SKASE Journal of Translation and Interpretation, 2025; 18(2): 38–54
doi: 10.33542/JTI2025-S-3

Evaluation of translations into Plain German produced by humans and MT systems including ChatGPT

Sarah Ahrens[*], Silvana Deilen[*], Sergio Hernández Garrido[*], Ekaterina Lapshinova-Koltunski[*], and Christiane Maaß[*]

[*]University of Hildesheim

Abstract
In this paper, we present the results of a study evaluating intralingual machine translations of health information texts into Plain German. In our study, we compare machine-translated simplified texts with those simplified manually by human translators, as well as with the original, unsimplified texts. We compare the output of four different machine translations systems and assess the translation quality using various criteria, including translation errors, readability, and syntactic complexity. The study reveals that from the four analysed machine translation systems, ChatGPT performed worst. Our results also suggest that fine-tuning a model with task-specific and domain-specific data improves the translation quality.

Keywords: *Plain Language; simplified German; text simplification; health communication; Large Language Models; ChatGPT*

1. Introduction

This paper focuses on the analysis of human and machine translations of health information texts into Plain German. As a form of accessible communication, Plain Language has become increasingly significant in health communication (Schaeffer et al. 2018; Deilen et al. 2024a; Ahrens 2025; Maaß 2024a). Translating a text from standard German into Plain Language belongs to the field of intralingual translation (Hansen-Schirra et al. 2020), which is defined as "an interpretation of verbal signs by means of other signs in the same language" (Jakobson 1959: 233).

Despite the urgent need for Plain Language translations, there is a shortage of qualified and experienced human translators as well as a lack of computer-aided translation (CAT) tools and high-quality machine translation (MT) systems for this type of intralingual translation. Even though more and more MT systems are currently being developed and released, there is limited information on the performance of existing systems for translating texts into Plain Language, particularly in the context of health communication.

Recent studies have revealed that over half of Germany's population have low health literacy, i.e., find it difficult to access, understand, appraise, and apply health information (Schaeffer et al. 2021). Consequently, enhancing health literacy has become a crucial aim for the German healthcare system (Schaeffer et al. 2018). Almost 75% of the general population show low health literacy in the dimension "appraise" (Schaeffer et al. 2021: 26), and digital health literacy is particularly low (Schaeffer et al. 2021: 68). Research shows that accessible communication is relevant to promote health literacy. Accessible communication helps patients to navigate the health system and to better comprehend, engage with, and adhere to medical advice (Schaeffer et al. 2021).

The emergence of MT tools that use Large Language Models (LLMs) that are readily available to the public enhances the importance of digital health literacy. While understanding health information may become easier, its appraisal becomes harder, knowing that such translation tools may present false and even harmful information. Therefore, in this study, we analyse machine-translated texts in Plain German, comparing them with human translations. The goal is to show how the different models perform and what types of problems they show. We also demonstrate that methods of translation error analysis from interlingual translation can also be applied to intralingual translation. In doing so, we contribute to the methodological development of research on intralingual MT. Apart from looking into compliance with the rules of Plain Language (such as readability and syntactic simplicity), we specifically pay attention to the correctness of the content and the errors that occur in machine translations. We evaluate the machine-generated translations produced by four different models and compare them with the professional human gold-standard translations. Additionally, we also compare all translations with the original source texts written in standard German. Both the source texts and human gold standard translations (text in Plain Language) were taken from the website of the German health magazine *Apotheken Umschau*[1].

The following sections provide an overview of our findings. Section 2 reviews related research. Section 3 outlines our research design, including the corpus and methods used. Section 3.3. details the results of our analyses, and in Section 4, we discuss these results along with their limitations and potential avenues for future research.

2. Related work

2.1. Plain Language

Plain Language is a means to make expert information accessible to lay readers (Maaß 2020, 2024a). The target audience of Plain Language texts are non-experts with average or slightly below average language or reading skills (Maaß 2020). Intralingual translation is often required in settings with asymmetric knowledge distribution or knowledge gaps between the communication participants, for example in the field of medical and health communication. Even though in intralingual translation, there is "no classical language change in the sense of interlinguality, which is typically regarded as a criterion for the definition of translation" (Hansen-Schirra et al. 2020: 199), we agree with researchers such as Schubert (2013), Bredel & Maaß (2016), or Hansen-Schirra et al. (2020), who propose a broader definition of translation (including intralingual, interlingual and intersemiotic translation).

Plain Language is flexible. Its linguistic features can be adapted to the presumed reading skills of the intended audience (Maaß 2020). Easy Language, on the other hand, is maximally reduced on all language levels (i.e. on the morphological, lexical, phrasal, syntactic and textual level) and therefore characterized by a more fixed set of linguistic rules (Bredel & Maaß 2016; Maaß 2020). It is mainly intended for people with communication impairments and disabilities (Bredel & Maaß 2016; Maaß 2020). In contrast to Easy Language, Plain Language has no stigmatising features but is also much less comprehensible. Therefore, situated between the

[1] For the source texts, see: https://www.apotheken-umschau.de (accessed 2025-03-14) and for the human gold standard translations, see https://www.apotheken-umschau.de/einfache-sprache/ (accessed 2025-03-14).

varieties Easy and Plain Language, Maaß (2020) models Easy Language Plus, which balances comprehensibility and acceptability.

