined as ”an interpretation of verbal signs by means of other signs of the same language” (Jakobson, 1959, p. 233). In this context, it refers to translating a text from standard language into a complexity-reduced language variety of the same language as described in Maaß (2020, p. 171ff) and in Maaß (2024, p. 265ff). Plain Language translation is also related to text simplification, which is an automatic procedure of changing complex structures into simple ones. However, from the perspective of translation studies and translation practice, this is a type of translation that involves more than reducing surface complexity. Unlike in interlingual translation (i.e. translating a text from one language to another), in intralingual translation the use of CAT tools and machine translation systems is still not established (Maaß et al. 2014, Deilen et al. 2023).

However, especially in health communication, there is a high need for technological assistance, which is especially due to the population's alarmingly low health literacy, as Schaeffer et al. (2017) point out. Their findings lead to the National Action Plan of the German Federal Government to promote health literacy (*Nationaler Aktionsplan Gesundheitskompetenz*, Schaeffer et al., 2018a) that lists Plain Language among the instruments to secure better access to information as the basis for better health literacy (see Section 2.2 below).

In our study, we analyse machine translations of medical texts into Plain German. The texts were taken from the website of the German health magazine *Apotheken Umschau*, which publishes healthcare articles and health information both in standard German and in Plain German. We evaluate the machine-translated output comparing it with human translations from the magazine's website, as well as with the underlying sources. We present the results of the qualitative and quantitative analysis.

2. Related Work

2.1. Plain German

Both Easy Language and Plain Language are complexity-reduced language varieties which aim to improve readability and comprehensibility of

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 44–53

texts (Bredel and Maaß, 2016; Maaß, 2020). They are used in different communication scenarios, e.g. in legal communication (Maaß and Rink, 2021) or health communication (see the contributions in Ahrens et al., 2022), and have different target groups (Maaß and Schwengber, 2022). While Easy Language is characterized by a maximally reduced complexity on all language levels and is mainly intended for people with communication impairments and disabilities, the grammatical and lexical features of Plain Language are only slightly less complex than in standard language and are mainly a means to open expert contexts for lay people (Maaß, 2020). Therefore, the main target audience of Plain Language is lay people with average or slightly below average language or reading skills (Maaß, 2020). In Germany, Easy Language has become a subject of scientific research since 2014 with rapidly growing output of publications in the following years. The studies point in two basic directions: studies on text qualities and possible barriers in various forms of communication on the one side (see, for example, Rink 2019) and studies on comprehensibility and recall by different target groups on the other (see, for example, Gutermuth 2020, Deilen 2021).

Unlike Easy Language, Plain Language is a dynamic variety. Plain Language does not have a fixed set of rules, but the linguistic complexity of Plain Language texts is adapted to the needs of the intended audience in a specific target situation (Bredel and Maaß 2016, Maaß 2020). Therefore, Plain Language is a flexible concept that varies depending on the presumed reading skills of its target group (for a more detailed distinction between the two varieties see Maaß 2020). In comparison to Easy Language, Plain Language has the advantage of not stigmatising the target audience (Maaß, 2020), which is one of the reasons why it is is also more acceptable than Easy Language. However, due to the higher degree of linguistic complexity, Plain Language texts are far less comprehensible than Easy Language texts and therefore not necessarily accessible for people with very low literacy skills (Maaß, 2020). Maaß (2020) therefore models the variety Easy Language Plus, which is situated between Easy Language and Plain Language and strikes a balance between comprehensibility and acceptability.

In Germany, Plain Language is used in different fields and different settings, such as by *Deutschlandfunk*, a public-broadcasting radio station that publishes weekly news in Plain Language for a broad audience with reading difficulties or reduced language skills. However, one of the most prominent application area of German Plain Language is health communication (Ahrens et al., 2022).

