Towards Holistic Human Evaluation of Automatic Text Simplification

Luisa Carrer[1], Andreas Säuberli[2], Martin Kappus[1], Sarah Ebling[2]
[1]School of Applied Linguistics, ZHAW Zurich University of Applied Sciences
[2]Department of Computational Linguistics, University of Zurich
cars@zhaw.ch, andreas@cl.uzh.ch, kapm@zhaw.ch, ebling@cl.uzh.ch

Abstract

Text simplification refers to the process of rewording within a single language, moving from a standard form into an easy-to-understand one. Easy Language and Plain Language are two examples of simplified varieties aimed at improving readability and understanding for a wide-ranging audience. Human evaluation of automatic text simplification is usually done by employing experts or crowdworkers to rate the generated texts. However, this approach does not include the target readers of simplified texts and does not reflect actual comprehensibility. In this paper, we explore different ways of measuring the quality of automatically simplified texts. We conducted a multi-faceted evaluation study involving end users, post-editors, and Easy Language experts and applied a variety of qualitative and quantitative methods. We found differences in the perception and actual comprehension of the texts by different user groups. In addition, qualitative surveys and behavioral observations proved to be essential in interpreting the results.

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

1. Introduction

Text simplification is a form of intralingual translation, or rewording, within one language, i.e. from the standard variety into a simplified variety (cf. Hansen-Schirra et al., 2020). Easy Language and Plain Language are two examples of easy-to-understand varieties aimed at optimizing the readability and comprehensibility of texts for a wide and heterogeneous target audience. More specifically, Easy Language is a strongly controlled form of language and is based on strict sets of rules (Maaß, 2020; Bock and Pappert, 2023). Primary target groups include persons with intellectual disabilities, persons with functional illiteracy, L2 learners as well as persons with dementia, prelingual hearing impairments, and aphasia (Bredel and Maaß, 2016). As a natural language processing task, automatic text simplification (ATS) has increasingly gained traction in recent years (Štajner, 2021). However, there is no consensus on best practices for evaluating simplified texts, resulting in inconsistencies in the methods used (Grabar and Saggion, 2022). Most commonly, automatic evaluation metrics are used, which have been shown to be unreliable (Alva-Manchego et al., 2021).

Studies that involve human evaluation typically employ experts or crowdworkers to rate different aspects of the output text such as simplicity, fluency, and adequacy on Likert-style scales (Štajner, 2021). However, those approaches have several shortcomings: first, they are not representative of the primary target groups of simplified texts. Second, they do not include other stakeholders such as post-editors. Third, they heavily rely on subjective ratings, which may not be indicative of the functionality of the simplified texts, i.e., enhanced comprehensibility.

In this paper, we contribute to the current debate on best practices for human evaluation by exploring different ways of measuring the quality of automatically simplified texts. Our methods span the quantitative to the qualitative, the subjective to the objective, and our raters range from Easy Language professionals to end users. Specifically, we conduct three evaluation studies: an end-user comprehensibility evaluation (Section 3), a post-editing productivity study (Section 4), and an expert evaluation (Section 5). Finally, we discuss the benefits of such multi-faceted evaluations of ATS and provide recommendations for future work.

2. Background and Related Work

2.1. Human Evaluation of Text Simplification

In terms of human evaluation, previous research has primarily relied on Likert-scale ratings of simplicity, fluency, and adequacy or meaning preservation for evaluating the quality of ATS output (Al-Thanyyan and Azmi, 2021; Stodden, 2021; Ryan et al., 2023; Martin et al., 2022; Štajner and Nisioi, 2018; Mallinson et al., 2020). The raters in these studies are typically researchers, students, or crowdworkers.