In the area of natural language processing, translation into Plain Language can be linked to the studies on automatic text simplification. This area of research has focused on automatically converting complex lexical and syntactical constructions in a text into simpler ones (see, for instance, Saggion 2017; Martin et al. 2020; Sheang & Saggion 2021; Maddela et al. 2021, Ebling et al. 2022, Weiss & Meurers 2022, amongst others). Most existing approaches have focused on the differences based on language command or educational level. However, the differences between readers of Easy and Plain Language and further differentiation within the subgroups have not been addressed so far. At the same time, it is known that target audience-oriented data helps to build better automatic text simplification models (as stated by Scarton & Specia 2018). One method to perform text simplification is by using machine translation tools. Specific problems of automatic systems for intralingual translation have been described by Säuberli et al. (2020) and Spring et al. (2023). Overall, with the emergence of generative language models, there has been a considerable improvement in performance for the task of intralingual translation, e.g., translation from standard German into Easy or Plain German.

At the same time, despite the improvements in the intralingual MT systems, studies revealed that target groups are not yet safe to use the outputs (Anschütz et al. 2023; Maaß 2024b). Still, the tools can be used to provide professional translators with a draft that is post-edited in a next step. Further studies (Deilen et al. 2023; Deilen at al. 2024a, 2024b) also show that intralingual MT systems can be useful for professional translators to reduce their workload, but that they cannot be used safely without post-editing. Deilen at al. (2024a, 2024b) evaluated the performance of the MT system SUMM AI. SUMM AI is an intralingual translation tool that not only uses an LLM but also employs Easy and Plain Language parameters. The authors analysed the output of three SUMM AI models: baseline model, model 1 and model 2. The baseline model was a generically trained LLM while model 1 and model 2 used LLMs *and* were fine-tuned using the texts from *Apotheken Umschau* (see 3.1.). The baseline model and model 1 used the same LLM; however, unlike the baseline model, model 1 was also fine-tuned (Deilen et al. 2024b). Fine tuning was done with 170 texts from the *Apotheken Umschau* website. Both source texts and human gold standard texts were used to fine tune the LLM. Model 2 was fine-tuned in the same manner but is based on a different LLM. They compared the models' output with the existing human translations and the underlying source texts. Their study revealed that even though the fine tuning improved the output, the evaluation of the correctness revealed severe misinformation and content-related mistakes for all models (Deilen et al. 2024a: 47, 2024b: 473).

Similarly to Deilen et al. (2023), who tested ChatGPT for German Easy Language translations, Madina et al. (2024) investigated the feasibility of using ChatGPT to create Spanish Easy-To-Read (E2R) texts. Their study revealed that the generated output does not follow the E2R text rules (for example the rule regarding sentence splitting) and that the tool did not anticipate the knowledge of users, for example by adding examples and explanations when necessary. Therefore, they conclude that the output is still not suitable for the target audience, in this case users with cognitive disabilities.

2.2. Accessibility in the medical domain

Studies existing in the area of health literacy show that readers have trouble appraising the quality, trustworthiness, commercial interest, and applicability of health information to their own situation (Schaeffer et. al 2021: 69). The German *National Action Plan Health Literacy* (Schaeffer et al. 2018) suggests using Plain Language to make health information more accessible to large portions of the population. Health information in Plain Language may address population groups whose health literacy is measured to be particularly low, i.e., population groups with lower socio-economic status, lower education level, higher age, or migration experiences. People with chronic illnesses and disabilities, too, can benefit from health information in Plain Language, as they need to navigate the health system more extensively than the average person. Plain Language may alleviate some of the complexity in the health system. To translate more health texts into Plain Language, translators may leverage intralingual MT tools, such as those using LLMs, to speed up the translation process. This may make it feasible to meet the demand for accessible health information.

However, the use of LLM-based tools to translate health information into Plain Language is not without risks. Language models predict the likelihood of an utterance, but do not assess its accuracy in a given context. Weidinger et al. (2022) therefore identify the risk area "misinformation harms". Generating false information may threaten a person's autonomy, increase their trust in false beliefs, amplify distrust in "society's shared epistemology" (Weidinger et al. 2022: 218), and even cause bodily harm (Weidinger et al. 2022: 219), so that LLM-based tools are unsafe for users (Maaß 2024b).

Wilhelm et al. (2023) compare the health information output of four LLMs (for example GPT-3.5-Turbo) in terms of the criteria of completeness, correctness, and harmfulness. The models were asked "How to treat [disease]". The four models showed promise, for example in presenting balanced and unbiased information, but did not usually present the risks and potential harms of suggested treatments (completeness) and usually presented also incorrect and harmful information. The authors thus conclude that professional proof-reading is needed. GPT-3.5-Turbo performed best in terms of correctness and was the only MT-tool to not present any harmful health information.

2.3. Error analysis in the field of text simplification

Automatic text simplification aims to make given texts easier to read and understand while preserving its original meaning. Even though there is a large amount of work on text simplification, the evaluation of the simplification output still remains understudied (Grabar & Saggion 2022). Grabar & Saggion (2022) discuss the current state of the evaluation of automatic text simplification, including some crucial factors such as the role of end users, the domain of source documents, and the evaluation measures. They point out that one of the main challenges in the field of text simplification is the fact that the output is subjective, which makes it difficult to evaluate. The main reasons for this lack of objective measures are that there are no native simplified-language speakers, the guidelines for simplification are often vague, and the simplification requirements often vary depending on the target group of the text. With the rise of Large Language Models, more and more studies have been conducted that investigated the feasibility of using LLMs for text simplification and evaluation tasks in different languages. Most studies conclude that LLMs show great potential for text simplification, especially after

prompting and fine-tuning (see for example Klöser et al. 2024 for German; Martínez et al. 2024 for Spanish; or Nozza & Attanasio 2023 for Italian).

