# Evaluation of intralingual machine translation for health communication

**Silvana Deilen[1], Ekaterina Lapshinova-Koltunski[1], Sergio Hernández Garrido[1], Christiane Maaß[1], Julian Hörner[2], Vanessa Theel[3], Sophie Ziemer[4]**
[1] University of Hildesheim, [2] Wort & Bild Verlag, [3] SUMM AI, [4] Johannes Gutenberg University Mainz
[1]deilen,lapshinovakoltun,hernandezs,maass@uni-hildesheim.de,
[2]j.hoerner@wubv.de, [3] vanessa@summ-ai.com, [4] sziemer@students.uni-mainz.de

## 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 outputs from three models based on different criteria, such as correctness, readability, and syntactic complexity. We compare the outputs of the three models under analysis between each other, as well as with the existing human translations. The study revealed that system performance depends on the evaluation criteria used and that only one of the three models showed strong similarities to the human translations. Furthermore, we identified various types of errors in all three models. These included not only grammatical mistakes and misspellings, but also incorrect explanations of technical terms and false statements, which in turn led to serious content-related mistakes.

## 1 Introduction

In Germany, according to recent studies in the field of Public Health, over half of the population reports having difficulties with health-related topics (Schaeffer et al., 2021). For that reason, the promotion of health literacy (knowledge and competences to access, understanding, appraise and apply medical information) has turned into an important task for the German health system (Schaeffer et al., 2018) (for an extensive definition of health literacy, see Sørensen et al. (2012)). In this context, recent research has underlined the need for accessible communication in the medical domain to effectively promote health literacy and consequently assist patients navigating the health system and improve patient understanding, engagement and compliance with medical recommendations (Ahrens et al., 2022; Blechschmidt, 2021; Schaeffer et al., 2021). Plain German is a prominent form of accessible communication that has gained relevance in health communication scenarios (Schaeffer et al., 2018).

Although there is an urgent need for translations into Plain German, there is also a gap in qualified and experienced human translators (Maaß, 2020). Moreover, there is a lack in computer-aided translation (CAT) tools and machine translation systems for this kind of intralingual translation. Unfortunately, little is known about existing systems and their performance for different texts that are required to be translated into Plain German, as for instance, texts in health communication that we focus on.

In our study, we evaluate machine translations of medical texts into Plain Language. The source texts, as well as reference human translations, are derived from the website of the German health magazine *Apotheken Umschau*. We analyse machine-translated texts produced with three models comparing them with human translations from the magazine's website. Besides that, we compare all translations with the underlying sources.

In the following, we present the results of the qualitative and quantitative analysis. Section 2 de-





*Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 469–479
June 24-27, 2024 ©2024 European Association for Machine Translation

scribes related work. In Section 3, we present our research design including corpus and the methods used. Section 3.3 presents the results of our analyses, while we discuss those results, as well as limitations and possible extensions in the future in Section 4.

## 2 Related Work

### 2.1 Plain Language

Both Easy Language and Plain Language are complexity-reduced language varieties which aim to improve readability and comprehensibility of 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 (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 textual features of Plain Language are closer to the standard language and are mainly a means to open expert contexts for lay people (Maaß, 2020). Therefore, the main target audience of Plain Language are non-experts 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 (Maaß et al., 2021; Deilen et al., 2023b). The studies point in two basic directions: studies on text qualities and possible barriers in various forms of communication on the one side (Rink, 2019) and studies on comprehensibility and recall by different target groups on the other (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 (Maaß, 2020). In comparison to Easy Language, Plain Language has the advantage of having less to no stigmatizing features, which is one of the reasons why it 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) therefore models the variety Easy Language Plus, which is situated between Easy Language and Plain Language and strikes a balance between comprehensibility and acceptability[1].

