

Evaluating LLM-Based Text Simplification for German: Effects on Post-Editing Effort, Quality Ratings, and User Comprehension

Luisa Carrer¹, Andreas Säuberli², Martin Kappus¹, Lukas Fischer³, Sarah Ebling³

¹School of Applied Linguistics, ZHAW Zurich University of Applied Sciences

²MaiNLP, Center for Information and Language Processing, LMU Munich

³Department of Computational Linguistics, University of Zurich

luisa.carrer@zhaw.ch, ebling@cl.uzh.ch

Abstract

Automatic text simplification (ATS) seeks to automate the process of rewording within the same language to enhance readability and comprehension. Current evaluation practices for ATS systems predominantly rely on automatic metrics or assessments by experts and crowdworkers, often excluding the intended end users and other stakeholders, and thus limiting insights into the actual effectiveness of ATS models. In this study, we address this gap by conducting a multi-faceted, mixed-method evaluation of two LLM-based ATS systems for German (capito.ai and GPT-4o) and by involving end users, post-editors, and Easy Language experts. The findings highlight the effectiveness of the LLM-based ATS systems examined across several dimensions, including post-editing efficiency, expert quality assessments, and, in the case of GPT-4o-generated simplifications, user comprehension. Post-editing effort metrics, in particular, show an increase in productivity of around 30% compared to full manual simplification. Moreover, the results reveal substantial differences in perception and understanding among participant groups. These outcomes clearly indicate that ATS for German has recently made considerable progress and, crucially, underscore the importance of incorporating multiple stakeholders into ATS evaluation to better align system performance with accessibility goals.

Keywords: automatic text simplification, large language models, Easy Language, post-editing, human evaluation, reading comprehension

1. Introduction

Easy Language is a simplified language variety that is designed to optimize comprehensibility, particularly for persons with limited reading skills. Primary target groups include persons with cognitive impairments and learning disabilities, prelingually deaf persons, and second language learners, among others (Maaß, 2020).

The automatic translation of standard language into simplified language varieties (automatic text simplification, ATS) has been a topic of increasing interest among natural language processing (NLP) researchers (Al-Thanyyan and Azmi, 2021; North et al., 2025). However, the evaluation of simplified texts poses a significant challenge. In particular, the limitations of automatic evaluation metrics have been discussed (Alva-Manchego et al., 2021; Grabar and Saggion, 2022; Madina et al., 2024; Cripwell et al., 2024). Human evaluation is no less challenging, especially when Easy Language is the target variety, where the end users are very diverse and often difficult to recruit. Nevertheless, the importance of including the primary target groups in such evaluations has been repeatedly stressed (Maaß, 2020; Štajner, 2021; Gooding, 2022; Carrer et al., 2024; Madina et al., 2024).

In addition to end users, a realistic use case of Easy Language translation also involves other stakeholders, including post-editors who manually

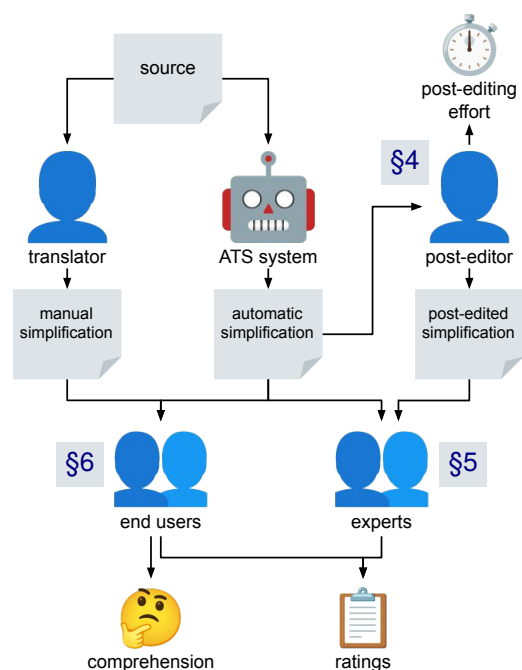


Figure 1: Overview of our study. Source texts are simplified through manual translation, automatic simplification, and post-editing. We measure post-editing effort, test end-user comprehension, and collect quality ratings to determine the effectiveness of ATS.

correct automatically simplified texts and experts who are familiar with the language-specific guidelines for Easy Language (Carrer et al., 2024). In NLP research, however, these additional stakeholders are often ignored (cf. Al-Thanyyan and Azmi, 2021). It is therefore problematic to get an accurate picture of the actual effectiveness of ATS models, especially for less-researched languages such as German.

Against this backdrop, (Carrer et al., 2024) report that no measurable productivity gains for post-editors emerged when working with neural ATS systems, raising fundamental questions about the added value of such systems in real-world settings. At the same time, the latest generation of transformer-based language models has shown substantial improvements in text simplification adequacy, including at the document level (Qiang et al., 2025; Kew et al., 2023; Beks van Raaij et al., 2024). These advances suggest that earlier conclusions may no longer fully reflect the current technological landscape. It remains unclear, however, whether these technical improvements translate into tangible benefits in authentic text simplification production processes—particularly with regard to post-editing effort, workflow efficiency, and overall effectiveness across stakeholders. This points to a methodological gap: we lack robust, multi-stakeholder, document-level evaluations that capture real-world text simplification scenarios (cf. Reiter, 2025). In this paper, we address this gap by conducting what to our knowledge is the first multi-stakeholder evaluation of LLM-based ATS for German.

Our central research question is: **How effective are latest-generation LLM-based systems in real-world document-level ATS scenarios?** Specifically, our contributions are (also visualized in Figure 1):

1. a post-editing evaluation, where we compare the effectiveness of semi-automatic simplification (i.e., ATS followed by post-editing) with that of full manual simplification (Section 4);
2. an expert evaluation, where the simplified texts are judged by professional Easy Language service providers (Section 5);
3. an end-user evaluation, where we measure end users' perception and actual comprehension of both original (or source) and simplified (or target) texts (Section 6).

2. Related work

2.1. Human Evaluation of Text Simplification

The importance of human evaluation has long been emphasized in ATS research (Alva-Manchego et al., 2020; Štajner, 2021; Gooding, 2022). However, there is no consensus on best practices, resulting in a wide range of methods, including ratings of simplicity, fluency, and meaning preservation (Štajner and Nisioi, 2018; Mallinson et al., 2020; Stodden, 2021; Martin et al., 2022; Ryan et al., 2023), comprehension testing (Leroy et al., 2013, 2022; Fajardo et al., 2014; Alonzo et al., 2021; Agrawal and Carpuat, 2024; Beks van Raaij et al., 2024; Säuberli et al., 2024; Guidroz et al., 2025), behavioural measurements (Rello et al., 2013a,b,c; Crossley et al., 2014; Saggion et al., 2015; Gooding et al., 2021), and qualitative approaches (Martínez et al., 2024). Some studies include persons with disabilities as end users (e.g., Huenerfauth et al., 2009; Saggion et al., 2015; Alonzo et al., 2021), while others often rely on crowdworkers, researchers, or students (cf. Stodden, 2026, pp. 124-130).

ATS research for German has lagged behind English, in part due to a lack of training data (Ebling et al., 2022). Recently, the few-shot and zero-shot capabilities of LLMs have enabled substantial improvements in output quality. This has prompted more research evaluating these technologies for German ATS—particularly in the area of Easy Language and Plain Language translation: Deilen et al. (2023) evaluated different prompting strategies for generating German Easy Language texts with ChatGPT, Deilen et al. (2024) conducted an expert evaluation of the commercial ATS system SUMM-AI on Plain Language translation, and Ahrens et al. (2025) compared Plain Language translation errors between two systems. However, the effectiveness of German LLM-based ATS for end users and post-editors has not been evaluated yet. This is the research gap that we are addressing in the present work.