2.2. Accessibility in Medical Domain in Germany

In 2016, findings from the Health Literacy Survey (HLS-GER) revealed that over half of the German population (54,3%) experiences significant difficulties in locating, comprehending, evaluating and effectively using health-related information (Schaeffer et al., 2017). These results, which were ”significantly worse than expected” (Schaeffer et al., 2020, p. 2), led to an increased awareness of the need for accessible health information and resulted in the development of the National Action Plan Health Literacy, which was published in 2018 (Schaeffer et al., 2018b). According to the National Action Plan, one strategy to promote health literacy in Germany is the use of Plain Language, which ”aims to adapt complex texts to the literacy skills of large population groups” (Schaeffer et al., 2018b, p. 43); the National Action Plan cites the model put forward in Bredel and Maaß (2016) for reference. Considering new data from the second Health Literacy Survey (HLS-GER 2) in 2021, Plain Language in German health communication becomes even more relevant, for even more persons (58,8 %) experience difficulties navigating the health system (Schaeffer et al., 2021). One of the most prominent examples of implementing this strategy is the *Apotheken Umschau*[1]. The *Apotheken Umschau*, which is Germany's leading health publisher and the largest consumer medium in the German-speaking area with a traffic of 6.94 m. visits and 64.42 m. page impressions per month[2], has so far published more than 220 texts in Plain Language on their website in a co-operation with the Research Centre for Easy Language (University of Hildesheim)[3]. By publishing information in both standard German and Plain German, they aim to “make reliable and helpful information on diseases, medications and preventive health care accessible to everyone with as few barriers as possible” (Hörner, 2022, p. 77). The project is based on the linguistic model for Plain Language by Bredel and Maaß (2016) and Easy Language Plus by Maaß (2020).

2.3. NLP for Plain Languages

Although the potentials of using computer-aided translation (CAT) tools for Plain Language translation were discussed almost a decade ago (Maaß et al., 2014), the role of automation and CAT tools in this area is still a major research desideratum. These potentials were re-explored and extended by Hansen-Schirra et al. (2020). In gen-

[1] `https://www.apotheken-umschau.de`

[2] `https://ausweisung-digital.ivw.de`, retrieved 10.10.2023

[3] `https://www.uni-hildesheim.de/leichtesprache`

eral, intralingual translation poses a number of challenges for CAT tools: terminology management and sentence alignment (see e.g. Kopp et al. 2023) differ from those common in interlingual translation and, therefore, pose additional workload for translators instead of decreasing it. The theoretical set-up for a CAT tool for intralingual translation was suggested by Welch and Sauberer (2019). However, to our knowledge, such tools, as well as their analysis, are still missing.

While there are plenty of studies on automatic text simplification methods that aim to automatically convert a text into another text that is easier to understand but ideally conveys the same message as the source text which contributes to textual accessibility (Sheang and Saggion, 2021; Maddela et al., 2021; Martin et al., 2020; Saggion, 2017), most of them do not consider the needs of the target audience. Scarton and Specia (2018) showed that using target audience oriented data helps to build better models for automatic text simplification using the Newsela corpus[4]. However, this corpus contains news texts only, whereas we are looking into the medical discourse, where texts in Plain Language enable accessibility to health literacy.

Biomedical lay summarization is also related to automatic translation into Plain Language. Goldsack et al. (2023) present results of a shared task on lay summarization of biomedical research articles (BioLaySumm 2023). In this case, medical information in expert language (expert-to-expert communication) is summarized for non-experts (expert-lay-communication). However, it is important to state that Plain Language translation, even if translators select and add information as described in Bredel and Maaß (2016, p. 202 ff.), is not the same textual practice as text summarization.

Specific problems of automatic systems of intralingual translation, e.g. copying source segments into the output, were addressed by Säuberli et al. (2020) and Spring et al. (2023) who showed that pretrained and fine-tuned NMT models have promising results in automatic text simplification. However, as stated by Anschütz et al. (2023), even though there are improvements in the systems of automated intralingual translation, the outputs might not be used by the target groups directly. Nevertheless, they may serve as a draft for professional intralingual translators to reduce their workload.

Deilen et al. (2023) drew similar conclusions for the outputs produced with ChatGPT. The authors investigated the feasibility of using this tool for intralingual translation. They analysed the quality of the generated texts according to such criteria as correctness, readability, and syntactic complexity. Their results indicated that the generated texts were easier than the standard texts, but the content was not always rendered correctly. Besides that, the automated intralingual output did not fully meet the standards which human translators follow.

In the present study, we follow a similar approach. However, while the authors analysed intralingual translation into German Easy Language, a simplified, controlled language variety adapted to the needs of people with reading impairments, we focus on translation into Plain German. Besides that, we focus on medical texts, whereas the authors translated citizen-oriented administrative texts. Moreover, we investigate the feasibility of a tool which was specifically trained for intralingual translation into Easy and Plain Language instead of using a chatbot designed for various tasks.