### 2.2 Accessibility in Medical Domain

In 2016, the Health Literacy Survey (HLS-GER) revealed that more than half of the German population (54.3%) encounters significant challenges in locating, understanding, appraising, and effectively using health-related information. These findings, which according to Schaeffer et al. (2017) were "significantly worse than expected" increased the awareness of the need for accessible health information, which in turn led the development of the National Action Plan Health Literacy (Schaeffer et al., 2018). According to this plan, one approach to increase health literacy in Germany is providing information in Plain Language, i.e., in a complexity-reduced variety of German. With the release of updated data from the second Health Literacy Survey (HLS-GER 2) in 2021, the importance of Plain Language in German health communication has been underscored, as it has shown that even more people (58.8%) encounter difficulties navigating the healthcare system. As a remedy for low health literacy, both practitioners and researchers increasingly advocate for the use of Plain Language. One of the most prominent examples of implementing this approach is the *Apotheken Umschau*. The *Apotheken Umschau*, which is Germany's leading health publisher and the largest consumer medium in the German-speaking area with a traffic of 6.68 m. visits and 49.11 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). By offering information in both standard German and Plain German, their goal is to provide accessible and reliable information on illnesses, medications, and preventive healthcare

[1] It should be noted that Plain Language is an international concept and not language-bound. However, in this paper we only focus on Plain Language in Germany, also called Plain German.

[2] `https://ausweisung-digital.ivw.de/`, retrieved 15.03.2024

with minimal barriers for all individuals.

### 2.3 NLP for Plain Languages

In Easy and Plain Language translation, which both belong to the domain of intralingual translation (Hansen-Schirra et al., 2020a), the potentials of using CAT tools are still a major research desideratum. There are some studies that have discussed the challenges of using CAT tools in intralingual translation compared to interlingual translation (Hansen-Schirra et al., 2020b; Spring et al., 2023; Kopp et al., 2023). For example, in contrast to interlingual translation, in intralingual translation there is usually no 1:1 correspondence between source and target sentences, which in turn means that the sentence alignment process has to be done or corrected manually by the translator, which increases the workload for translators instead of reducing it.

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, while ideally conveying 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[3]. 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.

Ondov et al. (2022) surveyed the literature in the field of automated methods for biomedical text simplification and found that one major challenge in this field is the lack of high-quality parallel text data, which so far impedes the development of fully automated biomedical text simplification methods.

Specific problems of automatic systems for 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, so far, 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. (2023a) drew similar conclusions for the outputs produced with ChatGPT. The authors investigated the feasibility of using ChatGPT 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 texts produced by ChatGPT were easier than the standard source 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 maximally simplified and strictly controlled language variety adapted to the needs of people with reading impairments, we focus on translation into Plain German (see 2.1). 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

Our dataset contains 200 parallel texts selected from the website of the German health magazine *Apotheken Umschau*[4]. The texts cover a broad range of topics, such as breast cancer, vaccination, long COVID, food poisoning, first aid and others. For all texts in the sample, a human translation into Plain Language was already available. 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. Furthermore, the human translations also comply with a practical concept for Plain Language for this specific health information scenario, which was established by the Research Centre for Easy Language[5]. We split the data into

[3] https://newsela.com/data

[4] apotheken-umschau.de

[5] www.uni-hildesheim.de/leichtesprache

test and train sets: 30 texts were selected to serve as test data in our evaluation study and the remaining 170 texts were used as training data for two of the three tested systems. The sample of 30 texts was translated using the machine translation system SUMM AI[6]. At the time of the study, SUMM AI was the only tool known to us for intralingual translation from standard German into Easy and Plain German and the only one with a specific focus on health communication texts.

In our study, we compared three different models of SUMM AI: the baseline model and two further fine-tuned models. The baseline model of our study was the already existing beta-model for Plain Language provided by SUMM AI. The model is a fine-tuned large language model (LLM) that was trained with in-house data and further rule-based approaches. In comparison to the baseline model, two further models (model 1 and model 2 hereafter) were enriched by SUMM AI with the training data (170 parallel texts out of 200 selected). The data was aligned and adapted according to the practical concept of the Research Centre for Easy Language. While the baseline model and model 1 have the same underlying LLM, model 2 is distinguished by a different underlying LLM.