Štajner (2021) argued that evaluating ATS output quality should include the usability by target readers. However, evaluation including target groups of Easy Language are rare. Notable ex-

Proceedings of the HumEval @LREC-COLING 2024, pages 71–80
Turin, Italy. May 21st, 2024. ©2024 European Language Resources Association (ELRA)

ceptions include studies involving deaf and hard-of-hearing adults (Alonzo et al., 2021), persons with intellectual disabilities (Huenerfauth et al., 2009; Saggion et al., 2015) or dyslexia (Rello et al., 2013b,a,c), and language learners (Crossley et al., 2014). In some cases, comprehensibility is assessed based on comprehension tests, e.g., using multiple-choice questions (Leroy et al., 2013, 2022; Fajardo et al., 2014; Charzyńska and Dębowski, 2015; Alonzo et al., 2021), cloze tests (Charzyńska and Dębowski, 2015; Redmiles et al., 2019), or free recall questions (Leroy et al., 2013, 2022). More rarely, measurements of reading behavior such as reading speed (Alonzo et al., 2021; Crossley et al., 2014; Saggion et al., 2015; Rello et al., 2013a), scrolling interactions (Gooding et al., 2021), or eye movements (Rello et al., 2013a,c) are obtained.

2.2. Evaluation of Post-editing Effort

The widespread use of post-editing in interlingual translation has spurred significant research interest in how translators engage in this task and the level of effort involved. Since Krings' (2001) seminal work, it has been widely recognized that post-editing effort encompasses three main dimensions: temporal, technical, and cognitive (cf. Alvarez-Vidal and Oliver, 2023). Temporal effort is easily quantifiable and directly influences productivity and is thus used to determine translators' post-editing rates. Technical effort pertains to the editing actions performed during post-editing, such as text productions, text eliminations, replacements, and shifts, often analyzed using keylogging data and specialized software. Finally, cognitive effort refers to the mental processes underlying post-editing, even when no tangible changes are made to the raw machine translation (MT) output. Measuring cognitive effort is challenging due to its complexity, 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. To the best of our knowledge, the present paper represents the first evaluation of post-editing effort for text simplification.

3. End-user Evaluation

In this section, we describe an evaluation involving two groups of end users (with and without intellectual disabilities). We measured text comprehensibility with comprehension questions and perceived difficulty of automatically simplified German texts and compared those measurements to the original (non-simplified) source texts and manually created reference simplifications of those texts. The end-user evaluation was already described in more detail in Säuberli et al. (2024) and will only be summarized here.

3.1. Materials and Methods

3.1.1. Texts and Comprehension Questions

The texts we used in this study are part of a parallel corpus of original and simplified German texts. The corpus was made available to us by a commercial 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 political advertisements. Their lengths range between 100 and 600 words.

Each text exists in three versions: (1) the original source text, (2) a reference simplification, which was manually created by the provider, and (3) an automatically simplified version. We generated the latter with a transformer-based model fine-tuned on data from the same parallel corpus using the approach described in Rios et al. (2021).

Based on the source and reference texts, we created four multiple-choice comprehension questions for each of the 12 texts. One of the questions was about the overarching topic of the text (with four answer options), while the remaining three asked about specific details in the text (with three answer options each).

Since the ATS model sometimes omits information from the source text, and the comprehension questions were written only based on the source and reference texts, some of the questions are not answerable based on the automatically simplified version. Therefore, we added a fourth answer option "Information does not appear in the text" to the detail questions.