Recently, Carrer et al. (2024) proposed a multidimensional approach to evaluating text simplification, covering comprehension testing, ratings, and post-editing efficiency, and involving end users, post-editors, and experts. We apply and extend their approach to evaluate and compare two systems based on large language models (LLMs) for German text simplification.

2.2. Evaluation of Post-editing Effort

Post-editing has recently been identified as the strongest growth target among language service providers in Europe (ELIS Research, 2025). With its widespread use in interlingual machine trans-

lation, research has increasingly focused on how translators engage in this task and what is the level of effort involved (cf., among others, Krings, 2001; Koponen, 2016; Scarton et al., 2019; Cumbreño and Aranberri, 2021; Alvarez-Vidal and Oliver, 2023). It has been widely recognized that post-editing effort encompasses three highly interconnected dimensions: temporal, technical, and cognitive effort (Krings, 2001). Temporal effort (i.e., time spent on post-editing) is easily quantifiable and directly influences productivity. Technical effort pertains to the editing actions taken by the editor (e.g., text productions, text eliminations, replacements, and shifts) and is often captured using keylogging data. Cognitive effort refers to the mental processes underlying post-editing, e.g., detecting errors and planning the necessary corrections. Measuring cognitive effort is challenging due to its complexity (cf. Vieira, 2017), but pauses have emerged as indicative of cognitive load. Lacruz et al. (2012, 2014) proposed measuring clusters of short pauses, which revealed a clear correlation with post-editing effort, and established the pause threshold at 300 ms. This study is one of the first to specifically measure post-editing effort for text simplification, with few known comparable attempts (cf. Carrer et al., 2024).

3. Evaluated systems and texts

Our evaluation includes two LLM-based systems:

- **capito.ai**¹ is a commercial ATS system that offers automatic simplifications at three levels, aligned with levels A1, A2, and B1 of the Common European Framework of Reference for Languages (CEFR). We evaluated simplifications at the A2 level, which allows for greater information density than A1 but still focuses on essential information and avoids complex jargon to maximize comprehensibility.
- **GPT-4o** (OpenAI) is a language model designed for chat interactions. We used the model snapshot `gpt-4o-2024-08-06` via the OpenAI API.

We opted for closed-source systems because they are currently still showing stronger performance for German ATS than open LLMs (Yousaf et al., 2025).

3.1. Texts used for evaluation

The texts used in this study originate from a parallel corpus of source and automatically simplified German documents. The corpus was made available to us by an Austrian-based commercial

¹<https://app.capito.ai/>

provider of text simplification services in the context of a large-scale research project on automatic text simplification. The texts span various topics and genres, including news, administrative texts, and tourist advertisements. Eight source texts (hereafter: SOURCE) were selected according to the following criteria (cf. Säuberli et al., 2023): (a) length between 100 and 300 words, (b) large variety of topics, but no topics known to be familiar to a wide audience, and (c) no need for extensive additional context for comprehension. The corresponding automatically simplified versions (CAPITO) were generated at CEFR level A2 using *capito.ai*. Additionally, for each of the eight source texts, four further simplified versions were included in the sample set: (1) out-of-the-box zero-shot simplification using the GPT-4o model² (GPT-4O), (2) manual simplification (HUMAN), (3) post-edited version of CAPITO (CAPITO+PE), and (4) post-edited version of GPT-4O (GPT-4O+PE), the latter three versions being produced as part of the post-editing evaluation (Section 4). The sample texts in Austrian German underwent minimal lexical and orthographic adjustments, i.e. in three cases, words that might be unfamiliar to the Swiss end users were identified and replaced following Dürscheid et al. (2015) (e.g. *Initiator* > *Initiant*); orthography was also adapted to Swiss German standards (e.g. *Straße* > *Strasse*). Our findings are therefore applicable to all standard varieties of German. Refer to Appendix A for examples of SOURCE and corresponding target versions.

3.2. Comparison of simplifications

Table 1 compares the mean source text measures with those of the corresponding automatically simplified, post-edited, and manually simplified versions used in the evaluations.

These measures indicate that both LLM-generated simplified versions (CAPITO and GPT-4O) were very similar with respect to character length, word length, and vocabulary density, while HUMAN, CAPITO+PE, and GPT-4O+PE were of greater length and lower lexical variety. The former suggests that explication strategies (cf. Maaß, 2020) were applied in the post-editing and manual simplification processes, presumably with the addition of clarifications and examples to improve comprehensibility for the target group. Mean readability levels as measured by the Amstad Flesch-Index³ (Amstad, 1978) show a much more homogeneous pattern,

²Prompt: “Übersetze den folgenden Text in Leichte Sprache. Bitte gib nur eine Vereinfachung an, ohne Einleitung, Alternativen oder Kommentare” (English: “Translate the following text into Easy Language. Please only provide a simplification, without introduction, alternatives, or comments.”)

³Flesch reading ease is calculated based on the numbers of sentences, words, and syllables in a text. A low

Measure	SOURCE	CAPITO	GPT-4O	CAPITO+PE	GPT-4O+PE	HUMAN
Length in characters (excl. spaces)	1331	875	899	1435	1448	1548
Length in words	200	142	147	239	236	254
No. of unique words	130	88	93	131	126	135
Type/token ratio	0.657	0.625	0.639	0.568	0.555	0.556
Avg. sentence length (in words)	13.2	7.3	8.6	5.9	7.3	6.0
Avg. word length (in characters)	6.5	5.9	5.8	5.7	5.9	5.8
Amstad Flesch-Index	36	57	56	57	54	59

Table 1: Characteristics of the SOURCE, CAPITO, GPT-4O, CAPITO+PE, GPT-4O+PE, and HUMAN texts used in the study (mean values).

with all five simplified versions (rightmost columns) scoring between 54 and 59. Although this is a clear indication of greater readability compared to the source texts (mean=36), it should also be noted that readability scores below 60 are generally indicative of moderately difficult texts that require a secondary level of education.

4. Post-editing evaluation

This section describes our post-editing evaluation, in which we observed the process of human translators post-editing the output of two LLM-based ATS systems and measured their productivity.

4.1. Participants

Four German-speaking professional translators from a commercial provider of text simplification services were recruited through self-selection sampling. Participants reported having limited professional experience in post-editing following intralingual manual simplification, ranging from one to two years of experience to less than one year.

4.2. Procedure

This evaluation replicated and extended Carrer et al.'s post-editing study. The following methods were used: (a) a pre-task questionnaire to collect professional background data as well as attitudinal data on participants' practices in both inter- and intralingual translation settings; (b) automatic recording of participants' typing process in manual simplification (MS) and post-editing (PE) tasks; (c) a post-task questionnaire to investigate how the participants rated their productivity during the tasks. Each participant performed two MS and two PE tasks in their workplace and post-edited one CAPITO text in the first PE task and one GPT-4O text in the second PE task. The translators followed the provider's

reading ease (0-30) corresponds to very high difficulty, whereas a high reading ease (90-100) corresponds to very low difficulty.

internal criteria catalogue⁴ for text simplification in German. To prevent learning effects during the PE activity, each participant manually simplified and post-edited two different pairs of texts. Keystroke logging (GenoGraphiX-Log; Caporossi et al., 2023) and screen recording of both processes were employed.