We investigate which of the three models yield better results in translating standard German into Plain German. For this, the 30 texts from the test set are translated with the three models under analysis[7].

## 3.2 Data Analysis

### 3.2.1 Analysis Steps

The resulting machine translations (three per each texts) are also compared with the existing human translations and the underlying sources. We follow the evaluation criteria suggested by Deilen et al. (2023a), which is one of the few studies known to us that evaluates intralingual machine translation. We assess machine translations for the correctness of the content, the readability of the texts, and their syntactic complexity. Readability as well as syntactic complexity are also assessed for human translations and source texts. We then compare sources, human, and machine translations according to these two criteria.

### 3.2.2 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: translators, content providers, or Plain German end users (for an overview over end users, see Bredel/Maaß, 2016 and Maaß 2020). 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.

### 3.2.3 Readability

We also compared the readability of human and the 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). As Plain German is more complex than Easy German, we suggest setting the benchmark for Plain German at 16 points.

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

[7]The whole dataset will be published on GitHub. The GitHub repository will contain the selected texts (sources, human and machine translations), including the raw data, the parsed data (conllu) and the Textlab analyses per text.

#### 3.2.4 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)[8] 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 components (ccomp), clausal subjects (csubj), open clausal elements (xcomp) and parataxis relation (parataxis). These selected categories are all listed under the clause dependents[9] 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 sub-corpora are.

#### 3.2.5 Automatic Evaluation Measures

We also applied SARI (Xu et al., 2016), which is a quantitative measure to evaluate automatic text simplification systems. The metric "compares system output against references and against the input sentence" (Xu et al., 2016) and is normally used for evaluation of automatic text simplification models but could also be used to evaluate intralingual translation.

While SARI is normally calculated on a sentence basis, this is not possible in the case of Plain Language since there usually is no sentence-to-sentence alignment but rather an alignment on paragraph level. To calculate these metrics, we aligned the source texts, machine translations, and human translations on a paragraph level and assessed their alignment quality. Since the translation into Plain Language compared to interlingual translation is significantly more liberal in terms of which information is translated, adequate alignment was difficult and only possible for 263 of 946 segments.

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

[9] https://universaldependencies.org/u/dep/

#### 3.2.6 Translation Comparison

In the last step, we compared the performance of the systems taken all criteria together. After that, we compared them with the existing human translations, as well as the underlying source texts. For this, we used an explorative multivariate technique called Correspondence Analysis (CA) performed with the package `ca` in R environment (R Core Team, 2017, R version 3.6.1).

Correspondence analysis (Greenacre, 2007) helps to explore relations between variables in a data set (both those constituting the rows and those in columns) and summarises and visualises data in a two-dimensional plot. We use CA to see which variables, in our case subcorpora representing source texts (source), human translations (human) and the three machine-translated outputs (baseline, model 1 and model 2), have similarities and how these subcorpora correlate with the analysed features (HIX values, syntactic structures) contributing to the similarities. Weighted Euclidean distances, termed the $\chi^2$ distances are measured on the basis of the distributions of these feature across the five subcorpora under analysis. The row (subcorpora) and the column (features) projections are then plotted on the same graph. The larger the differences between the subcorpora, the further apart they are on the map. Proximity between subcorpora and features in the merged map is an approximation of the correlation between them. The position of the dots (subcorpora) and triangles (features) indicates the relative importance of a feature for a subcorpus (see Figure 4). With the help of this technique, we will observe which texts are more similar between each other.

### 3.3 Results

#### 3.3.1 Correctness

The analysis of the correctness of the machine translations showed that from the baseline model, only one of the 30 texts was correctly translated. The other 29 texts showed problems with regard to their correctness in different aspects. Model 1 yielded similar results, with only two out of thirty texts being classified as correct. For model 2, however, we found that 15 out of 30 texts were translated correctly. 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. We encounter grammatical errors and misspellings, omissions of relevant pronouns or words, incorrect explanations of technical terms, incorrect statements and advice, wrong segmentation of compounds, 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 Section 4). Some examples of the errors we found are given in the following.