3. Research Design

3.1. Data Collection

We selected thirty texts from the website of the German health magazine *Apotheken Umschau*. The texts cover a broad range of topics such as insect bite, vaccination, cistitis, lumbago, food poisoning, heel spur and others. For all texts in the sample, a translation in Plain Language was already available, which was done by human translators. Both the source texts and the human translations were reviewed by medical or pharmaceutical professionals from the editorial team of *Apotheken Umschau* and comply with the guidelines of evidence-based medicine. Content accuracy is therefore guaranteed for the sample. This sample of thirty texts was translated using the machine translation system SUMM AI[5].

Then, we analysed machine-translated texts comparing them with human translations, as well as with the source texts following Deilen et al. (2023). For this, we used three different criteria, namely the correctness of the content (see 3.2.1), the readability of the texts (see 3.2.2), and their syntactic complexity (see 3.2.3). The first criterion was applied to the machine translations only, the second and the third criteria were applied on all the three subcorpora (source texts, human translations, and machine translations)[6].

[4] `https://newsela.com/data`

[5] SUMM AI (`https://summ-ai.com/en/`) is a tool for translating texts into Easy German and Plain German. The company SUMM AI offers different licenses for freelancers, authorities and companies.

[6] The whole dataset we analysed is published on GitHub, i.e., the selected texts (sources, human and machine translations), including the raw data, the parsed data (conllu) and the Textlab analyses per text, and can be accessed under `https://github.com/katjakaterina/MT4plainDE`.

3.2. Data Analysis

3.2.1. Correctness

The content of the machine-generated texts was first analysed for correctness. This content evaluation was done manually, whereby each text was assessed independently by two researchers, who checked whether the medical information in the target text is still valid despite reduction of complexity and shortening of information. In cases where an accurate assessment required specialized knowledge, a healthcare professional from the *Apotheken Umschau* team was consulted. No quantitative error analysis was performed. Consequently, a translation was already considered incorrect if it contained one content-related error. This is because the study seeks insights into who artificial intelligence (AI) powered translation tools are suitable for: For translators, content providers, or end users? In order for machine translation into Easy or Plain Language to be safely usable by end users, the target texts must not contain errors. The presence of errors in the target texts therefore indicates usability for users other than the end users.

3.2.2. Readability

We also compared the comprehensibility of the human and machine translations, as well as of the source texts. For this, we use the Hohenheim Comprehensibility Index (HIX). The HIX is a meta index that calculates the readability of a text taking into account the four major readability formulas common in Easy Language Research (Bredel and Maaß, 2016, p. 61ff). They include the Amstad index, the simple measure of gobbledygook (G-SMOG) index, the Vienna non-fictional text formula (W-STX) and the readability index (LIX), with a HIX of 0 indicating extremely low comprehensibility and a HIX of 20 extremely high comprehensibility (for further details see: `https://klartext.uni-hohenheim.de/hix`). The benchmark for a text to be classified as a text in Easy German, which is the least complex variety of German, is set at 18 points (Rink 2019, p. 77). As Plain German is more complex than Easy German, we suggest setting the benchmark for Plain German at 16 points.

3.2.3. Syntactic Complexity

We operationalised syntactic complexity as a distribution of specific syntactic relations, i.e. specific clauses. We automatically identified syntactic relations using dependency parsing that we obtained with the Stanford NLP Python Library Stanza (v1.2.1)[7] with all the models pre-trained on the Universal Dependencies v2.5 datasets. Our list of selected structural categories include adnominal clauses or clausal modifiers of noun (acl), adverbial clause modifiers (advcl), clausal complement (ccomp), clausal subjects (csubj), open clausal elements (xcomp) and parataxis relation (parataxis). These selected categories are all listed under the clause dependents[8] in the Universal Dependency. More details on dependency relations and their definitions across languages can be found in (De Marneffe et al., 2021). We collected and compared the distribution frequencies of these categories in the three subcorpora under analysis (source texts, human translations, and machine translations). We interpreted the results based on the assumption that the higher the number of these dependency relations in the corpus, the more complex the texts contained in these subcorpora are.