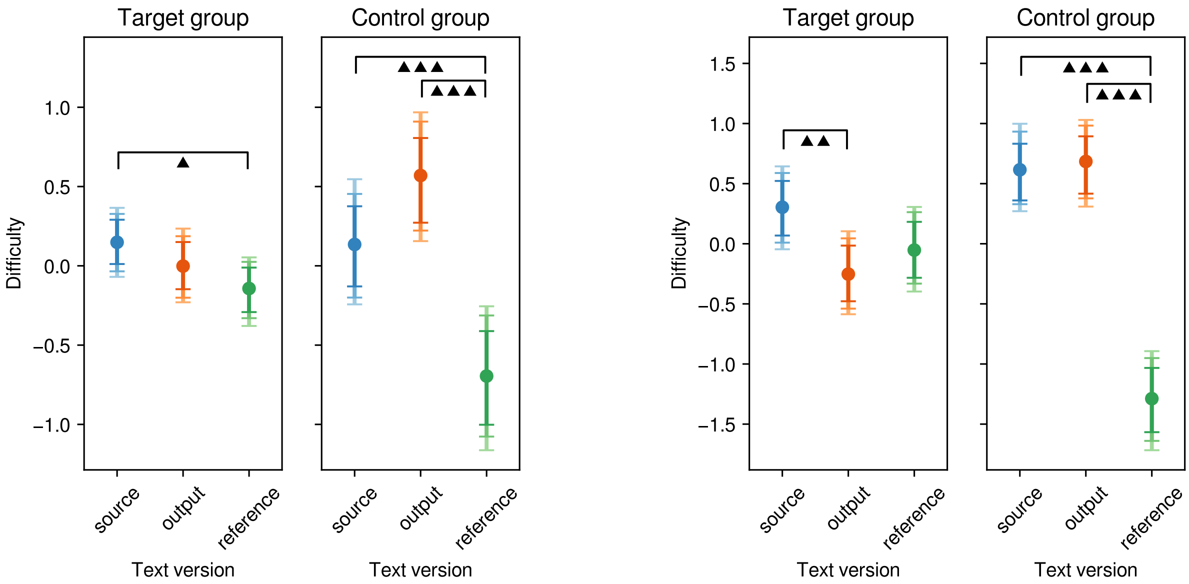
3.1.2. Participants

To compare comprehensibility among different populations, we recruited two groups of participants. The target group consisted of 18 persons with intellectual disabilities, i.e. a primary target group of Easy Language. The control group consisted of 18 native German speakers without intellectual disabilities. All participants took part on a voluntary basis and were compensated monetarily.

3.1.3. Procedure

Data collection was conducted using a mobile app which allowed participants to read and rate the texts and answer the comprehension questions. The texts were randomly assigned to participants such that each participant read exactly one version of each of the 12 texts.

After reading a text, participants were asked to rate the difficulty of the text on a five-point scale



(a) Comprehension question responses

(b) Perceived difficulty ratings

Figure 1: Difficulty estimates of the text versions based on responses to the comprehension questions and ratings. Points are posterior medians, error bars are 80%, 90% and 95% credible intervals (CI). A bracket with ▲ indicates that the 80% CI of the difference between the two parameters does not include zero (i.e., we are 80% confident that there is a difference). Similarly with ▲▲ for 90% CI and ▲▲▲ for 95% CI.

(from “very difficult” to “very easy”). The text was then shown again, and the comprehension questions were displayed at the same time. Apart from the responses to the rating and comprehension questions, we also collected behavioral data, including reading speed and response time.

3.2. Results

In this section, we will highlight the results for comprehension questions and difficulty ratings. Refer to Säuberli et al. (2024) for more detailed results.

To estimate the difference in comprehensibility between the three text versions based on the responses to the comprehension questions and the ratings, we applied Bayesian one-parameter logistic item response models (also known as Rasch models; Fox, 2010) and modeled the difficulty of the text version as an additional latent trait (cf. Linacre, 1989). We used separate models for each group.

The difficulty estimates based on the comprehension questions and the difficulty ratings are shown in Figure 1. For the target group, the ATS output was not significantly different from the source or the reference in terms of difficulty. For the control group, the output was even slightly more difficult than the source. The difficulty ratings also show remarkable differences between the two groups. While the target group rated the ATS output as being simpler than the source text, the control group’s ratings suggest that the output was equally difficult as the source and significantly more difficult than the reference.

3.3. Discussion

Several remarkable differences can be observed between the results of the two groups. First, the estimated effects are smaller and more uncertain in the target group. This is likely due to the heterogeneity of the target group, but also due to noisier data. The behavioral measurements show that reading speeds varied widely within the target group, suggesting that some participants did not read the texts carefully before rating, leading to less predictable responses. This interpretation is supported by the fact that the target group’s difficulty ratings did not differ significantly between the source and reference texts, while the control group’s ratings did.