- Missing segmentation signs: In some cases, segmentation signs would facilitate the processing, but the tool fails to apply them. This is especially true for polymorphemic compounds (i.e., compounds consisting of at least three free morphemes), such as *"Nasennebenhöhlenentzündungen"* (model 2) *(inflammation of the sinus cavities)*, in which indicating the morpheme boundaries would have reduced the compound's complexity.

- Redundancies and unreasonable statements:
  - *"Bei Männern kann eine Blasenentzündung auch die Prostata entzünden. Oder die Prostata entzündet sich. Dann kann sich die Prostata entzünden."* (model 1) *(In men, a bladder infection can also inflame the prostate. Or the prostate becomes inflamed. Then, the prostate becomes inflamed.)*
  - *"Dann kann eine Person eine Nierenbeckenentzündung oder eine Nierenbeckenentzündung bekommen."* (model 1) (*Then, a person can get an urinary tract infection or an urinary tract infection.)*
  - *"Eine Insekten-Stich ist eine allergische Reaktion auf einen Insekten-Stich."* (model 1) *(An insect bite is an allergic reaction to an insect bite.)*
  - *"Frauen haben oft eine Blasenentzündung, weil sie oft auf Toilette müssen."* (model 1) *(women often have a bladder infection because they often have to go to the toilet.)*

- Lexico-semantic errors:
  - *"Viele Menschen nehmen zu wenig Schlaf"* (model 1) *(Many people take too little sleep.)*
  - *"Wenn andere Menschen sich Sorgen um Sie machen, ist auch ein Zeichen."* (baseline) *(When other people are worried about you, is also a sign.)*

- Omission of reflexive pronouns:
  - *"Dann können sie gut konzentrieren"* (model 1) *(Then they can concentrate well.)*
  - *"Vielleicht haben Sie auch zu tief gebückt"* (baseline) *(Maybe you have bent over too far.)*

  In German, both verbs (*"concentrate"* and *"bend over"*) require the reflexive pronoun "*sich*" (themselves or yourself), which, however, the tool omitted.

- Incorrect statements and advice:
  - *"Nehmen Sie Ihren Helm ab"* (model 2) *(take off your helmet)*. In this text about first aid, the tool erroneously capitalized the pronoun "Ihren", which therefore refers to the second person singular instead of the third person singular. The correct spelling would be lowercase ("ihren").
  - *"Sie können die Pille auch in der Schwangerschaft nehmen"* (model 1) *(You can also take birth control pills during pregnancy.)*
  - *"Und Sie sollten alles tun, was Ihren Gelenken schadet."* (model 1) *(And you should do anything that harms your joints)*. The verb of the source text "meiden" (avoid) was translated with its antonym "tun" (do). Thus, the reader is even given harmful advice.
  - *"Bei etwa 14 Prozent der Patienten [. . . ] ist die Herz-Kranz-Gefäße verengt."* (model 1) *(In about 14 percent of patients, the coronary arteries is narrowed)*. In this case, not only the verb form is incorrect (singular instead of plural), but the tool also failed to translate the negated statement of the source text *("findet sich [...] keine Verengung der*

*Koronargefäße)" (no narrowing of the arteries is found).*

- Incorrect explanations of technical terms: *"Die Zeit, in der man krank ist, nennt man Inkubationszeit."* (baseline) *(The period during which one is sick is called incubation period)*. This is incorrect because the incubation period is the time between the infection and the manifestation of symptoms.