3.2.4. Automatic Evaluation Measures

We also used other indices that are commonly used in the field of automatic text simplification. Specifically, we applied SARI (Xu et al., 2016), which is a quantitative measure to evaluate automatic text simplification systems. SARI is suitable for evaluation of automatic text simplification models and could so be also suitable for the task of evaluating intralingual machine translation. In order to be able to compute these metrics, we aligned the source texts, machine translations and human translations on a paragraph level and scored them with respect to their grade of alignment. Out of the 935 analysed paragraphs more than 70%, namely 676 paragraphs, had no alignment between source text and human translations. This means that for 547 paragraphs in the source text no matching simplification could be allocated in the human translation and for 123 paragraphs in the human translation no matching source paragraph could be identified.

3.3. Results

3.3.1. Correctness

The analysis of the correctness of the machine translations showed that only one of the 30 texts was correctly translated. The other 29 texts showed problems with regard to their correctness in different aspects. Overall, the results are disparate and inconsistent. The texts do not follow a uniform structure and are not action-oriented. In practice, they would have to be completely post-edited. In some cases, the source texts (ST) are more stringent and comprehensible than the target texts (TT). We encounter grammatical errors and misspellings, omissions of relevant prefixes or negation, incorrect explanations of technical terms, incomplete listings, contradictory state-

[7] `https://stanfordnlp.github.io/stanza/index.html`

[8] `https://universaldependencies.org/u/dep/`

ments etc. It should be emphasized once again that no quantitative evaluation was performed because the mere presence of the errors themselves was considered a risk for the primary users. Furthermore, so far we have not classified or ranked the error types based on severity levels, but we plan to do so in our future work (see 4).
Some examples of the errors we found are given in the following.
Grammatical errors and misspellings are for example:

- homophonic but not homographic words are not correctly selected: ”dass” (the connective ”that”) vs. ”das” (the article ”the” or the relative pronoun ”that”). They are used in German to differentiate a function and they are not interchangeable (1):

  1.*”das” instead of ”dass”: ”Die Zahl 1 bedeutet,* ***das*** *der Tumor weniger als 1 Millimeter dick ist.” [”The number 1 means* ***the*** *the tumor is less than 1 millimeter thick” instead of ”***that*** *the tumor”]*

  Another example is given in (2):

  2.*”isst” (”eat”) vs. ”ist” (”is”): ”Wenn man nüchtern* ***isst****, geht es sehr schnell” [”When you* ***eat*** *sober, it happens very quickly” instead of ”When you* ***are*** *sober”]*

- wrong prepositions: ”durch” (”through”) vs. ”von” (”of”) (3):

  3.*”Erkrankung* ***von*** *Lebensmitteln” [”illness* ***of*** *food” instead of ”***through*** *food”]*

- wrong genus of nouns (4):

  4.*”***Das*** *Rücken wird dann immer schwerer und schwerer” [”The back then becomes heavier and heavier”], ”Es soll* ***kein*** *Rückfall geben” [”There should be no relapse”]*

Other errors contain verb numerus, genus of nouns, the syntax of clauses or sentences, especially in the passive voice, and other.
Semantic errors or inaccuracies are for example:

- Wrong explanations (5):

  5.*”am Tage mehrfach* ***wegdösen*** *” [ST, ”***dozing off*** *several times during the day”] vs. ”man* ***fällt*** *am Tag mehrmals* ***weg*** *und muss dann wieder aufwachen” [TT, ”***falling away*** *several times during the day and then having to wake up again”]*

- Terminology inaccuricities, e.g. (6):

  6.*”Ein erhöhtes Schlafbedürfnis am Tag, eine sogenannte Hypersomnie, ist eine oft kennzeichnende Folge solcher nächtlichen unbewusst oder bewusst erlebten Unterbrechungen” [ST, ”Hypersomnia is characterized by increased need for sleep during the day caused by sleep interruptions at night.”].* In the translation, this connection is no longer clear due to the information being abridged: *”Manchmal kann man nachts aufwachen und dann nicht mehr einschlafen. Das nennt man dann Hypersomnie.” [TT, ”Sometimes you can wake up at night and then not go back to sleep. This is then called hypersomnia”]*

- Polysemous words errors: In German, ”Satz” means ”sentence”, but also ”leap” (7):