4. Post-editing Evaluation

4.1. Methods

In our post-editing (PE) evaluation, we observed human translators post-editing the output of the automatic text simplification model (cf. Section 3.1.1) and we quantified PE effort. The following methods were employed: (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' unfolding typing process in manual simplification (MS) vs. PE tasks; (c) a post-task questionnaire to investigate how the participants rated their productivity during the tasks; and, finally, (d) a comparative analysis of production time and effort required.

In line with common practice in translation process (cf. Alves, 2003; Kappus and Ehrensberger-Dow, 2020) and post-editing research (cf. Krings, 2001; Alvarez-Vidal and Oliver, 2023), several quantitative measures were used to determine the effort involved in manually simplifying the eight source texts and in post-editing the corresponding automatically simplified target texts. More specifically, effort was quantified in terms of task duration, number of keyboard and mouse-based user events (as a measure of addition, change, regression or navigation), number of cognitive pauses (i.e., pauses with duration greater than 2000 milliseconds), and total pause time.

4.2. Participants and Procedure

Four German-speaking professional translators from a commercial provider of text simplification services were recruited through self-selection sampling. Each participant was given detailed step-by-step instructions to perform two MS and two PE tasks in their workplace. Eight source texts were used in this phase of the study. Texts were selected from an *ad-hoc* pool of texts used for all human evaluations (cf. Section 3.1.1). To prevent bias during the PE activity, each participant manually simplified and post-edited two different pairs of texts. Keystroke logging (GenoGraphiX-Log; cf. Caporossi et al., 2023) and screen recording of both processes were employed.

4.3. Results

Data from the pre-task questionnaire showed that participants (P) had three to five years of professional experience in text simplification and various degrees of expertise in interlingual MT and PE, with responses spreading evenly across choices (from 'no experience' to '3-5 years' of experience). On the other hand, participants' background in automatic intralingual text simplification and PE was significantly lower, with three out of four respondents having less than one year of experience.

Table 1 shows the total and mean values for each measure of effort compared between the MS and PE tasks. 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 1, no significant difference could be determined for any of the effort measures considered. This means that the statistical data do not suggest any significant decrease in effort in either MS or PE activity.

However, it should be emphasized that borderline statistical values relating to three user events categories were extracted (see Table 2 for a descriptive user events analysis). The mean number of text productions (i.e., textual inputs) in the MS tasks was higher (i.e., 2105, range of 952–3429) than the mean number of text productions in the PE tasks (i.e., 1417, range of 513–3043; $t = 1.77$; $p = 0.06$). On the other hand, the mean number of cursor navigations (i.e., navigation key presses) in the MS tasks was lower (i.e. 118, range of 4–548) than the mean number of cursor navigations in the PE tasks (i.e. 724, range of 4–2265; $t = 1.85$; $p = 0.05$). Similarly, study participants in the MS tasks made on average fewer mouse clicks (i.e., 55, range of 2–92) than they did in the PE tasks (i.e., 94, range of 7–226; $t = 1.81$; $p = 0.09$).

In the post-task questionnaire, respondents were asked to rate their perceived productivity on a 5-point Likert scale (5 = very high). Self-assessed productivity reached an average rating of 4.25 in the MS tasks. In their comments, respondents reported that they could generally maintain a high concentration during the task and that the source texts were "readily comprehensible". In the PE tasks, self-assessed productivity reached an average rating of 3.00. Three out of four respondents were unanimous in pointing out that "cognitive pauses" were often necessary, ultimately affecting productivity. Despite admitting that the automatic output provided a helpful "rough structure" and seemingly good translation solutions, respondents reported that the target texts lacked coherence. In addition, they emphasized that a painstaking source-target comparison was necessary to validate the adequacy of the automatic output, which resulted in higher time expenditure and lower productivity. On the other hand, one respondent stated that the post-editing activity required little effort overall, as the source texts were "relatively easy".