- Wrong relation: Source text: *"Deshalb ist hier unbedingt ein Arztbesuch angeraten. Auch wenn die Symptome länger als drei Tage anhalten [...] wird der Besuch beim Arzt unumgänglich" (Therefore, seeing a doctor is strongly recommended here. Also, if the symptoms persist for more than three days, a visit to the doctor is inevitable.)* vs. Target text: *"Bei diesen Menschen kann eine Lebens-Mittel-Vergiftung schwerer verlaufen. Deshalb sollten Sie bei diesen Anzeichen sofort zum Arzt gehen: Die Beschwerden dauern länger als 3 Tage."* (model 1) *(In these individuals, food poisoning can progress more severely. Therefore, if you experience these symptoms, you should see a doctor immediately: The symptoms last longer than 3 days.)*. In the source text, the word "*hier*" (here) refers to vulnerable groups of people; however, it is erroneously translated with "symptoms". As a result, the target text states that these individuals should only see a doctor when they experience one of the following symptoms, while the source text indicates that vulnerable people have to see a doctor in any case.

- Homophonic but not homographic words are not correctly selected: *"Dann 7 Sie den Saft durch ein Tuch oder einen Kaffeefilter."* (model 2) *(Then strain the juice through a cloth or a coffee filter)*. In German, the word "*sieben*" is both a verb (strain) and a number (7).

Correctness is not yet present for the different systems 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 different corpora revealed that, as expected, the source texts were the least comprehensible texts (mean: 10.46, SD: 2.76). Model 2 had the highest comprehensibility, with a mean HIX value of 19.5 (SD: 0.76). While this is a slight improvement compared to the baseline model (mean: 19.15, SD: 0.49), Figure 1 shows that the HIX value for model 1 (mean: 17.71, SD: 1.41) was considerably lower than that of the baseline model, i.e., based on the HIX, the model's comprehensibility was not improved. However, as seen from the boxplot, human translations also yielded a lower HIX value (mean: 17.74, SD: 1.67) than the baseline model, and both the human and the model 1 translations reveal a much greater variation in the HIX values than the baseline and model 2 translations. While from the model 1 translations, only 93% of the texts, and from the human translations, only 83% of the texts reached the predefined Plain German benchmark, all of the baseline and model 2 texts could be classified as Plain German texts. However, when interpreting the HIX value, it should be kept in mind that this is only a quantitative analysis that focuses only on comprehensibility features on the text surface (i.e. overt complexity) and the textual level is mainly ignored. For this reason, HIX values only represent a starting point for the analysis and it has to be complemented by a qualitative analysis (e.g. Section 3.3.1).

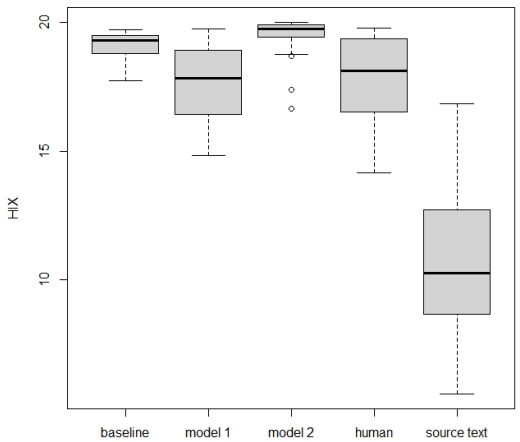


**Figure 1:** HIX values of the three machine translations, the human translations, and the source texts.

### 3.3.3 Syntactic Complexity

As seen from Figure 2, human translations contain the least number of complex syntactic con-