  7.*”Ansonsten vermutlich der übliche* ***Satz*** *morgens aus dem Bett” [ST, ”Otherwise probably the usual* ***leap*** *out of bed in the morning”] vs. ”Sonst ist es wahrscheinlich der übliche* ***Satz****, den Sie morgens sagen” [TT, ”Otherwise it is probably the usual* ***sentence*** *you say in the morning.”].*

- Sentences with conditional meaning have a particularly high error rate, like in the following example (8):

  8.*”Bei fortgeschrittenen Tumorstadien […] ist eine umfassendere Behandlung notwendig” [ST, ”In advanced tumor stages [...] more comprehensive treatment is necessary”] vs. ”Wenn der Hautkrebs schon weiter fortgeschritten ist, gibt es mehr Möglichkeiten zur Behandlung” [TT, ”If the skin cancer is already more advanced, there are more options for treatment”]*

Correctness is not yet present for the system under study to the extent that texts would be usable without post-editing. The human translation corpus does not have such errors, but has a high degree of correctness.

3.3.2. Readability

Comparing the comprehensibility of the human and machine translations, as well as of the source texts, revealed that the machine translations had the highest comprehensibility, with a mean HIX value of 19.15 (SD: 0.49). In comparison, the human translations yielded a mean HIX value of 17.74 (SD: 1.67). Based on the HIX, the source texts were the least comprehensible (mean: 10.46, SD: 2.76). Given the low variance in the machine translations (see Figure 1), all of the 30 texts could be classified as Plain Language texts.

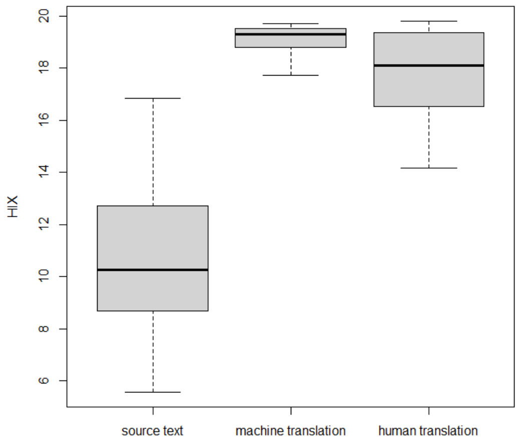

Figure 1: HIX values of the source text, the machine translation, and the human translation.

From the human translations, however, only 83% of the texts reached the predefined Plain German benchmark. As seen from the boxplot, human translations reveal a much greater variation in the HIX values than the machine-translated texts.
It is important to highlight that HIX values only consider overt complexity. Therefore, these values represent a starting point for evaluating comprehensibility, but have to be complemented with further qualitative analysis.

3.3.3. Syntactic Complexity

In the next step, we analysed the distribution of the dependency relations in human and machine translation, as well as in the source texts. We summarise the results (frequencies normalised per 10000) in Figure 2.
The distribution numbers reveal that both the source texts and machine translations seem to contain a higher number of complex syntactic relations than the human translation. For the latter, we observed higher number for parataxis relations only. This goes in hand with the results by Deilen et al. (2023) who reported similar tendencies for machine-generated texts - they contained more complex constructions than those generated by humans. However, the authors did not compare machine-generated texts with human translations. In our study, we had human translations at our disposal and observed that they are the most simplified amongst the subcorpora under analysis. Interestingly, sources contain more clausal subjects (csubj) as well as clauses modifying nouns (acl) than machine-translated texts do. It is also interesting to note that human translations do not contain any clausal subjects at all. At the same time, clausal complements of verbs and adjectives (ccomp) along with clauses modifying verbs and adjectives (advcl) predominate in machine translations. So do subjectless clausal complements (xcomp), whose number is significantly higher in machine translated-texts if compared to the other two subcorpora.
A sentence with a clausal subject that is frequent in source texts is illustrated in example 9a (*Wer... [Who...]*). Its corresponding machine translation in 9b contains a clause modifying a verb (*Wenn man merkt/If you realise*) complement of a verb and a clausal verb complement (starting with *dass.../that*). The only subclause contained in the human translation (in 9c) is parataxis. The other parts are simple sentences.