4.4. Discussion

The mean productivity value that participants self-reported was 1.25 points higher in the MS tasks than in the PE tasks. Nonetheless, our statistical analysis did not suggest any significant increase in productivity in either manual simplification or post-editing activity. Factors that may have affected the results include participants' main expertise in manual simplification (vs. post-editing). Furthermore, a different working environment – that is, the use of keystroke logging software to perform the MS and PE tasks – may have had an impact on the participants' translation activity and/or their perceived productivity.

Effort measure	Mean (per P per task) MS	PE	Student's t-test
Task duration (h:m:s)	00:21:15	00:22:14	p = 0.43 | t = 1.78
User events	2760*	2556	p = 0.39 | t = 1.79
Number of cognitive pauses	87.57*	88.25	p = 0.48 | t = 1.78
Pause time (h:m:s)	00:18:47*	00:19:22	p = 0.45 | t = 1.77

Table 1: Total and mean values for each measure of effort compared between manual simplification (MS) and post-editing (PE) tasks (* P1 completed both MS tasks but did not submit keystroke logging data for the second MS task). Right column: statistical analysis of each measure of effort in manual simplification (MS) and post-editing (PE) tasks according to two-sample Student's t-tests assuming unequal variances.

	P1		P2		P3		P4	
User events	MS*	PE	MS	PE	MS	PE	MS	PE
Text productions	1708	1092	5697	3064	2772	1308	4559	4665
Text eliminations	87	428	894	767	321	122	773	1450
Cut events	0	1	3	7	2	1	2	3
Copy events	0	2	1	1	0	3	1	0
Paste events	0	4	6	8	2	6	12	4
Cursor navigations	153	3171	16	66	95	9	567	2553
Mouse events	2	23	156	274	80	124	150	333
Misc. events (e.g., modifier keys)	133	136	567	336	181	83	383	407
Total user events	2083	4857	7340	4523	3453	1656	6447	9415

Table 2: Events analysis per study participant (P) in manual simplification (MS) and post-editing (PE) tasks. Combined values (i.e., two MS tasks and two PE tasks) per study participant (* P1 completed both MS tasks but did not submit keystroke logging data for the second MS task).

5. Expert Evaluation

5.1. Methods

In our expert evaluation, we obtained translation quality ratings from experts in German Easy Language translation. In this phase, we employed an online evaluation questionnaire in which four evaluators performed a source-based direct assessment (cf. Graham et al., 2013; Federmann, 2018) of the target texts. The questionnaire was developed with LimeSurvey[1] and comprised eleven items, of which three collected professional background data, and eight presented two parallel texts each, i.e. one source text and one corresponding target text. For each source text used in the post-editing productivity study (cf. Section 4.2), four corresponding simplified versions were employed, i.e. one reference text, one automatically simplified text, one manually simplified text, and one post-edited text – the latter two being produced during the post-editing study (cf. Section 4). Based on the experts' evaluations, the end quality of the experimental units was then analyzed and compared.

5.2. Participants and Procedure

Four Swiss-based German-speaking experts in Easy Language translation were recruited through purposive sampling. A 4x8 Latin square gave us a total of 32 experimental units and secured an unbiased response. Evaluators were asked to assign simplicity, adequacy, and fluency scores on 5-point scales (5 = maximal quality; cf. Grabar and Saggion, 2022) to each target text (see Table 3) and, if desired, insert comments. Evaluators were not provided with any information about how the target texts had been produced.

5.3. Results

All four evaluators had over five years of professional experience in text simplification and regularly provided a wide portfolio of Easy and Plain Language services, including intra- and interlingual translations, text production, and text validation in collaboration with the target groups. A comparison of the average simplicity ratings for each target text category shows that the manually simplified texts produced during the post-editing productivity study (cf. Section 4) were rated higher (i.e., 4.38, range of 4–5) than the other three categories (see Table 4). A similar pattern emerged for the adequacy and fluency ratings: the manually simplified texts

[1] https://www.limesurvey.org/

Rating	Simplicity Q1: How does the target text differ from the original text?	Adequacy Q2: Does the target text reflect the content of the original text?	Fluency Q3: Is the target text fluent and grammatical?
5	much easier	completely	fluent
4	easier	mostly	mostly
3	equally difficult	partially	partially
2	more difficult	mostly not	mostly not
1	much more difficult	not at all	not at all fluent

Table 3: Simplicity, adequacy and fluency scales used in the expert evaluation questionnaire (Q = question).