The only studies known to us that deal with intralingual (i.e. Plain Language) translation evaluation are Deilen et al. (2023) and Deilen et al. (2024a, 2024b). In Deilen et al. (2023), the researchers used ChatGPT-3.5 as a MT tool to translate administrative texts into German Easy Language. Two linguists trained in Easy and Plain Language manually checked the ChatGPT-output independently of one another. In case of discrepancies, the errors were discussed by six Easy Language experts. The authors followed a similar approach in Deilen et al. (2024a, 2024b). Here, the focus was on health communication in German Plain Language. Standard language texts from the *Apotheken Umschau* website were machine translated using the intralingual translation tool SUMM AI[2] and two researchers conducted the same manual error-check. When specialised knowledge of the medical domain was needed, Deilen et al. consulted the editorial team of the *Apotheken Umschau*. Deilen et al. (2023, 2024a, 2024b) report only whether the outputs were correct or incorrect. If one content-related error was found, the output was considered incorrect. This was done to first evaluate the suitability of machine-translated Plain Language output for end users. In health communication, texts are unsafe if they contain any content-related mistakes as health literacy is low – especially considering the dimension *appraisal*. From the baseline model, 29 translations were incorrect; model 1 performed similarly, having translated only two out of 30 texts correctly; model 2 had the best performance with 15 correct translations (Deilen et al. 2024b). None of the ChatGPT-4o translations were correct. The authors do not provide any information on error categories nor error severity.

Established schemes for error analysis in interlingual (machine) translation are widely used (see, e.g., multidimensional quality metrics – MQM, Lommel et al. 2014). They normally describe error typology with various linguistic categories, as well as error scoring and severity levels. However, to the best of our knowledge, an error evaluation scheme for intralingual translation is still missing. We chose the translation error framework MQM. Relevant criteria are chosen from this framework to assess the translation quality in a given project. For example, using the harmonised Dynamic Quality Framework and Multidimensional Quality Metrics (DQF-MQM), Rodríguez Vázquez et al. (2022) compared three interlingual MT tools for various language pairs. Easy Language texts regarding the legal, medical, and political domain were translated from French to German, English, Spanish, and Farsi. The authors compared adherence to Easy Language rules and translation quality according to DQF-MQM. The translations from French into German Easy Language most notably contained translation errors regarding linguistic conventions (like spelling, punctuation, or grammar) and accuracy (like mistranslations, over-, or under-translation, Rodríguez Vázquez et al. 2022). Translation of health information texts, too, contained most notably errors in linguistic conventions, but also style errors (i.e. awkward, unidiomatic, or inconsistent style, Rodríguez Vázquez et al. 2022). Concerning Easy Language rules, German texts contained most rule violations on the word level (for example choice of words and numbers) and the rule to segment compounds was not implemented at all (Rodríguez Vázquez et al. 2022: 40). In all domains, most rules on the word level contained errors. In health information texts, errors on the phrase level (for example negation and use of present tense) were also common (Rodríguez Vázquez et al. 2022: 39). The

[2] summ-ai.com. (accessed 2025-03-14). SUMM AI is a machine translation tool for translating texts into Easy German and Plain German. The company SUMM AI offers different licenses for freelancers, authorities and companies.

authors conclude that their results can be used to devise rules for the pre- and post-editing processes that aim to solve the most prominent accessibility issues in the MT-generated texts. Other studies utilise LLMs to automatically apply MQM for translation evaluation (Fernandes et al. 2023; Kocmi & Federmann 2023). Since application of MT for Easy and Plain Language is growing, there is a need for studies and frameworks focusing on error typology in intralingual translation.

3. Research design

3.1. Data collection

We use the dataset from Deilen et al. (2024b), which is based on the 30 texts from the *Apotheken Umschau* website. The texts cover various topics ranging from long COVID, diabetes to birth control. The 30 texts were chosen randomly out of the 200 texts for which a Plain Language translation was available, with the only condition that they cover as many different categories as possible (medication, diseases, therapies, first aid, contraception). For each of these 30 texts, the dataset contains a professional human translation into Plain Language and three machine-translated outputs produced with three different SUMM AI models (see 2.3.). The human translations were produced by Plain Language experts who were translation students at the University of Hildesheim and who professionally translated health information at the Research Centre for Easy Language (University of Hildesheim). The human translations underwent a three-stage proof-reading process. The rough translations were proofread by a peer. The translator then improved the translation according to the feedback. A project manager proofread the translation again before sending it to the *Apotheken Umschau* editorial team that ensured the correctness of the content. If the translation contained any mistakes, the translations were again improved by the project manager. Consequently, these translations can be considered as reference, i.e., a human gold standard.

In our study, we add a ChatGPT-4o corpus and compare the machine translations (four for each text) with the existing human translations and the source texts. The main difference between ChatGPT and SUMM AI is that the latter was specifically developed for Easy and Plain German translation (see above), while ChatGPT is a generic tool. In addition, ChatGPT is a chatbot that can be prompted whereas SUMM AI does not contain any prompting functions. Translations with ChatGPT were generated after a two-step prompt. The prompt was developed based on findings from Deilen et al. (2023), stating that assigning a role, setting a context and asking for background information seemed to improve the output. First, ChatGPT was asked to define *Plain Language* in the field of accessible communication and inclusion: “ChatGPT, wie wird im Bereich der Barrierefreien Kommunikation und der Inklusion ‘Einfache Sprache’ definiert?” (‘ChatGPT, how is *Plain Language* defined in the field of Accessible Communication and inclusion?’). Then, ChatGPT was prompted to take on the role of a Plain Language expert:

> ChatGPT, du bist jetzt ein Experte für Einfache Sprache. Wir brauchen Unterstützung in der Übersetzung eines Textes der Gesundheitskommunikation in Einfache Sprache. In dem Text geht es um [hier Thema einfügen]. Die Übersetzung soll für Menschen mit Leseschwierigkeiten leicht verständlich sein. Übersetze bitte den folgenden Text:

> 'ChatGPT, you are now an expert in Plain Language. We need help translating a health communication text into Plain Language. The text is about [insert topic]. The translation should be easily comprehensible for people who have difficulty reading[3]. Please translate the following text:'.