> 9a. *Wer bei sich Probleme im Umgang mit Alkohol feststellt, sollte daher unbedingt das Gespräch mit dem Arzt suchen. [...]* (source)
> 9b. *Wenn man merkt, dass man mit Alkohol Probleme hat, sollte man unbedingt zum Arzt gehen. [...]* (machine translation)
> 9c. *Sie glauben: Ich bin vielleicht alkoholsüchtig? Dann sprechen Sie mit Ihrem Arzt.* (human translation)

3.3.4. Automatic Evaluation Measures

In the last step, we analysed the SARI score of machine translated texts which is a quantitative measure of text simplification. The boxplot visualising the SARI score computed on all aligned segments is displayed in Figure 3. As already mentioned in Section 3.2.4 above, SARI compares machine translated output with the human translations and the sources measuring added, deleted or kept words. Higher SARI values indicate better machine translated outputs.
The system used in the analysis achieves an average SARI-Score of 40.67 (SD: 6.79), which is in line with state-of-the-art text simplification models reported by Sheang and Saggion (2021). We also see from the box plot that our maximum values achieved by the system are around 55. Moreover, the data contains many outliers, i.e. segments with the score of over 55.

4. Discussion and Future Work

The present paper evaluates the use of a machine translation system for translating medical texts into Plain German. Our results showed that in terms of readability, the machine translations are much easier than the source texts and even easier than the human translations. Analysing the syntactic complexity, however, revealed that machine translations contain a significantly higher number of

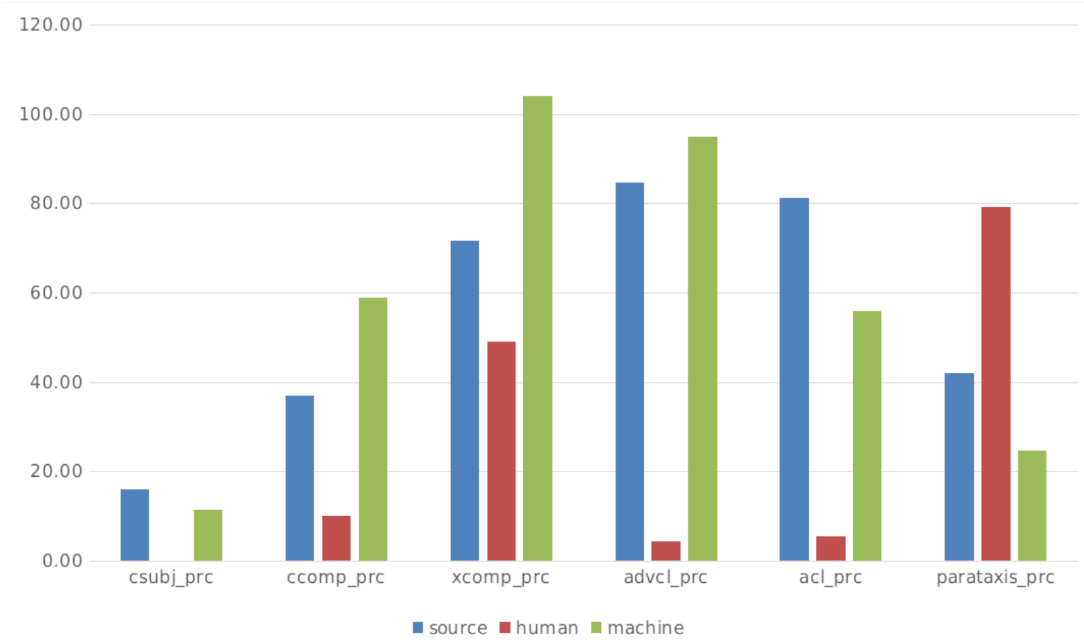

Figure 2: Distribution of syntactically complex dependency relations in the source texts, human and machine translations (normalised frequencies per 10000).