In line with common practice in translation process (Alves, 2003; Kappus and Ehrensberger-Dow, 2020) and post-editing research (Krings, 2001; Alvarez-Vidal and Oliver, 2023), several quantitative measures were used to determine the effort involved in performing the tasks at document level. More specifically, effort was quantified in terms of task duration (i.e., minutes to produce the entire target text), number of keyboard and mouse-based interaction events (as a measure of addition, change, regression or navigation), number of cognitive pauses (i.e., pauses with duration greater than 300 ms), and total pause time (cf. Section 2.2).

4.3. Results and discussion

Student's t-tests with unequal variances were used for statistical analysis. All statistical tests were one-tailed with a 5% level of significance ($p < 0.05$). As can be seen from Table 2, no significant differences were observed between HUMAN and CAPITO+PE across any metric (all $p > 0.22$), indicating comparable levels of editing effort. In contrast, comparison between HUMAN and GPT-4O+PE showed a trend toward reduced effort, particularly for task duration ($p = 0.08$).

Interestingly, by comparing these results with those reported by Carrer et al. (2024)—who could not determine any meaningful decrease in effort in either MS or PE activity—, we observed a larger gain in efficiency across all effort metrics, especially in the GPT-4O+PE condition. This implies that post-editing is becoming more efficient, reducing task duration, user interactions, and idle time. While these effects remain statistically non-significant, they in-

⁴<https://www.capito.eu/en/capito-method/>

PE effort metrics	HUMAN	CAPITO+PE	GPT-4O+PE
Task duration (h:m:s)	00:22:34	00:17:12 p=0.22	00:15:41 p=0.08
Interaction events	2827	2379 p=0.33	1986 p=0.13
Cognitive pauses	392	318 p=0.24	315 p=0.15
Pause time (h:m:s)	00:17:44	00:15:48 p=0.38	00:13:43 p=0.17

Table 2: Mean values for each measure of effort compared between full manual simplification (HUMAN) vs. two post-editing conditions (CAPITO+PE and GPT-4O+PE). Statistical analysis according to two-sample Student’s t-tests assuming unequal variances (one tailed).

dicating that newer ATS systems may increasingly ease cognitive load in post-editing workflows.

Data from the post-task questionnaire—in which respondents were asked to rate their perceived productivity on a 5-point Likert scale (5=very high)—do not seem to clearly corroborate the above results. Average self-assessed productivity was similar in MS (mean=3.50) and PE activities (mean=3.75). In addition, subjective ratings revealed contrasting individual differences: productivity assessments were mixed across post-editing conditions, with one participant favoring CAPITO+PE, two GPT-4O+PE, and one reporting no difference. In contrast, perceptions of automatic translation quality showed the opposite tendency, with two participants favoring CAPITO and one favoring GPT-4O.

5. Expert evaluation

In the following section, we describe our expert evaluation, in which we obtained simplification quality ratings from experienced Easy Language service providers through an online questionnaire.

5.1. Participants

Four Swiss-based German-speaking experts in text simplification were recruited through purposive sampling. Three evaluators had over five years of professional experience in the field, and one over three. All of them regularly provided a wide portfolio of Easy and Plain Language services.

5.2. Procedure

Building on Carrer et al.’s approach, we employed an online evaluation questionnaire in which the four experts performed a source-based direct assessment (cf. Graham et al., 2013; Federmann, 2018) of the simplified texts in our sample set. For each SOURCE text, five corresponding simplified versions were employed, i.e., two automatically simplified versions (CAPITO and GPT-4O), two post-edited versions resulting from the PE evaluation (CAPITO+PE and GPT-4O+PE), and one manually simplified version (HUMAN) also resulting from the PE evaluation

(cf. Section 4). Experts evaluated one CAPITO+PE and one GPT-4O+PE version each.

Evaluators were asked to assign simplicity, adequacy, and fluency scores on 5-point scales (5=maximal quality; cf. Grabar and Saggion, 2022) to each simplified version and, if desired, insert comments. A 4×8 Latin square gave us a total of 32 experimental units and secured an unbiased response. Evaluators were not provided with any information about how the target texts had been produced.

5.3. Results and discussion

A comparison of the average simplicity ratings for each target text category shows that CAPITO+PE and GPT-4O+PE were rated higher (mean=4.50, range of 4–5) than the other three categories. The manually simplified texts obtained a slightly lower average score (mean=4.38), albeit with the same range of values (range of 4–5; cf. Figure 2). A similar pattern emerged for the adequacy and fluency ratings. CAPITO+PE were rated highest overall (mean=4.75, range of 4–5), while both automatically simplified versions were assigned the lowest ratings: CAPITO scored an average of 3.38 (range of 2–4) for adequacy, and GPT-4O received 3.75 (range of 2–5) for fluency (cf. Figure 2). The ratings are consistent with the experts’ qualitative feedback, which described the two LLM-based versions as generally lacking the required adequacy and simplicity. Although HUMAN and CAPITO+PE were often rated as being “very good”, it was also pointed out that both versions could have been simplified to a greater extent to ensure accessibility to the primary target group (cf. text measures in Section 3.2). Refer to Appendix B for examples of experts’ comments.

These results are in line with expert ratings reported in Carrer et al. (2024), where automatically simplified texts without post-editing were also assigned the lowest average scores and manual simplifications were rated in the higher end of the scale. Crucially, however, the models examined in the present study achieved stronger overall performance and facilitated more effective post-editing, with CAPITO+PE in fact obtaining the highest adequacy and fluency ratings.

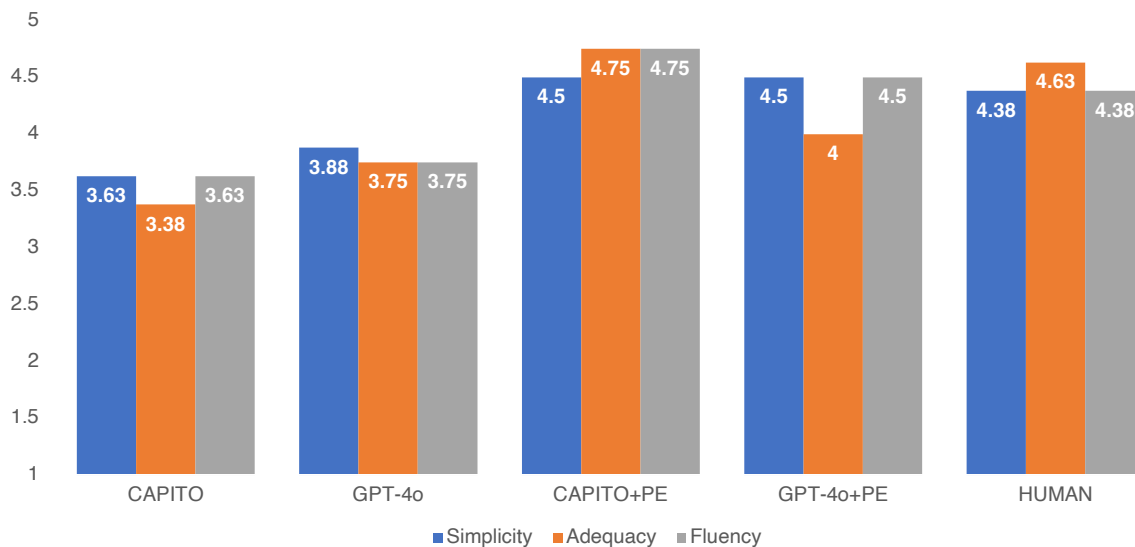


Figure 2: Average simplicity, adequacy, and fluency ratings (5= maximal quality) for each target text category (sample size: 40 texts, of which eight SOURCE, eight CAPITO, eight GPT-4O, four CAPITO+PE, four GPT-4O+PE, eight HUMAN.)