As a result, for each source text we have one human gold standard and four machine translations produced with four different models that are labelled as baseline, model 1, model 2 and ChatGPT in the analysis below[4].

3.2. Data analysis

We follow Deilen et al. (2023, 2024a, 2024b) to evaluate the machine translated outputs in terms of readability, syntactic complexity, and correctness. Easy Language and Simple Language have specific characteristics at all linguistic levels. We will limit ourselves here to the example of syntactic complexity because it provides a good indication of the profiles of the target texts. We add to the results reported in Deilen et al. (2024b) by comparing them with the ChatGPT-4o corpus. We will also add an analysis of error typology, not only for the ChatGPT-4o outputs, but also for the SUMM AI outputs. This analysis is described in section 3.2.2., and preliminary results from the current stage of error analysis are presented in 3.3.3.

3.2.1. Readability

We automatically evaluated readability using the Hohenheim Comprehensibility Index (HIX). The HIX, which is commonly used in German Easy Language Research, takes into account the four major readability formulas validated for the German language (Bredel & Maaß 2016: 61–62): the Amstad index, the simple measure of gobbledygook (G-SMOG) index, the Vienna non-fictional text formula (W-STX), and the readability index (LIX). A HIX of 0 indicates extremely low comprehensibility and a HIX of 20 means extremely high comprehensibility[5]. At a HIX of 18, a text can be classified as Easy German, i.e. the least complex variety of German (Rink 2020; Maaß 2024). The benchmark for Plain German is at 16 points (Kröger et al. 2025).

3.2.2. Syntactic complexity

We use dependency parsing (using the Stanford NLP Python Library Stanza (v1.2.1)[6]) to automatically measure the distribution of specific syntactic relations with the goal to assess syntactic complexity. Syntactic relations 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). The higher the number of these syntactic relations in the corpus, the more complex the texts. The distribution frequencies of these syntactic relations are calculated for each corpus: source texts, human translations, and the four machine translated outputs.

[3] We used "people who have difficulty reading" as an umbrella term for the different Plain Language target groups (see Maaß 2020).

[4] The labels for the SUMM AI systems were taken from Deilen et al. (2024b).

[5] see https://klartext.uni-hohenheim.de/hix (accessed 2025-03-14)

[6] https://stanfordnlp.github.io/stanza/index.html (accessed 2025-03-14)

3.2.3. Correctness and error typology

Correctness as a criterion for automatic intralingual Easy and Plain Language translation was reported in Deilen et al. (2024a, 2024b). In Plain Language translation, a text can be classified as correct if it does not contain any errors, neither in terms of content nor in terms of language (i.e. spelling, grammar). Common errors in the corpus of Deilen et al. (2024a, 2024b) were redundancies, unreasonable and incorrect statements, lexico-semantic errors, missing segmentation signs, as well as missing reflective pronouns. However, in Deilen et al. (2024a, 2024b), the researchers simply mentioned the errors they encountered but they did not employ a structured error analysis, as is done in this study.

To check the correctness in a more structured manner that allows for a later quantitative error analysis, we first adapted and chose between relevant parts of the MQM to fit the present intralingual translation study. We then analysed the four main error categories: terminology, accuracy, linguistic conventions, and audience appropriateness. Table 1 presents and defines these four categories (in bold letters) and its error types in more detail.

As stated before, the error typology is based on the categories proposed by MQM, expanded and adapted with error types fitting intralingual translation analysis. For instance, the typology "wrong or missing explanation" was added under accuracy: if an explanation was present and relevant but did not reflect the information from the source text, it constitutes a "wrong addition". If an explanation was needed, but not present, it was a "missing addition". The adapted MQM is used as a coding frame to manually analyse the translation errors within the MT output.

An initial analysis using the coding frame focused on the ChatGPT-4o output. Here, researchers compared the source texts with the ChatGPT-4o output. They annotated any errors according to the coding frame. This initial analysis was useful to evaluate which parts of the coding frame were usable and which parts needed further adaptation. Definitions for each error category were specified to fit the present corpora, and examples were added to make the coding frame more usable. The initial analysis was also needed to train the researchers in error analysis. At this stage, errors were documented thoroughly in each text and the number of errors was added up according to the four main categories of terminology, accuracy, linguistic conventions, and audience appropriateness (see Section 3.3.3.).

In a next step, the improved coding frame was used to again analyse three texts in the ChatGPT-4o subcorpus and to analyse three texts in the model 1 subcorpus. Again, the MT output was compared with the source texts, and any errors were annotated according to the coding frame. The researchers who had previously analysed the ChatGPT-4o output now analysed the model 1 output. The researchers discussed cases in which they disagreed. Due to time constraints, this was completed for one of the model 1 texts and for all three ChatGPT-4o texts which are much shorter than the SUMM AI generated texts. After this process, the coding frame was again adapted: definitions were written more clearly and more distinctly from each other (see Table 1), examples were corrected, clarified, or added. Errors were again documented thoroughly. After this second step, the coding frame was finalised and is currently used to analyse all MT outputs.