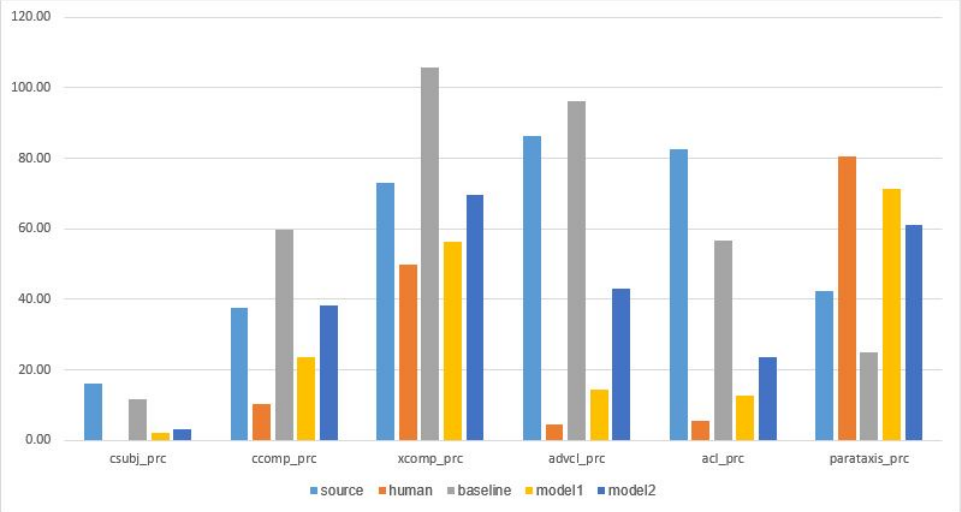


**Figure 2:** Distribution of syntactically complex dependency relations in the source texts, human and machine translations (normalised frequencies per 10000).

structions, except for parataxis.

In machine translation outputs, we observe the following pattern: model 1 contains the least number of complex syntactic constructions, followed by model 2. Here again, the only exception is the distribution of the parataxis constructions. Translations with the baseline model are much more complex in terms of syntax if compared to the other two systems. Remarkably, for some structures, they are even more complex than the source texts.

### 3.3.4 Automatic evaluation measures

In the final stage, we examined the text simplification metric SARI. Figure 3 displays box-

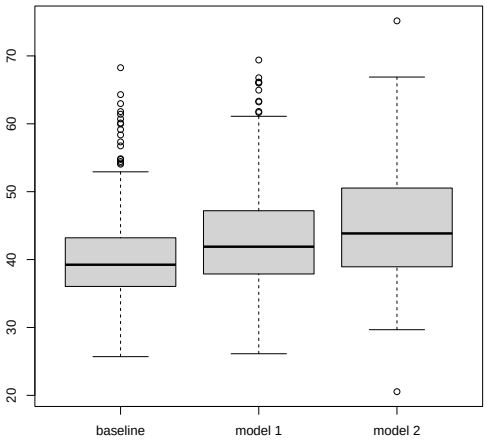


**Figure 3:** SARI score of the machine translation output from the baseline model, model 1 and model 2.

plots, comparing the SARI of the machine translation output from the baseline model, model 1 and model 2. Higher SARI values indicate better machine translated outputs.

The system utilized in our analysis achieves an average SARI score of 40.61 (SD: 6.78) for the baseline model, 43.49 (SD: 7.84) for model 1 and 45.13 (SD: 8.15) for model 2. All models are therefore in line with with state-of-the-art text simplification models reported by Sheang and Saggion (2021).

### 3.3.5 Translation Comparison

We now summarise the results of all evaluation criteria for the system outputs. Table 1 illustrate the system ranking depending on the used criteria.

| system | corr | HIX | synt | SARI |
|---|---|---|---|---|
| **baseline** | 3 | 2 | 3 | 3 |
| **model 1** | 2 | 3 | 1 | 2 |
| **model 2** | 1 | 1 | 2 | 1 |

**Table 1:** System ranking according to the evaluation criteria: corr=correctntess, HIX=readability, synt=syntax, and SARI=text simplification.