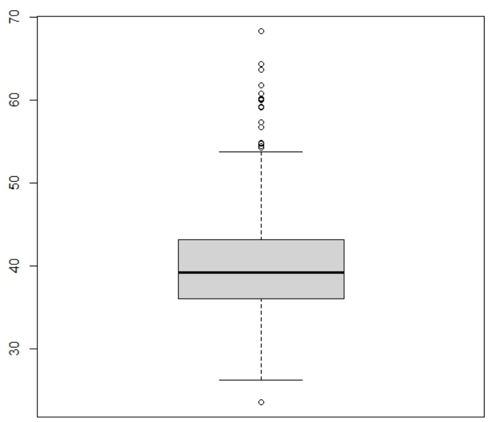

Figure 3: SARI score of the aligned source text, machine translation and human translation.

complex syntactic relations than human translations. Particularly interesting and against our expectations was the result that in most cases, the machine translations are even more complex than the source texts. Furthermore, our analysis revealed that the machine-translated texts contained various types of mistakes.

In our further research, we will proceed to classify the different types of machine translation errors and also look into different cases of partial correctness, where only some pieces of information were incorrect or missing. Furthermore, the present study only focused on the text perspective. However, to draw reliable conclusions about the functionality of a translation, not only the text but also the user perspective needs to be considered. Therefore, in our future work, we plan to conduct empirical studies, consisting of eye-tracking and reading experiments, to gain insights into the cognitive processing costs of the target groups when reading machine translated texts. In addition, we plan to use questionnaires to investigate whether the end users accept the generated texts.

All in all, we conclude that the analysed tool is a promising text simplification tool, however, in terms of correctness and syntactic complexity, it still does not reach the human parity. The machine translation system showed its limitations in the field of selecting and prioritizing information, including adequate examples and images, and adapting the content to the prior knowledge of the target groups, i.e. adding for example explanations of difficult words and concepts. Human translators are therefore still indispensable. It becomes very clear that machine translated Plain Language texts cannot do without post-editing, but need intensive revision. The translation tools at hand are therefore not yet suitable for end users, but are rather to be used as CAT tools for professional translators or experts in the relevant domain.

Another aspect we want to point out is the aspect of liability: When pondering the use of AI in intralingual translation, the translator or company also has to keep in mind that the human translator still assumes full liability for the translation (since machines are not liable). This is especially important in "safety-critical domains", which Canfora and

Ottmann (2020) define as ”those domains where translation errors can lead to injuries. Examples of safety-critical domains in translations are health-care, mechanical engineering, the chemical industry and power generation” (Canfora and Ottmann, 2020, p. 61). Thus, in medical discourse there is a high risk of safety-critical errors, which can result in serious damage. One of the reasons why these mistakes are especially dangerous is that post-editors seem to have difficulties to detect them in the raw machine translation output (Canfora and Ottmann, 2020). This underlines the importance of professional post-editing competences. Translators must be trained to detect and correct different types of errors, especially those that are critical for user safety.

Still, by constant training, SUMM AI is currently working towards improving their machine translation system using in-domain data. To investigate whether the trained and improved version of their machine translation system yields better results than the current one, we also plan to conduct a second, comparative study. A machine translation system that has evolved through several iterations and has achieved a satisfactory level of liability, coupled with the professional post-editing skills of a translator or a suitably trained editor would represent a breakthrough for the editorial process:

Editors would be able to publish a much larger volume of texts in Plain Language with greater frequency. Scientific review, however, would still have to be done with the same meticulousness as with human translations. But the translation process would be much faster. This could be a real milestone in the field of accessible health communication. As in the future, all essential questions – even current ones – on diseases, medications and preventive health care should also appear in the accessible and at the same time acceptable form of Plain Language. According to the National Action Plan for Health Literacy (Schaeffer et al., 2018b) this could contribute significantly to promoting health literacy in the population.

5. Acknowledgements

This research was funded by the Ministry of Science and Culture of Lower Saxony and its program ”zukunft.niedersachsen”.

6. Bibliographical References

Sarah Ahrens, Rebecca Schulz, Janina Kröger, Sergio Hernández Garrido, Loraine Keller, and Isabel Rink. 2022. Accessible communication and health literacy. *Accessibility–Health Literacy–Health Information: Interdisciplinary Approaches to an Emerging Field of Communication*, 13:9.

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

Ursula Bredel and Christiane Maaß. 2016. *Leichte Sprache: Theoretische Grundlagen, Orientierung für die Praxis*. Dudenverlag.

Carmen Canfora and Angelika Ottmann. 2020. Risks in neural machine translation. *Translation Spaces*, 9(1):58–77.