Table 1: Overview of the MQM error categories from the current project stage[7]

Error type	Definition
Terminology	**The use of a term does not fit to the field conventions, is incorrectly used in the target text or is not equivalent to the term in the source text.**
Inconsistent terminology	Multiple terms are used to describe the same concept when just one term is needed or appropriate.
Wrong term	Multiple terms are used to describe the same concept when just one term is needed or appropriate.
Accuracy	**Content in the target text does not match the propositions from the source text.**
Mistranslations	Target content does not accurately represent the source content.
Ambiguous content	Ambiguity is introduced where specification is needed.
Hallucination	Machine translation produces an output that is totally decoupled from the source text.
Wrong or missing explanation	Explanation is necessary and added but does not represent the information from the source text (*wrong*) or an explanation is needed but is not present in the target text (*missing*).
Incomplete information	Relevant information from the source text is missing in the target text.
Linguistic conventions	**Errors related to the linguistic level of the source text.**
Grammar	Grammatical rules are violated in the source text.
Punctuation	Punctuation is used incorrectly.
Spelling	Words are misspelled.
Cohesion and coherence	Connectors necessary to understand the text as a whole are missing or incorrect (*cohesion*). Semantic relationships within the text are not clear (*coherence*)
Audience appropriateness	**Content in the target text is not valid, appropriate or acceptable for the target audiences.**
Inaccurate advice	Target text contains advice that is not in the source text or that is not suitable for the situation in question.
Stigmatising content	Content can lead to stigmatization of end users.

[7] Current project stage refers to the fact that we are currently working on adapting the MQM error categories to the field of intralingual translation and that these are the categories used for the presented analyses. However, this work can be regarded as work in progress, i.e. while analysing the errors, we also adapt the scheme accordingly.

3.3. Results

3.3.1. Readability

Figure 1 illustrates the HIX values for each corpus. As expected, the source text corpus scored the lowest HIX values of the six corpora (mean: 10.46, SD: 2.76). Model 2 achieved the highest HIX values, with a mean HIX value of 19.5 (SD: 0.76). The ChatGPT-4o texts yielded the lowest HIX value among all machine translations (mean: 16.1, SD: 3.26). Figure 1 also shows that the ChatGPT translations had a much greater variation in the HIX values than all other texts.

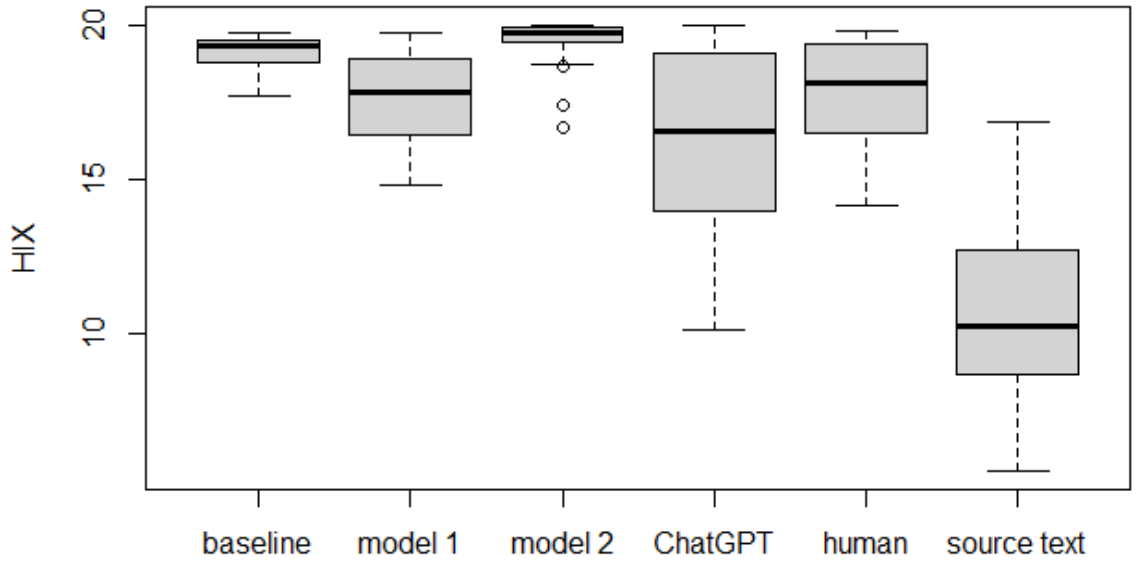

Figure 1: HIX values of the four machine translations, the human translations, and the source texts

All texts in the baseline corpus and in the model 2 corpus could be classified as Plain German – meaning all texts reached a HIX value of at least 16 points. 93% of the model 1 corpus and 83% of the human translation corpus reached this benchmark. Only 57% of texts in the ChatGPT-4o corpus could be classified as Plain German. Thus, in terms of readability, ChatGPT-4o performed worse than the other MT systems.

Still, it is important to keep in mind that the HIX value is only a quantitative measure that focuses purely on readability features on the text surface (i.e. overt complexity). This means that the textual level and aspects like cohesion, coherence, or information structure are not assessed. Consequently, HIX values can only be regarded as a first step in the analysis and have to be supplemented with further qualitative analysis. For example, how is the complexity of the text contents reflected in the very homogeneously high HIX values of model 2?

3.3.2. Syntactic complexity

Figure 2 shows the distributions of complex dependency relations in all the subcorpora under analysis. Similarly to the results reported by Deilen et al. (2024b), none of the machine-translated outputs is syntactically as simple as the human translations.