As seen from the table, the worst outputs according to our evaluation criteria are found with the baseline system. The results for models 1 and 2 vary depending on the evaluation criteria. For instance, model 1 performs better in terms of syntax, as its outputs reveal not so many complex syntactic constructions. This system seems to be very close to human translations as well. However, many of the texts translated with model 1 are not correct. In terms of correctness, as well as readability scores and text simplification scores, model 2 is the win-

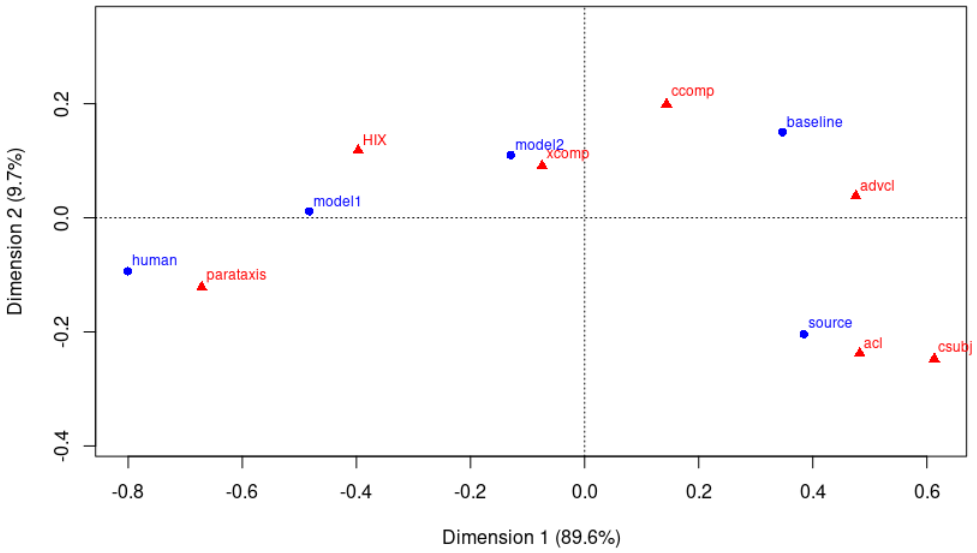


**Figure 4:** Correspondence analysis based on HIX scores and distribution of syntactic structures.

ner.

For the comparison of the machine-translated outputs with the human translations and sources, we only use HIX scores and distributions of the complex syntactic structures. The results of the correctness analysis are not numeric. The resulting two-dimensional graph is shown in Figure 4.

The most obvious information we can obtain from this graph is that the difference is most strongly pronounced along the x-axis between the two groups of subcorpora: source texts and translations with the baseline model on the right side vs. human translations and outputs of the two other models (model 1 and model 2) on the left side. This difference is considerable, as the dimension along the x-axis explains a very high proportion (89.6%) of the data variance. We also see that translations with model 1 are the closest to the human translations. On the y-axis, we see a separation between human- vs. machine-authored texts. However, this difference is not big, as it explains only 9.7% of the data variance in our dataset.

## 4 Discussion and Future Work

The present paper evaluates three different models of a machine translation system for translating medical texts into Plain German. It covers one of the first steps towards the implementation of these tools in accessible health communication in Germany and it discusses first methodological approaches, which we intend to expand on in further research.

Model 2 seems to achieve the best results according to most criteria. At the same time, model 1 seems to be more similar with the human translations at hand. While in terms of syntactic complexity and text readability, the models yielded promising results, the evaluation of the correctness revealed severe misinformation for all three models, the consequence being that the texts cannot be safely used by end users. At the same time, the tool under analysis can be used as a CAT tool for professional translators and content providers with an expertise in Plain Language and post-editing. As our study has clearly revealed that so far machine translated Plain Language texts cannot do without post-editing, but need intensive revision, professional post-editing competences are more important than ever. This means that translators and experts working on machine translated text must be trained to detect and correct different types of errors, especially those that are critical for user safety. In further steps, a guide for post-editing in intralingual translation will be developed, exposing the necessary competences and factors to be considered when using machine translation into Plain German.