Target texts	Simplicity	Adequacy	Fluency
RT	4.25	3.75	3.75
AS	3.38	3.63	3.13
MS	4.38	4.25	4.25
PE	4.25	4.25	3.50

Table 4: Average simplicity, adequacy, and fluency ratings for each target text category (RT = reference target text, AS = automatically simplified text, MS = manually simplified text, PE = post-edited text).

consistently obtained the highest average ratings (i.e., 4.25, range of 3–5), while the automatically simplified texts were assigned the lowest average ratings (i.e., 3.63, range of 1–5, and 3.13, range of 2–4, respectively) (see Table 4).

The ratings of simplicity, adequacy, and fluency are consistent with the experts' comments in the evaluation questionnaire, in which seven out of eight experts reported finding the automatically simplified texts mostly not adequate. Simplification techniques were also considered to be only partially effective. On the other hand, the manually simplified texts were often reported as being "very good", even though it was also emphasized that they did not consistently comply with German Easy Language guidelines. As for the post-edited texts, most evaluators remarked on several simplicity as well as adequacy issues. Refer to Appendix A for examples of expert comments.

5.4. Discussion

The outcomes indicate that the automatic text simplification model under examination is not ready for deployment with or without post-editing, mainly due to weak simplification capabilities. As previously highlighted, the automatically simplified texts obtained, in fact, the lowest ratings across all evaluation metrics. In retrospect, it would have been beneficial to collect additional background data to identify the specific sets of Easy Language guidelines that experts commonly employed and referred to in their evaluations. Such data could have provided support for both the quantitative (i.e., ratings) and qualitative (i.e., comments) findings.

6. Overall Discussion

6.1. Advantage of Multi-stakeholder Involvement

Involving multiple stakeholders in ATS processes and assessments holds significant value for ensuring the ultimate quality and functionality of simplified texts. In our study, experts played a pivotal role by evaluating the adherence of texts to established guidelines, thereby offering critical insights into the simplicity, adequacy, and fluency of the simplified content (cf. Section 5). Conversely, post-editors contributed valuable feedback regarding productivity gains (cf. Section 4). Additionally, the perspectives of end users were indispensable for gauging the comprehensibility and acceptability of simplified texts in real-world contexts (cf. Section 3). It is crucial to acknowledge that linguistic complexity pertains to individual cognitive costs (Hansen-Schirra et al., 2020; Pallotti, 2015), and that text simplification efforts cater to highly diverse target groups. Hence, the active involvement of target audiences and the consideration of individual variability are paramount in optimizing the effectiveness and inclusivity of text simplification.

6.2. Advantage of Mixed-method Approaches

Mixed-method approaches offer several advantages in ATS studies. Given the inherent challenge of directly measuring reading comprehension behaviors, triangulating multiple proxies becomes imperative. The use of rating scales poses similar challenges, as interpretations may vary among participants (Stodden, 2021). Equally, qualitative findings regarding end users' perceptions of complexity may diverge from quantitative metrics (Säuberli et al., 2024; Carrer, 2021; Benson-Goldberg et al., 2024), highlighting the importance of discerning discrepancies between perception and actual comprehension. Therefore, the adoption of mixed-method approaches not only enhances the robustness of research findings but also enables a more nuanced exploration of complex behaviors.