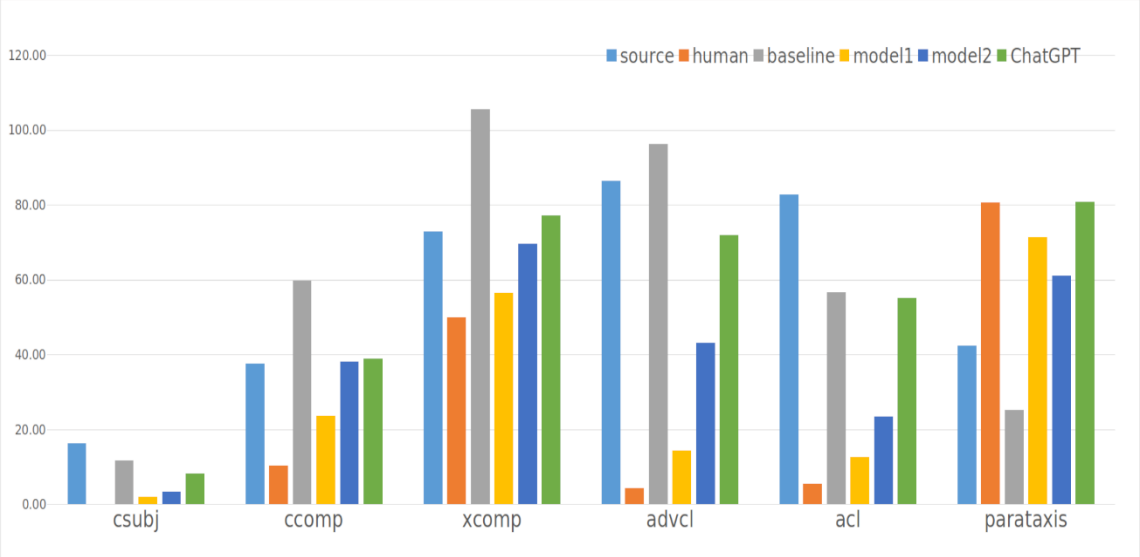

Figure 2: Distribution of syntactically complex dependency relations in the source texts, human and the four machine translations (normalised frequencies per 10000)

The best performing system in terms of syntactic simplicity was model 1 with the only exception of the parataxis relation. This type of relation was automatically assigned to sentential parenthetical or a clause after a colon or a semicolon placed side by side without any explicit coordination or subordination. This type of construction, especially the one with colon, is frequently used in Plain German, as shown in the human translation in example (1): *Sie glauben* ('You think') is followed by a colon and the verb *haben* ('have') is tagged as a parataxis relation to the verb *glauben* ('think') by the parser.

(1) **Sie glauben**: Ich habe eine seelische Erkrankung? ('**You think**: I have a mental illness?')

This demonstrates the appropriateness of the use of this relation in translations into Plain Language and also explains why we observe an opposite tendency here if compared to other syntactically complex relations. Since syntactic embedding depth in particular poses a risk to comprehension, parataxis in Plain Language texts to a moderate extent does not pose a problem. Accordingly, it also appears frequently in the human-translated gold standard texts, which supports once more that parataxes are not difficult to process and can therefore be used in Plain Language texts. Interestingly, translations with ChatGPT-4o were similar to human translations in terms of the use of parataxis relations.

However, this is not the case for the other syntactic relations. The ChatGPT-4o outcomes are positioned between the performance by baseline and the models 1 and 2, with the exception of complement clauses, where ChatGPT-4o achieved a similar performance as the most successful model 2. An example of a complement clause is illustrated in example (2).

(2) a. Ein gesundes Gelenk hat eine Schicht aus Knorpel und Gelenkflüssigkeit. Diese sorgen dafür, dass die Knochen nicht aneinander reiben. (ChatGPT-4o)

'A healthy joint has a layer of cartilage and synovial fluid. These ensure that the bones do not rub against each other.'

b. Zwischen den Knochen sind Knorpel und Gelenkflüssigkeit. Sie dienen dem Schutz der Knochen: So reiben die Knochen nicht aneinander. (Human)

'Between the bones are cartilages and synovial fluid. They serve to protect the bones: This way the bones do not rub against each other.'

The translation with ChatGPT-4o in (2a) contains the complement clause introduced with the complementizer *dass* ('that'). The human gold standard translation in (2b) contains a parataxis construction with a colon instead.

3.3.3. Correctness and Error Typology

The initial MQM-based error analysis of the ChatGPT-4o translation corpus shows several issues along all four main categories (see section 3.2.2.), indicating that most of the texts have to be professionally post-edited to be functional and action-oriented. In this first analysis, a total amount of 244 errors were found in 30 ChatGPT-4o generated texts, averaging 8.13 errors per text (SD: 7.3; mode: 2). Most errors were found in the main categories *accuracy* (110 errors) and *linguistic conventions* (109 errors). A "mistranslation" is illustrated in the following example from our MQM coding frame:

(3) a. Zudem gibt es eine Früherkennungsuntersuchung, die ab einem Alter von 35 Jahren alle zwei Jahre von den Krankenkassen bezahlt wird. (source text)

'Additionally, there is a screening that from age 35 onward is paid by health insurance.'

b. Ab 35 Jahren kannst du alle zwei Jahre eine Hautuntersuchung bei deiner Krankenkasse machen lassen. (ChatGPT-4o)

'From the age of 35 onward, you can get a skin examination at your health insurance.' (using the informal *you*)

The correctness and error typology analysis indicates the utmost importance of post-editing by professional translators or domain experts. Classifying error severity and thus also the resulting post-editing workload remains an open question, which will be addressed in future phases of this research project.

4. Summary and future work

ChatGPT-4o generated texts displayed the lowest and the most varied HIX values. The three SUMM AI models were better able to produce readable text that reached the Plain Language benchmark more often. While ChatGPT-4o did not perform as well as model 1 in terms of syntactic complexity, it resembled human translation in its use of parataxis. In other syntactic relations, it mostly performed better than the baseline model, but worse than models 1 and 2. All 30 ChatGPT-4o generated texts contained errors.