6. End-user evaluation

The goal of the end-user evaluation was to compare the comprehensibility of the SOURCE texts and three corresponding simplified versions—the HUMAN simplification, the CAPITO simplification, and the GPT-4O simplification.

We followed the methodology of Säuberli et al. (2024) and measured (a) end-users’ actual comprehension using multiple-choice questions, and (b) perceived difficulty using Likert-style ratings.

6.1. Comprehension questions

For each text, we created four multiple-choice questions with four options and one correct answer each. The first question concerned the high-level topic of the text, while the remaining three focused on more specific details. Attention was paid to ensuring that the questions were comprehensible and grammatically correct, unambiguous, and independent of one another (i.e., answering a question did not depend on obtaining the correct answer to any previous question). The questions were also designed to be answerable without requiring additional world knowledge, while at the same time not solvable through world knowledge alone. In addition, each question was reviewed by three co-authors to verify compliance with these criteria and then pilot tested by a person with cognitive impairments.

In order not to risk biasing the questions towards one of the simplification methods, we designed the multiple-choice test based exclusively on the source text and then showed the same questions with all text versions. However, this approach re-

sulted in some questions becoming unanswerable based on the simplified versions (e.g., due to systems omitting or changing some information from the source). In those cases, we excluded the responses from the analysis (1 question for CAPITO, 2 for GPT-4O, and 1 for HUMAN).

6.2. Participants

To explore group-specific differences, two cohorts from different populations were recruited. All participants took part on a voluntary basis, gave informed consent, and received monetary compensation.

6.2.1. Target group

Following approval by the institutional ethics committee, we recruited 14 participants⁵ from a recreational program for persons with cognitive impairments in German-speaking Switzerland. Seven were female, six were male and one was non-binary, ranging in age from 29 to 72 (mean=45) at the time of recruitment. All participants had some form of cognitive impairment with an onset before or after the age of 18 and therefore represented a primary target group for simplified language. Furthermore, all participants were legally entitled to sign the informed consent forms independently.

In a background questionnaire administered prior to the first session, most respondents (n=10) stated that they engaged in reading activities on a daily basis, while a smaller number (n=4) reported doing

⁵Two participants dropped out of the study and completed only half of the experiment (i.e., each submitted 16 out of the 32 expected responses).

so infrequently. Similarly, only a few participants stated that they regularly read texts in Easy Language. The majority (n=12) also reported daily use of touchscreen devices, with a few (n=2) indicating less frequent use. All participants listed German as their native language.

6.2.2. Control group

To compare the effects of text simplification on people outside the primary target group, we additionally recruited 15 people without cognitive impairments—current undergraduate and graduate students—through simple random sampling at a university for applied sciences. Twelve were female, one was male, and two were non-binary, ranging in age from 22 to 47 (mean=27.6). Twelve participants were native German speakers, three did not list German as their native language but were proficient at CEFR level C1 or higher. Similarly to the target group, two-thirds of the participants in the control group were not used to reading documents in Easy Language.

6.3. Procedure

Participants completed the experiment using the *Okra* app (Säuberli et al., 2023) on Apple iPads. Each participant read all eight texts in exactly one version (two texts per version). Text order was randomized for each participant. Control group participants read all texts in a single session, while the experiment was split into two sessions for target group participants to avoid fatigue.

Following an initial reading, participants were asked to rate how difficult the text was to read, how interesting they found the text, and how familiar they were with the content of the text. The text was then shown again and participants were asked to answer the four comprehension questions. We measured the reading time during the initial reading and the response time for each comprehension question.

6.4. Statistical analysis

The data collection resulted in a total of 403 individual comprehension question responses and 104 ratings from the target group, and 465 responses and 120 ratings in the control group (after removing unanswerable questions as mentioned in Section 6.1).

For statistical analysis, we applied Bayesian many-facets Rasch models (Linacre, 1989) from item response theory (IRT). These are essentially binary or ordinal regression models that predict response probabilities from a linear combination of several latent variables.

For the comprehension questions, we fitted logistic models that predict the probability of a correct re-

sponse based on the participant's proficiency level, the item's difficulty, and the difficulty contributed by the text version. For the difficulty ratings, we used graded response models (Samejima, 1997) predicting the probability for each rating level from the participant's severity, the text's difficulty (independent of version), and the text version's contribution. We fitted separate models for the control and target groups.

6.5. Results and discussion

Figure 3a shows the difficulty estimates based on the comprehension data. In the target group, the simplifications from GPT-4O were the easiest (easier than source with 90% credibility), while the other simplifications did not have a strong effect on response accuracy compared to the source. Interestingly, for the control group, we see the opposite trends, although there is a strong ceiling effect (only 4.1% incorrect responses, compared to 31.5% in the target group), leading to less reliable estimates.

Regarding the perceived difficulty ratings (shown in Figure 3b), the results from the target group largely follow the same pattern as the comprehension results, but the differences are less strong. In the control group, the source texts were rated significantly more difficult than all the simplified versions, with no significant differences between the two ATS systems.

These results are strikingly different from those reported by Säuberli et al. (2024). The ATS outputs are much more successful at reducing difficulty, which can be attributed to the stronger performance of the LLM-based systems employed in the study at hand. In fact, GPT-4O-generated texts were better understood by the target group than the fullmanual simplifications⁶ (75% vs. 67% comprehension accuracy; difference with 80% credibility in the IRT model). This suggests that LLM-based ATS for German has made substantial progress in making texts more accessible to readers with cognitive impairments. On the other hand, we observe stark differences between the two participant groups, with a stronger preference for the manual simplification in the control group, which is in line with Säuberli et al. (2024).

On average, the target group exhibited substantially lower reading speeds (median=134 words per minute) than the control group (median=194 words per minute). Given an average reading speed of 179 words per minute for adults in German (Trauzettel-Klosinski et al., 2012), this suggests that most participants read the texts carefully before rat-

⁶Note that in Säuberli et al.'s study, the target group participants were generally much more experienced with reading Easy Language. This may have additionally contributed to the differences in results.