In a next step, the machine translation system will now be integrated into the editorial workflow of the *Apotheken Umschau* on a trial basis. This

practice test will serve to assess the time and effort that is needed to post-edit the machine translated texts. Adding machine translation to the editorial process could optimize the process and addresses the gap between the need for texts in Plain German and the lack of professional translators (see Section 1). The metrics from the practice test will be particularly interesting because the tool will only be permanently integrated into the workflow if the time and effort for post-editing the output is lower than for translating from scratch. Therefore, these metrics will determine the final decision for or against adapting the status quo.

In our future research, we will conduct a thorough analysis and classification of the various error types found in the machine translated texts. For example, we plan to investigate specific linguistic phenomena, such as the translation of compound words.

In addition, we also want to test and compare the output from both SUMM AI and other state-of-the-art systems to investigate which of the currently available systems is most suitable for intralingual translation into Plain German, in general and for specific subjects and text types. These systems include both freemium tools that offer both free and paid plans, such as ChatGPT and Google Gemini, and commercial tools, such as Klartext St. Pauli, capito digital and T2K (text2knowledge). In these future studies, we also plan to use other text types and texts from other domains, so that we are able to compare not only different tools but also different datasets.

## References


Ahrens, Sarah, 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.

Anschütz, Miriam, 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*.

Blechschmidt, Anja. 2021. Health literacy and multimodal adapted communication. *New Approaches to Health Literacy: Linking Different Perspectives*, pages 65–82.

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

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

Deilen, Silvana, Sergio Hernández Garrido, Ekaterina Lapshinova-Koltunski, and Christiane Maaß. 2023a. 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.

Deilen, Silvana, Silvia Hansen-Schirra, Sergio Hernández Garrido, Christiane Maaß, and Anke Tardel. 2023b. Emerging fields in easy language and accessible communication research. In *Emerging Fields in Easy Language and Accessible Communication Research*, pages 9–15. Springer.

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

Greenacre, Michael J. 2007. *Correspondence Analysis in Practice*. Chapman & Hall/CRC, Boca Raton.

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

Hansen-Schirra, Silvia, Walter Bisang, Arne Nagels, Silke Gutermuth, Julia Fuchs, Liv Borghardt, Silvan Deilen, Anne-Kathrin Gros, Laura Schiffl, and Johanna Sommer. 2020a. Intralingual translation into easy language–or how to reduce cognitive processing costs. *Easy Language Research: Text and User Perspectives. Berlin: Frank & Timme*, pages 197–225.

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

Kopp, Tobias, 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 Deilen, Silvana, 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.

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

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

Maaß, Christiane, Isabel Rink, Silvia Hansen-Schirra, Camilla Lindholm, and Ulla Vanhatalo. 2021. Easy language in Germany. *Handbook of Easy Languages in Europe*, 8:191.

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

Maddela, Mounica, 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, June. Association for Computational Linguistics.

Martin, Louis, 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, May. European Language Resources Association.

Ondov, Brian, Kush Attal, and Dina Demner-Fushman. 2022. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11):1976–1988.

R Core Team, 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

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

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

Säuberli, Andreas, 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.

Scarton, Carolina 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, July. Association for Computational Linguistics.

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

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

Schaeffer, Doris, Eva-Maria Berens, Svea Gille, Lennert Griese, Julia Klinger, Steffen de Sombre, Dominique Vogt, and Klaus Hurrelmann. 2021. Gesundheitskompetenz der Bevölkerung in Deutschland vor und während der Corona Pandemie: Ergebnisse des HLS-GER 2. Technical report, Universität Bielefeld, Interdisziplinäres Zentrum für Gesundheitskompetenzforschung.

Sheang, Kim Cheng 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, August. Association for Computational Linguistics.

Sørensen, Kristine, Stephan Van den Broucke, James Fullam, Gerardine Doyle, Jürgen Pelikan, Zofia Slonska, Helmut Brand, and (HLS-EU) Consortium Health Literacy Project European. 2012. Health literacy and public health: a systematic review and integration of definitions and models. *BMC public health*, 12:1–13.

Spring, Nicolas, 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.

Xu, Wei, 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.