7. Conclusions

We conducted an extensive evaluation study involving end users, post-editors, and experts as stakeholders, and using a combination of quantitative/qualitative and objective/subjective methods. The results showed that there are differences in comprehensibility and perception of simplified texts between different user groups. We also found that qualitative surveys and behavioral observations can be essential in interpreting the results. These differences need to be accounted for in human evaluations of ATS models. Specifically, the following recommendations emerged from our experiments:

- Whenever possible, include target readers to assess comprehensibility.
- Do not rely solely on perceived quality ratings and assess the quality and functionality of the output as directly as possible, e.g., by measuring comprehension or post-editing effort.
- Collect qualitative data (e.g., through surveys or interviews) and behavioral measurements (e.g., while reading or post-editing) to support the interpretation of quantitative results.
- When collecting expert ratings, clearly define the Easy Language guidelines to be taken as a reference and ask the evaluators to specify which rules were not observed.

As ATS research begins to harness the new potential of large language models (Kew et al., 2023), future research should adopt a more human-centric and holistic approach to evaluation. We believe that this is essential for ensuring that the technological advancements yield tangible benefits for the end users of those technologies.

8. Acknowledgements

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

9. References

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. 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. *Ampersand*, 11:100125.

Fabio Alves. 2003. *Triangulating Translation: Perspectives in process oriented research*. John Benjamins Publishing Company, Amsterdam.

Sofia Benson-Goldberg, Lori Geist, and Karen Erickson. 2024. Simplified COVID-19 guidance for adults with intellectual and developmental disabilities. *Journal of Applied Research in Intellectual Disabilities*, 37(3).

Bettina M. Bock and Sandra Pappert. 2023. *Leichte Sprache, Einfache Sprache, verständliche Sprache*. Narr Francke Attempto, Tübingen.

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

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. 2021. Translating into Easy Italian : an analysis of health-related texts and their impact on comprehension by people with intellectual disabilities. Thesis: Master, ZHAW Zürcher Hochschule für Angewandte Wissenschaften, Winterthur.

Edyta Charzyńska and Łukasz Jerzy Dębowski. 2015. Empirical verification of the polish formula of text difficulty. *Cognitive Studies*, 15:125–132.

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.

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.

Jean-Paul Fox. 2010. *Bayesian Item Response Modeling: Theory and Applications*. Springer New York, New York, NY.

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.

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

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.

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.

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.

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.

Gabriele Pallotti. 2015. A simple view of linguistic complexity. *Second Language Research*, 31(1):117–134.

Elissa Redmiles, Lisa Maszkiewicz, Emily Hwang, Dhruv Kuchhal, Everest Liu, Miraida Morales, Denis Peskov, Sudha Rao, Rock Stevens, Kristina Gligorić, Sean Kross, Michelle Mazurek, and Hal Daumé III. 2019. Comparing and developing tools to measure the readability of domain-specific texts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4831–4842, Hong Kong, China. Association for Computational Linguistics.

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.

Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A new dataset and efficient baselines for document-level text simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161, Online and in Dominican Republic. Association for Computational Linguistics.

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 simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Trans. Access. Comput.*, 6(4).

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).

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.

Andreas Säuberli, Franz Holzknecht, Patrick Haller, Silvana Deilen, Laura Schiffl, Silvia Hansen-Schirra, and Sarah Ebling. 2024. Digital comprehensibility assessment of simplified texts among persons with intellectual disabilities.

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.

A. Expert comments

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

A.1. Automatically simplified text (AS)

7 out 8 experts reported finding the AS texts mostly not adequate.

- “The TT contains some ‘information vagueness”’
- “In terms of content, these are two different texts”
- “The simplification is mainly achieved through a different text structure”
- “This text is not in Easy Language. At most, it is a shortened version of the original text”

A.2. Manually simplified text (MS)

Often reported as being "very good".

- “very good”
- “One of the better texts in this questionnaire”
- “Again, this text is not in Easy Language, although it does comply with many of the rules”
- “The rules are not consistently adhered to”

A.3. Reference text (RT)

More in line with Plain Language properties.

- “One of the better texts here and much easier to understand for the target group”
- “The target text combines two different language levels”
- “Here too: Target text is no Easy Language” / “This text is not in Easy Language”

A.4. Post-edited text (PE)

Experts reported simplicity/adequacy issues.

- “The target text is not written in Easy Language: Several rules are not observed”
- “Significant reduction in content”
- “The TT still contains some difficult words”