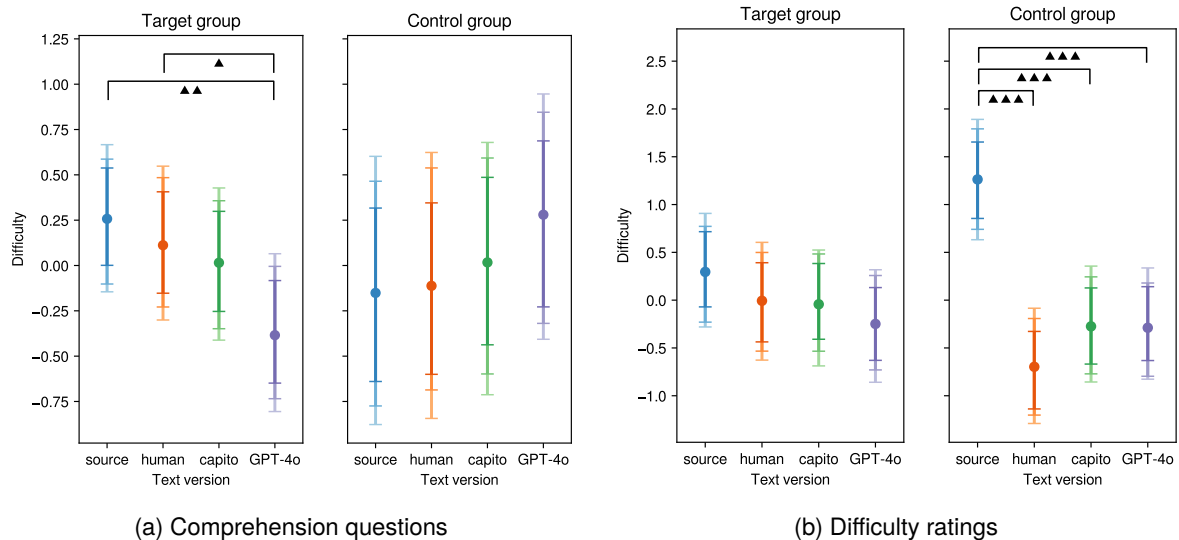


Figure 3: IRT-based difficulty estimates based on the response accuracies in the comprehension questions and the perceived difficulty ratings. Difficulty is measured in logits. Estimates for the four text versions are constrained to sum to zero. Error bars represent 80%, 90%, and 95% Bayesian credible intervals (CI). ▲/▲▲/▲▲▲ indicates that the 80/90/95% CI of the difference between the two text versions does not include zero.

ing them. However, there were also some outliers with implausibly high reading speeds in both groups, indicating that some texts were only skimmed or skipped entirely during the first reading.

7. Overall discussion and conclusions

We carried out an extensive mixed-method evaluation study involving post-editors, experts, and end users. The findings revealed clear variations in the perception and comprehension of the simplified texts, which reinforces the advantages of a multi-stakeholder, human-centric approach to the evaluation of ATS models, as proposed by Carrer et al. (2024). Returning to our central research question—how effective latest-generation LLM-based systems are in real-world document-level ATS scenarios—the effectiveness of the ATS outputs under review emerged across several dimensions, including post-editing efficiency, expert direct assessments and, in the case of GPT-4o, comprehensibility for the target group (increase of 12.5 percentage points in comprehension accuracy). Compared to Carrer et al. (2024), the results indicate that both CAPITO and GPT-4o enabled more efficient post-editing, with an increase in productivity of around 30% compared to full manual simplification, thereby compensating for qualitative shortcomings in the automatic outputs. Potential objections that post-editors' performance improved over time merely due to greater familiarity with the relevant workflow can be dismissed, since, as outlined in Section 4.1, participants had

only limited prior experience in post-editing ATS outputs.

When comparing evaluation perspectives, we observed that the control group's ratings—which identified the most difficult texts in the SOURCE documents—aligned with those of the experts, while the target group's perceived difficulty scores and response accuracy rates clearly favored the simplifications from GPT-4o. This finding is particularly noteworthy, as it highlights the considerable progress of recent LLM-based ATS systems for German in producing simplifications that measurably enhance accessibility for end users. It should be emphasized that, in comparison with Säuberli et al. (2024), respondents in our study exhibited far less skipping or skimming behavior, as evidenced by reading speed measurements (cf. Section 6.5). The extent to which target group characteristics (e.g., age, familiarity with Easy Language materials, or prior experience) influenced these results remains uncertain and warrants further investigation.

Future research will expand the current findings by incorporating more quantitative and qualitative data from additional experiments with diverse groups of target users. Integrating think-aloud protocols or post-task interviews into future holistic evaluations of ATS models could provide valuable insights into cognitive load, while also offering more reliable and nuanced perspectives into the end users' subjective preferences and perceptions. Furthermore, extending beyond metric and human evaluations based on test sets and conducting additional impact evaluations of LLM-based ATS in real-

world usage could facilitate a more profound comprehension of how deployed ATS systems followed by post-editing may enhance key performance indicators in complex real-world settings, such as text simplification agencies in multilingual Switzerland.

Acknowledgements

We are indebted to CFS GmbH (capito) for providing help and resources in conducting the post-editing evaluation. In addition, we would like to thank all study participants for their time and effort. This work was funded by the Swiss Innovation Agency (Innosuisse) Flagship IICT (PFFS-21-47).

Limitations

Recruiting Swiss-based participants for our end-user evaluation presented multiple obstacles which eventually affected our sample size. As often reported in the literature (cf. Becker et al., 2004), recruiting people with cognitive impairments for inclusive research projects poses considerable challenges. People's motivation to participate in research may be influenced by several factors including unfamiliarity with the research process. In our case, the fluctuating availability of recruited participants due to health reasons, professional commitments or loss of interest also proved challenging.

Moreover, multiple-choice testing formats have often attracted criticism for measuring test-taking skills more than receptive language ability (Jones, 2020; Stevenson, 2010). Administering multiple-choice formats may be even more problematic for people with cognitive impairments (Finlay and Lyons, 2001). As mentioned above, future evaluations should include broader qualitative measures to counterbalance such limitations.

Ethical considerations

End-user study. When designing the end-user study, we took particular care to adhere to ethical standards. For the target group, we asked them to read no more than four texts per session. Participants were allowed to interrupt or withdraw their participation at any point. Participant information and consent forms were translated to Easy Language. No sensitive information (e.g., diagnoses or specifics about participants' disabilities) were collected. The study was approved by an institutional review board.

Use of generative models. We used ChatGPT for brainstorming ideas as starting points for writing comprehension questions, GitHub Copilot for coding assistance in data analysis, and DeepL and

ChatGPT for improving wording during paper writing. All generated outputs were thoroughly and manually checked.

Bibliographical References

- Sweta Agrawal and Marine Carpuat. 2024. [Do text simplification systems preserve meaning? A human evaluation via reading comprehension](#). *Transactions of the Association for Computational Linguistics*, 12:432–448.
- Sarah Ahrens, Silvana Deilen, Sergio Hernandez Garrido, Ekaterina Lapshinova-Koltunski, and Christiane Maaß. 2025. [Evaluation of machine translation errors in German Plain Language texts in the domain of health information](#). In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 176–185, Hannover, Germany. HsH Applied Academics.
- Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. [Automated text simplification: A survey](#). *ACM Comput. Surv.*, 54(2).
- Oliver Alonzo, Jessica Trussell, Becca Dingman, and Matt Huenerfauth. 2021. [Comparison of methods for evaluating complexity of simplified texts among deaf and hard-of-hearing adults at different literacy levels](#). In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, Yokohama, Japan. Association for Computing Machinery.
- Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. [Data-driven sentence simplification: Survey and benchmark](#). *Computational Linguistics*, 46(1):135–187.
- Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. [The \(un\)suitability of automatic evaluation metrics for text simplification](#). *Computational Linguistics*, 47(4):861–889.
- Sergi Alvarez-Vidal and Antoni Oliver. 2023. [Assessing MT with measures of PE effort](#). *Amper-sand*, 11:100125.
- Fabio Alves. 2003. *Triangulating Translation: Perspectives in process oriented research*. John Benjamins Publishing Company, Amsterdam.
- Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Ph.D. thesis, University of Zurich, Zürich.
- Heather Becker, Greg Roberts, Janet Morrison, and Julie Silver. 2004. [Recruiting people with disabilities as research participants: Challenges and](#)