The SUMM AI models outperform ChatGPT-4o in almost all analysed criteria. This suggests that fine-tuning both with task-specific data (baseline system), meaning the task of intralingual translation into Plain Language, and with domain-specific data (health texts from *Apotheken Umschau*), improves MT into Plain German. Fine-tuned models that are trained specifically for intralingual translation tasks outperform the general model that uses prompting. In the case of ChatGPT-4o, we did not use all the prompting strategies proposed in

Deilen et al (2023): In addition to the holistic model that was used in this paper, they chose a second approach that they called "linguistic". This approach gives ChatGPT the order to simplify subsequently at different linguistic levels: explain terms, reduce syntactic complexity and similar. In Deilen et al. (2023), the holistic approach performed better, which is why it was favoured in the current study. However, it would be interesting to include this few-shot prompting approach again in the next steps of the study design. It is similar to the fine-tuning used for the different models of SUMM AI. Further insights into the successful use of prompt models could be expected here.

Preliminary results show that, similar to the results by Rodríguez Vázquez et al. (2022) who evaluated interlingual translations of Easy Language texts, most ChatGPT-4o-generated errors appear in the categories *accuracy* and *linguistic conventions*. In the current project stage, we once more apply the error analysis to the ChatGPT-4o corpus, but also to the SUMM AI models (baseline, model 1, model 2). Furthermore, following MQM settings and recommendations (The MQM Council), we aim to assess error severity. This will give us an insight into the usability of the examined MT tools in intralingual translation processes and will allow us to assess the time and effort that are needed for post-editing the output.

References

Ahrens, Sarah. 2025. *Einfache Sprache in der Gesundheitskommunikation. Patientinnenaufklärung für Frauen mit Deutsch als Zweitsprache.* Berlin: Frank & Timme.

Anschütz, Miriam & Oehms, Joshua & Wimmer, Thomas & Jezierski, Bartłomiej & Groh, Georg. 2023. Language models for German text simplification. Overcoming parallel data scarcity through style-specific pre-training. arXiv preprint arXiv:2305.12908. doi: 10.18653/v1/2023.findings-acl.74.

Bredel, Ursula & Maaß, Christiane. 2016. *Leichte Sprache. Theoretische Grundlagen, Orientierung für die Praxis*. Berlin: Duden.

Deilen, Silvana & Hernández Garrido, Sergio & Lapshinova-Koltunski, Ekaterina & Maaß, Christiane. 2023. Using ChatGPT as a CAT tool in Easy Language translation. arXiv preprint arXiv:2308.11563.

Deilen, Silvana & Lapshinova-Koltunski, Ekaterina & Hernández Garrido, Sergio & Maaß, Christiane & Hörner, Julian & Theel, Vanessa & Ziemer, Sophie. 2024a. 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*. 44–53.

Deilen, Silvana & Lapshinova-Koltunski, Ekaterina & Hernández Garrido, Sergio & Hörner, Julian & Maaß, Christiane & Theel, Vanessa & Ziemer, Sophie. 2024b. Evaluation of intralingual machine translation for health communication. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation* (Vol. 1). European Association for Machine Translation. 467–77.

Ebling, Sarah & Battisti, Alessia & Kostrzewa, Marek & Pfütze, Dominik & Rios, Annette & Säuberli, Andreas & Spring, Nicolas. 2022. Automatic Text Simplification for German. *Frontiers in Communication* 7: 706718. doi: 10.3389/fcomm.2022.706718.

Grabar, Natalia & Saggion, Horacio. 2022. Evaluation of automatic text simplification: Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*. Avignon: ATALA. 453–63.

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

Jakobson, Roman. 1959. On Linguistic Aspects of Translation. In Brower, Reuben Arthur (ed.), *On Translation*. Cambridge, Mass.: Harvard University Press. 233–39.

Kocmi, Tom & Federmann, Christian. 2023. GEMBA-MQM. Detecting Translation Quality Error Spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*. Singapore. Association for Computational Linguistics. 768–75. doi: 10.18653/v1/2023.wmt-1.64.

Klöser, Lars & Beele, Mika & Schagen, Jan-Niklas & Kraft, Bodo. 2024. German Text Simplification: Finetuning Large Language Models with Semi-Synthetic Data. arXiv preprint arXiv:2402.10675.

Kröger, Janina & Deilen, Silvana & Hörner, Julian & Lapshinova-Koltunski, Ekaterina & Maaß, Christiane. 2025. Proof of Concept. Entwicklung eines redaktionellen Workflows für die KI-gestützte Übersetzung von Gesundheitsinformationen in Einfache Sprache. *TransKom* 18(1): 380–404.

Lommel, Arle Richard & Burchardt, Aljoscha & Uszkoreit, Hans. 2014. Multidimensional Quality Metrics (MQM). A Framework for declaring and describing Translation Quality Metrics. *Tradumàtica tecnologies de la traducción* 12: 455–63. doi: 10.5565/rev/tradumatica.77.

Maaß, Christiane. 2020. *Easy language – Plain language – Easy language plus: Balancing comprehensibility and acceptability*. Berlin: Frank & Timme.

Maaß, Christiane. 2024a. Intralingual translation in easy language and plain language. In Pilliére, Linda & Berk Albachten, Özlem (eds.), *Routledge Handbook of Intralingual Translation*. London: Routledge. 234–51.

Maaß, Christiane. 2024b. Hi ChatGPT, Translate this text into Easy Language. Is the new Easy Language translator a machine? *VAKKI Publications* 16: 9–29. doi: 10.70484/vakki.145621.

Maddela, Mounica & Alva-Manchego, Fernando & Xu, Wei. 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,* Online. Association for Computational Linguistics. 3536–53. doi: 10.18653/v1/2021.naacl-main.277.

Madina, Margot & Gonzalez-Dios, Itziar & Siegel, Melanie. 2024. A Preliminary Study of ChatGPT for Spanish E2R Text Adaptation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024*). Torino, Italia: ELRA and ICCL. 1422–34.

Martin, Louis & de la Clergerie, Éric & Sagot, Benoît & Bordes, Antoine. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France. European Language Resources Association. 4689–98.