Silvana Deilen. 2021. *Optische Gliederung von Komposita in Leichter Sprache. Blickbewegungsstudien zum Einfluss visueller, morphologischer und semantischer Faktoren auf die Verarbeitung deutscher Substantivkomposita*. Frank & Timme.

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 (TSAR)*, RANLP, Varna, Bulgaria. ACL.

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.

Silke Gutermuth. 2020. *Leichte Sprache für alle?: eine zielgruppenorientierte Rezeptionsstudie zu Leichter und Einfacher Sprache*, volume 5. Frank & Timme GmbH.

Silvia Hansen-Schirra, Jean Nitzke, Silke Gutermuth, Christiane Maaß, and Isabel Rink. 2020. Technologies for translation of specialised texts into easy language. *Easy Language Research: Text and User Perspectives. Berlin: Frank & Timme*, pages 99–127.

Julian Hörner. 2022. Accessible and reliable health information on the internet. *Accessibility–Health Literacy–Health Information: Interdisciplinary Approaches to an Emerging Field of Communication*, 13:77.

Roman Jakobson. 1959. On linguistic aspects of translation. In *On translation*, pages 232–239. Harvard University Press.

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

Christiane Maaß. 2020. *Easy Language–Plain Language–Easy Language Plus: Balancing Comprehensibility and Acceptability*. Frank & Timme.

Christiane Maaß. 2024. Translation into Easy Language. *Handbook of Accessible Communication*, pages 261–289.

Christiane Maaß and Isabel Rink. 2021. Translating legal texts into easy language. *J. Open Access L.*, 9:1.

Christiane Maaß, Isabel Rink, and Christiane Zehrer. 2014. Leichte Sprache in der Sprach- und Übersetzungswissenschaft. *Sprache barrierefrei gestalten. Perspektiven aus der Angewandten Linguistik. Berlin: Frank & Timme*, pages 53–85.

Christiane Maaß and Laura Marie Schwengber. 2022. Easy Language and Plain Language in Germany. *Rivista internazionale di tecnica della traduzione= International Journal of Translation*.

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.

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.

Isabel Rink. 2019. *Rechtskommunikation und Barrierefreiheit: Zur Übersetzung juristischer Informationsund Interaktionstexte in Leichte Sprache*. Frank & Timme.

Horacio Saggion. 2017. Applications of automatic text simplification. In *Automatic Text Simplification*, pages 71–77. Springer.

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.

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718, Melbourne, Australia. Association for Computational Linguistics.

D Schaeffer, EM Berens, S Gille, L Griese, J Klinger, S de Sombre, D Vogt, and K Hurrelmann. 2021. *Gesundheitskompetenz der Bevölkerung in Deutschland vor und während der Corona Pandemie: Ergebnisse des HLS-GER 2*. Interdisziplinäres Zentrum für Gesundheitskompetenzforschung.

Doris Schaeffer, Svea Gille, and Klaus Hurrelmann. 2020. Implementation of the national action plan health literacy in Germany – lessons learned. *International Journal of Environmental Research and Public Health*, 17(12):4403.

Doris Schaeffer, Klaus Hurrelmann, Ullrich Bauer, and Kai Kolpatzik. 2018a. Nationaler Aktionsplan Gesundheitskompetenz. *Die Gesundheitskompetenz in Deutschland stärken. Berlin: KomPart*, 10:0418–1866.

Doris Schaeffer, Klaus Hurrelmann, Ulrich Bauer, and Kai Kolpatzik. 2018b. National Action Plan Health Literacy. Promoting Health Literacy in Germany. *Journal of Public Health*, 31:905—915.

Doris Schaeffer, Dominique Vogt, Eva-Maria Berens, and Klaus Hurrelmann. 2017. *Gesundheitskompetenz der Bevölkerung in Deutschland: Ergebnisbericht*. Bielefeld: Universität Bielefeld, Fakultät für Gesundheitswissenschaften.

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.

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 *Emerging Fields in Easy Language and Accessible Communication Research*, pages 339–369. Springer.

Birgit Welch and Gabriele Sauberer. 2019. Easy-to-read language and terminology: New needs, new rules, new software? In *Systems, Software and Services Process Improvement: 26th European Conference, EuroSPI 2019, Edinburgh, UK, September 18–20, 2019, Proceedings 26*, pages 647–658. Springer.

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.

7. Language Resource References

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.