- strategies to address them. *Mental Retardation*, 42(6):471–475.
- Nadine Beks van Raaij, Daan Kolkman, and Ksenia Podoyntsyna. 2024. [Clearer governmental communication: Text simplification with ChatGPT evaluated by quantitative and qualitative research](#). In *Proceedings of the Workshop on DeTermt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 152–178, Torino, Italia. ELRA and ICCL.
- Gilles Caporossi, Christophe Leblay, and Hakim Usoof. 2023. [GenoGraphiX-Log version 2.0. User guide](#). *Les Cahiers Du GERAD G-2020-68*, pages 1–63.
- Luisa Carrer, Andreas Säuberli, Martin Kappus, and Sarah Ebling. 2024. [Towards holistic human evaluation of automatic text simplification](#). In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 71–80, Torino, Italia. ELRA and ICCL.
- Liam Cripwell, Joël Legrand, and Claire Gardent. 2024. [Evaluating document simplification: On the importance of separately assessing simplicity and meaning preservation](#). In *Proceedings of the 3rd Workshop on Tools and Resources for People with READING Difficulties (READI) @ LREC-COLING 2024*, pages 1–14, Torino, Italia. ELRA and ICCL.
- Scott A Crossley, Hae Sung Yang, and Danielle S McNamara. 2014. What’s so simple about simplified texts? A computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language*, 26(1):92–113.
- Cristina Cumbreño and Nora Aranberri. 2021. [What Do You Say? Comparison of Metrics for Post-editing Effort](#), page 57–79. Springer International Publishing.
- Silvana Deilen, Sergio Hernández Garrido, Ekaterina Lapshinova-Koltunski, and Christiane Maaß. 2023. [Using ChatGPT as a CAT tool in Easy Language translation](#). In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 1–10, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Silvana Deilen, Ekaterina Lapshinova-Koltunski, Sergio Garrido, Julian Hörner, Christiane Maaß, Vanessa Theel, and Sophie Ziemer. 2024. [Evaluation of intralingual machine translation for health communication](#). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 469–479, Sheffield, UK. European Association for Machine Translation (EAMT).
- Christa Dürscheid, Stephan Elspaß, and Arne Ziegler. 2015. [Variantengrammatik des Standarddeutschen. Konzeption, methodische Fragen, Fallanalysen](#), pages 207–236. VR Unipress.
- Sarah Ebling, Alessia Battisti, Marek Kostrzewa, Dominik Pfütze, Annette Rios, Andreas Säuberli, and Nicolas Spring. 2022. [Automatic text simplification for German](#). *Frontiers in Communication*, Volume 7 - 2022.
- ELIS Research. 2025. [European language industry survey 2025: Trends, expectations and concerns of the European language industry](#). Accessed: 16 Oct 2025.
- Inmaculada Fajardo, Vicenta Ávila, Antonio Ferrer, Gema Tavares, Marcos Gómez, and Ana Hernández. 2014. Easy-to-read texts for students with intellectual disability: linguistic factors affecting comprehension. *Journal of applied research in intellectual disabilities*, 27(3):212–225.
- Christian Federmann. 2018. [Appraise evaluation framework for machine translation](#). In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.
- W. M. L. Finlay and E. Lyons. 2001. [Methodological issues in interviewing and using self-report questionnaires with people with mental retardation](#). *Psychological Assessment*, 13(3):319–335.
- Sian Gooding. 2022. [On the ethical considerations of text simplification](#). In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57, Dublin, Ireland. Association for Computational Linguistics.
- Sian Gooding, Yevgeni Berzak, Tony Mak, and Matt Sharifi. 2021. [Predicting text readability from scrolling interactions](#). In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 380–390, Online. Association for Computational Linguistics.
- Natalia Grabar and Horacio Saggion. 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*, pages 453–463, Avignon, France. ATALA.
- Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. [Continuous measurement scales in human evaluation of machine](#)

- translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.
- Theo Guidroz, Diego Ardila, Jimmy Li, Adam Mansour, Paul Jhun, Nina Gonzalez, Xiang Ji, Mike Sanchez, Sujay Kakarmath, Mathias MJ Bellaiche, Miguel Ángel Garrido, Faruk Ahmed, Divyansh Choudhary, Jay Hartford, Chenwei Xu, Henry Javier Serrano Echeverria, Yifan Wang, Jeff Shaffer, Eric, Cao, Yossi Matias, Avinatan Hassidim, Dale R Webster, Yun Liu, Sho Fujiwara, Peggy Bui, and Quang Duong. 2025. [LLM-based text simplification and its effect on user comprehension and cognitive load](#).
- Matt Huenerfauth, Lijun Feng, and Noémie Elhadad. 2009. [Comparing evaluation techniques for text readability software for adults with intellectual disabilities](#). In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '09, page 3–10, Pittsburgh, Pennsylvania, USA. Association for Computing Machinery.
- Glyn Jones. 2020. Designing multiple-choice test items. In *The Routledge Handbook of Second Language Acquisition and Language Testing*, chapter 9, pages 90–101. Routledge.
- Martin Kappus and Maureen Ehrensberger-Dow. 2020. [The ergonomics of translation tools: understanding when less is actually more](#). *The Interpreter and Translator Trainer*, 14(4):386–404.
- Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. [BLESS: Benchmarking large language models on sentence simplification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.
- Maarit Koponen. 2016. [Is machine translation post-editing worth the effort? A survey of research into post-editing and effort](#). *The Journal of Specialised Translation*, (25):131–148.
- Hans P. Krings. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes*. Translation studies. Kent State University Press, Kent, Ohio and London.
- Isabel Lacruz, Michael Denkowski, and Alon Lavie. 2014. [Cognitive demand and cognitive effort in post-editing](#). In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 73–84, Vancouver, Canada. Association for Machine Translation in the Americas.
- Isabel Lacruz, Gregory M. Shreve, and Erik Angelone. 2012. [Average pause ratio as an indicator of cognitive effort in post-editing: A case study](#). In *Workshop on Post-Editing Technology and Practice*, San Diego, California, USA. Association for Machine Translation in the Americas.
- Gondy Leroy, James E Endicott, David Kauchak, Obay Mouradi, and Melissa Just. 2013. User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention. *Journal of medical Internet research*, 15(7):e144.
- Gondy Leroy, David Kauchak, Diane Haeger, and Douglas Spegman. 2022. Evaluation of an online text simplification editor using manual and automated metrics for perceived and actual text difficulty. *JAMIA Open*, 5(2):ooac044.
- John Michael Linacre. 1989. *Many-faceted Rasch measurement*. Ph.D. thesis, The University of Chicago.
- Christiane Maaß. 2020. *Easy Language–Plain Language–Easy Language Plus: Balancing comprehensibility and acceptability*. Frank & Timme, Berlin.
- Margot Madina, Itziar Gonzalez-Dios, and Melanie Siegel. 2024. [Towards reliable E2R texts: A proposal for standardized evaluation practices](#). In *Computers Helping People with Special Needs*, pages 224–231, Cham. Springer Nature Switzerland.
- Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. [Zero-shot crosslingual sentence simplification](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126, Online. Association for Computational Linguistics.
- Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. [MUSS: Multilingual unsupervised sentence simplification by mining paraphrases](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.
- Paloma Martínez, Alberto Ramos, and Lourdes Moreno. 2024. [Exploring Large Language Models to generate Easy to Read content](#). *Frontiers in Computer Science*, Volume 6 - 2024.

- Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2025. [Deep learning approaches to lexical simplification: A survey](#). *Journal of Intelligent Information Systems*, 63(1):111–134.
- Jipeng Qiang, Minjiang Huang, Yi Zhu, Yunhao Yuan, Chaowei Zhang, and Kui Yu. 2025. [Redefining simplicity: Benchmarking Large Language Models from lexical to document simplification](#).
- Ehud Reiter. 2025. [We should evaluate real-world impact](#). *Computational Linguistics*, pages 1–13.
- Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013a. [Simplify or help? Text simplification strategies for people with dyslexia](#). In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility, W4A '13*, Rio de Janeiro, Brazil. Association for Computing Machinery.
- Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013b. [Frequent words improve readability and short words improve understandability for people with dyslexia](#). In *INTERACT '13*, pages 203–219, Cape Town, South Africa. Springer.
- Luz Rello, Susana Bautista, Ricardo Baeza-Yates, Pablo Gervás, Raquel Hervás, and Horacio Saggion. 2013c. [One half or 50%? An eye-tracking study of number representation readability](#). In *INTERACT '13*, pages 229–245, Cape Town, South Africa. Springer.
- Michael Ryan, Tarek Naous, and Wei Xu. 2023. [Revisiting non-English text simplification: A unified multilingual benchmark](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.
- Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. [Making it Simplex: Implementation and evaluation of a text simplification system for Spanish](#). *ACM Trans. Access. Comput.*, 6(4).
- Fumiko Samejima. 1997. [Graded response model](#). In Wim J. van der Linden and Ronald K. Hambleton, editors, *Handbook of Modern Item Response Theory*, pages 85–100. Springer New York, New York, NY.
- Andreas Säuberli, Franz Holzknrecht, Patrick Haller, Silvana Deilen, Laura Schiffli, Silvia Hansen-Schirra, and Sarah Ebling. 2024. [Digital comprehensibility assessment of simplified texts among persons with intellectual disabilities](#). In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.
- Scarton Scarton, Mikel L. Forcada, Miquel Esplà-Gomis, and Lucia Specia. 2019. [Estimating post-editing effort: a study on human judgements, task-based and reference-based metrics of MT quality](#). In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.
- Sanja Štajner and Sergiu Nisioi. 2018. [A detailed evaluation of neural sequence-to-sequence models for in-domain and cross-domain text simplification](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Marie Stevenson. 2010. [Researching reading](#). In Brian Paltridge and Aek Phakiti, editors, *Continuum Companion to Research Methods in Applied Linguistics*, pages 174–190. Continuum London, England.
- Regina Stodden. 2021. [When the scale is unclear – analysis of the interpretation of rating scales in human evaluation of text simplification](#). In *Proceedings of the First Workshop on Current Trends in Text Simplification (CTTS 2021) co-located with the 37th Conference of the Spanish Society for Natural Language Processing (SE-PLN2021)*, Online (initially located in Málaga, Spain), September 21st, 2021, volume 2944 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Regina Stodden. 2026. [Automatic German Text Simplification: Data, Evaluation, and Models](#), 1 edition, volume 21 of *Easy – Plain – Accessible*. Frank & Timme.
- Andreas Säuberli, Silvia Hansen-Schirra, Franz Holzknrecht, Silke Guterath, Silvana Deilen, Laura Schiffli, and Sarah Ebling. 2023. [Enabling text comprehensibility assessment for people with intellectual disabilities using a mobile application](#). *Frontiers in Communication*, 8.
- Susanne Trauzettel-Klosinski, Klaus Dietz, and the IReST Study Group. 2012. [Standardized Assessment of Reading Performance: The New International Reading Speed Texts IReST](#). *Investigative Ophthalmology & Visual Science*, 53(9):5452–5461.
- Lucas Nunes Vieira. 2017. [Cognitive effort and different task foci in post-editing of machine translation: A think-aloud study](#). *Across Languages and Cultures*, 18(1):79–105.

Sanja Štajner. 2021. [Automatic text simplification for social good: Progress and challenges](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652, Online. Association for Computational Linguistics.

Raeesa Yousaf, Marina Walther, and Michael Gertz. 2025. Accessible language simplification: Large Language Models for generating Easy German. In *KI 2025: Advances in Artificial Intelligence*, pages 204–217, Cham. Springer Nature Switzerland.

A. Sample texts with comprehension questions

The following is a sample SOURCE text along with four corresponding simplified versions (i.e. HUMAN, CAPITO, GPT-4O, and GPT-4O+PE) and four comprehension questions.

A.1. SOURCE

Back to the roots - die Post bleibt!

Nachdem Herr Bürgermeister Christian Sander von der Post AG Wien informiert wurde, dass die Postfiliale in der Hauptstrasse nicht kostendeckend geführt werden kann und Anfang Juni eine Entscheidung bezüglich der weiteren Vorgehensweise fällt, wurde er sofort tätig.

Nachdem die Post AG Stillschweigen verlangte, hat er dem natürlich Folge geleistet. Seine Kontakte zum Land und auch ins Ministerium konnten Früchte tragen. Das Wichtigste dabei war aber im Hintergrund die Zusammenarbeit mit der Post AG.

Nicht immer führen populistische Massnahmen zum Ziel. Im Gegenteil: Oft sind diese auch hinderlich in wichtigen Verhandlungen. Über den Verbleib der BAWAG ist dem Bürgermeister zum derzeitigen Zeitpunkt noch nichts bekannt!

In weiterer Folge wird im Herbst eine Umstrukturierung stattfinden und die Post auf ihren alten Standort - das Gebäude ist ja nach wie vor im Eigentum der Post - in die Ungerfeldgasse Nr. 4 übersiedeln.

Hier sind genügend Parkplätze vorhanden und auch der Postlieferwagen behindert nicht mehr in der Fussgängerzone. Ein Mitarbeiter wird in einer nach jetziger Frequenzerhebung gestalteten Arbeitszeit für Sie da sein und darüber hinaus entsteht in Kindberg eine der modernsten Postfilialen Österreichs!

Die Gemeinde wird möglichst zeitnahe über die weiteren Schritte informieren. Es freut uns, dass durch die gemeinsamen Bemühungen unseres Bürgermeisters und der Post AG unsere Post-Filiale in Kindberg erhalten bleibt.

A.2. HUMAN

Die Post bleibt in Kindberg!

In Kindberg gibt es in der Hauptstrasse eine Post-Stelle.

Sie wird geschlossen,

aber zum Glück bekommt Kindberg eine neue Post-Stelle.

Der Hauptsitz der Post ist in Wien.

Bürgermeister Christian Sander hat gesagt:

Die Post in Kindberg kostet zu viel Geld.

Wir entscheiden Anfang Juni,

ob wir die Post schliessen müssen.

Der Hauptsitz hat gefordert,

dass Christian Sander diese Information nicht veröffentlicht.

Der Bürgermeister hat sich natürlich daran gehalten.

Aber er wollte nicht, dass die Post schliesst.

Deshalb hat er sofort etwas unternommen.

Christian Sander hat mit Politiker*innen

vom Bundesland und vom Bund gesprochen.

Auch mit der Post hat er immer zusammengearbeitet.

Das zeigt: Es hilft nicht immer, wenn man

ein Problem öffentlich macht und darüber schimpft.

Das ist manchmal sogar schlecht für die Verhandlungen.

Christian Sander hat das also gut gelöst.

Wie geht es jetzt mit der Post weiter?

Die Post bleibt in Kindberg.

Aber im Herbst gibt es einige Änderungen.

Die Post kommt wieder dorthin,

wo sie früher war: in die Ungerfeldgasse 4.

Das ist kein Problem,

weil das Gebäude noch der Post gehört.

Hier gibt es genug Parkplätze.

Auch der Post-Lieferwagen hat genug Platz.