Martínez, Paloma & Ramos, Alberto & Moreno, Lourdes. 2024. Exploring Large Language Models to generate Easy to Read content. *Frontiers in Computer Science* 6: 1394705. doi: 10.3389/fcomp.2024.1394705.

Nozza, Debora & Attanasio, Giuseppe. 2023. Is it really that simple? Prompting language models for automatic text simplification in Italian. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023).* Venice, Italy: CEUR Workshop Proceedings. 322–33.

Rink, Isabel. 2020. *Rechtskommunikation und Barrierefreiheit. Zur Übersetzung juristischer Informations- und Interaktionstexte in Leichte Sprache.* Berlin: Frank & Timme.

Rodríguez Vázquez, Silvia & Kaplan, Abigail & Bouillon, Pierrette & Griebel, Cornelia & Azari, Razieh. 2022. La traduction automatique des textes faciles à lire et à comprendre (falc) : une étude comparative. *Meta* 67(1): 18–49. doi: 10.7202/1092189ar.

Saggion, Horacio. 2017. Applications of automatic text simplification. In Saggion, Horacio (ed.), *Automatic Text Simplification*. Cham: Springer. 71–77. doi: 10.1007/978-3-031-02166-4_7.

Säuberli, Andreas & Ebling, Sarah & Volk, Martin. 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). Marseille, France. European Language Resources Association. 41–48.

Scarton, Carolina & Specia, Lucia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Volume 2: Short Papers.* Melbourne, Australia. Association for Computational Linguistics. 712–18. doi: 10.18653/v1/P18-2113.

Schaeffer, Doris & Berens, Eva-Maria & Gille, Svea & Griese, Lennert & Klinger, Julia & de Sombre, Steffen & Vogt, Dominique & Hurrelmann, Klaus. 2021. *Gesundheitskompetenz der Bevölkerung in Deutschland vor und während der Corona-Pandemie: Ergebnisse des HLS-GER 2.* (Technical report.) Bielefeld: Interdisziplinäres Zentrum für Gesundheitskompetenz-forschung (IZGK), Universität Bielefeld. doi: 10.4119/unibi/2950305.

Schaeffer, Doris & Hurrelmann, Klaus & Bauer, Ullrich & Kolpatzik, Kai. 2018. National Action Plan Health Literacy: Promoting health literacy in Germany. Berlin: KomPart, 10: 0418–1866.

Schubert, Klaus. 2013. Bürgernahe Sprache. Überlegungen aus fachkommunikationswissenschaftlicher Sicht. *Synaps* 29: 48–57.

Sheang, Kim Cheng & Saggion, Horacio. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*. Aberdeen, Scotland, UK. Association for Computational Linguistics. 341–52. doi: 10.18653/v1/2021.inlg-1.38.

Spring, Nicolas & Kostrzewa, Marek & Fröhlich, David & Rios, Annette & Pfütze, Dominik & Battisti, Alessia & Ebling, Sarah. 2023. Analyzing sentence alignment for automatic simplification of German texts. In Deilen, Silvana & Hansen-Schirra, Silvia & Hernández Garrido, Sergio & Maaß, Christiane & Tardel, Anke (eds.), *Emerging Fields in Easy Language and Accessible Communication Research*. Berlin: Frank & Timme. 339–69. doi: 10.57088/978-3-7329-9026-9_1.

The MQM Council. n.d. The MQM Error Typology. (https://themqm.org/error-types-2/typology/) (Accessed 2024-08-07).

Weidinger, Laura & Uesato, Jonathan & Rauh, Maribeth & Griffin, Conor & Huang, Po-Sen & Mellor, John & Glaese, Amelia & Cheng, Myra & Balle, Borja & Kasirzadeh, Atoosa & Biles, Courtney & Brown, Sasha & Kenton, Zac & Hawkins, Will & Stepleton, Tom & Birhane, Abeba & Hendricks, Lisa Anne & Rimell, Laura & Isaac, William & Haas, Julia & Legassick, Sean & Irving, Geoffrey & Gabriel, Iason. 2022. Taxonomy of risks posed by language models. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*. Seoul, Republic of Korea. 214–29. doi: 10.1145/3531146.3533088.

Weiss, Zarah & Meurers, Detmar. 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). 141–53. doi: 10.18653/v1/2022.bea-1.19.

Wilhelm, Theresa I. & Roos, Jonas & Kaczmarczyk, Robert. 2023. Large language models for therapy recommendations across 3 clinical specialties. Comparative study. *Journal of Medical Internet Research* 25: 1–13. doi: 10.2196/49324.

Sarah Ahrens
University of Hildesheim
Institut für Übersetzungswissenschaft und Fachkommunikation
Universitätsplatz 1
31141 Hildesheim
Germany
e-mail: ahrenss@uni-hildesheim.de

Silvana Deilen
University of Hildesheim
Institut für Übersetzungswissenschaft und Fachkommunikation
Universitätsplatz 1
31141 Hildesheim
Germany
e-mail: deilen@uni-hildesheim.de

Sergio Hernández Garrido
University of Hildesheim
Institut für Übersetzungswissenschaft und Fachkommunikation
Universitätsplatz 1
31141 Hildesheim
Germany
e-mail: hernandezs@uni-hildesheim.de

Ekaterina Lapshinova-Koltunski
University of Hildesheim
Institut für Übersetzungswissenschaft und Fachkommunikation
Universitätsplatz 1
D-31141 Hildesheim
Germany
e-mail: lapshinovakoltun@uni-hildesheim.de

Christiane Maaß
University of Hildesheim
Institut für Übersetzungswissenschaft und Fachkommunikation
Universitätsplatz 1
D-31141 Hildesheim
Germany
e-mail: maassc@uni-hildesheim.de