Er stört nicht mehr in der Fussgänger-Zone,

wie es in der Hauptstrasse war.
Ausserdem schaut die Post jetzt:
Wann kommen viele Menschen?
Wann sollten wir offen haben?
Dann passt die Post ihre Öffnungszeiten an.
Die neue Post-Stelle wird
eine der modernsten Post-Stellen in ganz Österreich.
Wie geht es mit der BAWAG weiter?
In der Hauptstrasse war auch eine BAWAG
in der Post-Stelle.
Der Bürgermeister weiss noch nicht, wie es damit weitergeht.
Und jetzt?
Wir informieren Sie so schnell wie möglich,
was als nächstes passiert.
Unser Bürgermeister und die Post haben sich sehr bemüht.
Deshalb haben wir weiterhin eine Post in Kindberg.
Das freut uns sehr!

A.3. CAPITO

Die Post bleibt in Kindberg

Der Bürgermeister Christian Sander erfuhr von der Post AG Wien, dass die Postfiliale in der Hauptstrasse nicht genug Geld verdient.
Im Juni entscheidet die Post AG, was sie macht.
Der Bürgermeister handelte sofort.
Die Post AG wollte, dass er nichts sagt.
Er hielt sich daran.
Seine Kontakte halfen ihm.
Wichtig war die Zusammenarbeit mit der Post AG.
Populistische Massnahmen helfen oft nicht.
Sie stören wichtige Verhandlungen.
Der Bürgermeister weiss noch nichts über die BAWAG.
Im Herbst zieht die Post in die Ungerfeldgasse Nummer 4.
Dort gibt es genug Parkplätze.
Der Postwagen stört nicht mehr in der Fussgängerzone.
Ein Mitarbeiter arbeitet nach Bedarf.
In Kindberg entsteht eine moderne Postfiliale.
Die Gemeinde informiert bald über die nächsten Schritte.
Der Bürgermeister und die Post AG haben die Post-Filiale gerettet.

A.4. GPT-40

Die Post in Kindberg bleibt offen. Der Bürgermeister, Herr Christian Sander, hat mit der Post gesprochen. Die Post in der Hauptstrasse bringt nicht genug Geld. Es gab ein Geheimgespräch. Der Bürgermeister arbeitete mit dem Land und dem Ministerium zusammen. Im Herbst zieht die Post in die Ungerfeldgasse Nr. 4 um. Dort gibt es mehr Parkplätze. Die neue Post wird modern sein. Die Gemeinde informiert bald über die nächsten Schritte. Es ist gut, dass die Post in Kindberg bleibt.

A.5. GPT-40+PE

Die Post in Kindberg bleibt offen!

Die Post AG in Wien hat Bürgermeister Christian Sander informiert, dass die Post in der Hauptstrasse in Kindberg zu wenig Geld bringt. In Juni sollte es eine Entscheidung geben, ob sie zugesperrt wird oder nicht. Der Bürgermeister hat sofort etwas für die Post-Filiale getan. Das Wichtigste waren die Gespräche mit der Post AG in Wien. Aber auch seine Kontakte zum Land und zum Ministerium haben geholfen. Im Herbst zieht die Post in die Ungerfeldgasse Nr. 4 um. Das ist der alte Standort und das Gebäude gehört ja noch immer der Post. Dort gibt es genug Parkplätze und die Lieferwagen der Post fahren nicht mehr in der Fussgängerzone. Die neue Post-Filiale wird eine der modernsten Filialen in ganz Österreich sein.

Die Gemeinde wird Sie so schnell wie möglich informieren, wie es weitergeht. Es ist gut, dass es in Kindberg weiter eine Post-Filiale geben wird!

A.6. Comprehension questions

Q1. Um was geht es in dem Text?

- A) Die Post bleibt in Kindberg. (correct)
- B) Ein neues Rathaus wird gebaut. (incorrect)
- C) Die Fussgänger-Zone wird grösser. (incorrect)
- D) Die Post wird komplett geschlossen. (incorrect)

Q2. Was hat der Bürgermeister gemacht?

- A) Er hat mit seinen Kontakten gesprochen. (correct)
- B) Er hat selbst eine neue Post eröffnet. (incorrect)
- C) Er hat einen Protest organisiert. (incorrect)
- D) Steht nicht im Text. (incorrect)

Q3. Was passiert im Herbst mit der Post in Kindberg?

- A) Die Post wird geschlossen. (incorrect)
- B) Die Post ist wieder in der Ungerfeldgasse. (correct)
- C) Die Post bleibt in der Hauptstrasse. (incorrect)
- D) Steht nicht im Text. (incorrect)

Q4. Was bekommt Kindberg nach dem Umbau?

- A) Eine moderne Post. (correct)
- B) Ein neues Einkaufs-Zentrum. (incorrect)
- C) Ein neues Parkhaus neben der Post. (incorrect)
- D) Steht nicht im Text. (incorrect)

B. Expert comments

The following is a selection of comments from the expert evaluation. Comments were translated into English from German.

B.1. HUMAN

Although HUMAN were often rated as being “very good”, it was also pointed out that they could have been further simplified to ensure accessibility to the primary target group.

- “I think the target text is a very successful translation. All the information is there, it is easy to understand linguistically, and the use of ‘you’ makes it feel personal. The topic is well explained. The presentation and length are also good.”
- “All in all, a very successful translation.”
- “The text is not written in Easy Language, as it contains many subordinate clauses, third-person pronouns, genitive constructions, gender-neutral spelling, etc. The target audience for Easy Language would not understand this text.”
- “This text is not in Easy Language. It contains, for example, several statements per sentence, nominal style and participles. The sentences are difficult in structure, e.g. starting with an object.”

B.2. CAPITO

Experts reported several simplicity/adequacy issues.

- “Target text is unstructured, with disjointed sentences and poor prioritisation of the selected content, making it difficult to understand.”
- “English expressions, words without explanations and the graphic design hinder comprehension.”
- “The individual sentences are simpler, making the entire text easier to read. However, the use of many difficult words means that the content is actually hardly any easier to understand.”
- “The structure of the information (plot) is incorrect in the Easy Language version, and the whole thing seems very disjointed.”

B.3. CAPITO+PE

Although one expert rated CAPITO+PE as very good, another questioned their simplicity.

- “I really like the adapted structure. First, the project is explained. The information about the founder comes at the end. The subheadings provide good orientation. The sections also help to guide the reader. I think it’s a successful adaptation.”
- “The text is simple, and the layout (each sentence begins on a new line) is particularly helpful. Nevertheless, it is not yet Easy Language (subordinate clauses, third-person pronouns, etc.).”
- “No comments.”

B.4. GPT-4O

Experts reported simplicity/adequacy issues.

- “This target text is far from being written in Easy Language. It is good that a lot of unnecessary information has been removed. This would not be relevant for other target groups either. However, the text uses a lot of passive constructions, participle constructions, future tense, subordinate clauses, and so on. It would be too difficult even if it was labelled ‘plain language’.”
- “I have the impression that the target text does not accurately reflect the content.”
- “The sentences are simpler and the text is easier to read, but the use of difficult and abstract terms that are not explained makes it less comprehensible than the source text.”

B.5. GPT-4O+PE

Experts reported mainly simplicity issues.

- “Graphical representation too dense in the Easy Language text, long sentences, numerous words unchanged and without explanation.”
- “The target text could be simplified a little more. There are sentences with more than one piece of information and passive constructions.”
- “This is no text in Easy Language. It does not help to break lines after just a few words. This makes the sentences appear shorter, but in fact they are not.”