Exploring Cognitively Informed Sentence Simplification with Gaze-Guided Text Generation

Andreas Säuberli[1,2] **Diego Frassinelli**[1] **Barbara Plank**[1,2]
[1]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany
{andreas.saeuberli, diego.frassinelli, b.plank}@lmu.de

Abstract

Automatic text simplification has mostly relied on human judgments when it comes to what is considered easy or difficult to read. Eye movements while reading can offer a more direct and objective signal of processing effort and reading ease. In this paper, we explore gaze-guided text generation (GGTG), an approach to control reading ease in generated texts, and assess its use for sentence simplification. GGTG employs a gaze model that is trained to predict eye-tracking measures such as reading times or regression rates, which are then used to rerank next-token probabilities generated by a language model. We evaluated the approach on an English sentence simplification benchmark and found gains in automatic evaluation metrics, although the simplification operations are mostly limited to the lexical level. Its modular nature also allows GGTG to be combined with other simplification techniques such as prompting or fine-tuning.

Keywords: text simplification, eye tracking, cognitive modeling, controlled text generation

1. Introduction

Most research in automatic text simplification has relied on human judgments of simplicity or manual reference simplifications that are often crowdsourced (Coster and Kauchak, 2011; Paetzold and Specia, 2017; Alva-Manchego et al., 2020; Grabar and Saggion, 2022). This means that the human intuition of what *should* be considered easy to read is taken as a proxy of what is *actually* easy to read. However, recent research has shown that there can be a substantial discrepancy between subjective perception of difficulty and actual comprehension, and that both aspects can vary between different user groups (Alonzo et al., 2021; Carrer et al., 2024). In contrast, methods such as eye tracking can provide a more direct and objective signal of the processing effort required to read and understand a text (Just and Carpenter, 1980). While eye-movement data has been used to *evaluate* simplified texts, using this cognitive signal directly to *generate* simplifications remains underexplored. At the same time, the increasing amount of available eye-tracking-while-reading data, including multilingual corpora (Siegelman et al., 2022; Jakobi et al., 2025) and datasets involving diverse readers such as non-native or dyslexic readers (Kuperman et al., 2023; Siegelman et al., 2025; Hollenstein et al., 2022; Reich et al., 2024) makes it more feasible now to use such data for natural language processing (NLP) applications like text simplification.

Recent work by Säuberli et al. (2026) proposed **gaze-guided text generation** (GGTG) as a simple yet effective way to integrate eye-tracking data into the text generation process. It works by training a gaze model to predict eye-tracking measures and using these predictions to modify the next-token probabilities generated by a language model (LM). Their findings suggest that the level of control that can be achieved with this approach may be limited to shallow features affecting lexical processing, such as word length and frequency. In this paper, we explore to what extent GGTG can be used to simplify sentences, and whether the method can be pushed to induce more complex simplification operations at the syntactic level.

We build on and extend the experiments in Säuberli et al. (2026) in several ways:

- We explore five eye-tracking measures associated with different levels of processing and their ability to capture text complexity.
- We train gaze models that ignore word length and frequency and focus on higher-level aspects of text complexity.
- We evaluate GGTG on the ASSET benchmark for sentence simplification.

2. Related work

Controllable text simplification has been approached from several angles. Most prominently, models have been trained with special tokens or feature vectors to control characteristics like word frequency, dependency tree depth, and readability level (Scarton and Specia, 2018; Martin et al., 2020, 2022; Agrawal and Carpuat, 2023). Nishihara et al. (2019) used a lexical constraint loss to control lexical complexity. Kew and Ebling (2022) is the most similar to our approach. Instead of using eye-tracking data, they trained classifiers that predict the level of difficulty for next token candidates and modified the token probabilities accordingly.

Proceedings of The Second International Workshop onEye-Tracking Resources and Evaluation for Human-Aligned NLP, pages 16–23
May 12, 2026.

While cognitive data such as eye movements have been used to *evaluate* simplified texts or predicting readability (Rello et al., 2013; Singh et al., 2016; Vajjala et al., 2016; Ivchenko and Grabar, 2024; Gruteke Klein et al., 2025a,b), the present work is, to the best of our knowledge, the first to use gaze data directly in the simplification process.

3. Methods

3.1. Gaze-guided text generation (GGTG)

At its core, GGTG involves an ensemble of an off-the-shelf **language model** and a **gaze model**, which we train to predict word-level eye-tracking measures (e.g., first fixation time or regression rate). The LM predicts candidates for the next token, which are then re-ranked by the gaze model. The strength of the influence of the gaze model can be controlled via a **gaze weight**. This way, the LM output can be steered towards eliciting specific reading behaviors (e.g., longer/shorter fixation times or higher/lower regression rates), thereby manipulating reading ease.

We applied beam search with a beam size of 8 to decode simplified texts. We used the implementation by Säuberli et al. (2026) and refer to the corresponding paper for more details.

3.2. Language model and prompts

We chose the instruction-tuned Llama 3.2 model (3B; Grattafiori et al., 2024), as it is a small and efficient model with strong instruction-following performance. For the simplification task, we experiment with two different prompts. The first prompt instructs the LM to *paraphrase* the source sentence without changing its meaning, allowing us to assess the simplifying effect of the gaze model alone. The second prompt instructs the LM to *simplify*, to test whether GGTG still has a simplifying effect. See Appendix A for the precise wording.

3.3. Gaze models

3.3.1. Predicted eye-tracking measures

We selected five word-level eye-tracking measures that can plausibly be predicted from preceding context only (which is a requirement for autoregressive generation):

- **First fixation duration**: the duration of the first fixation on a word during the first pass (0 if the word is skipped in the first pass).
- **First-pass reading time**: the sum of all fixation durations on a word during the first pass (0 if the word is skipped in the first pass).
- **Go-past time**: the sum of all fixation durations from when the word was first fixated until the gaze moves past the word for the first time. This includes regression paths initiated on the word during first-pass reading.
- **First-pass skipping rate**: 1 if the word was skipped in the first pass, 0 otherwise.[1]
- **First-pass regression rate**: 1 if there was a regression during the first pass, 0 otherwise.

All measures are computed for readers individually and then averaged across readers for each word. The final measures are normalized to have a mean of 0 and a standard deviation of 1.

According to psycholinguistic research, some of these measures are associated with earlier cognitive processes such as word recognition, while others reflect later processing such as syntactic integration (Rayner, 1998; Godfroid, 2019). For example, first fixation duration is measured when the word is first looked at, while go-past time also includes time spent *after* the first fixation. Therefore, we expect earlier measures to mainly enable lexical simplification, while later measures may allow more syntactic simplification.

We fine-tuned the large variant of GPT-2 (774M; Radford et al., 2019) to predict the eye-tracking measures listed in Section 3.3.1. We trained separate models for each eye-tracking measure.

3.3.2. Model training

All models are trained on a mix of four publicly available eye-tracking corpora of naturalistic reading that cover a range of genres and difficulty levels, listed in Table 1. All of these datasets contain texts that span multiple sentences, so we perform automatic sentence splitting after calculating eye-tracking measures and train the gaze model on individual sentences.

We trained the models on 90% of each dataset and used the remaining 10% as a validation set for early stopping and to measure performance. We ensured that all sentences from the same text are assigned to the same data split to avoid data leakage. The remaining training procedure follows Säuberli et al. (2026). Performance on the validation set is reported in Table 2.

3.3.3. Residual models

Säuberli et al. (2026) found that their gaze model's reading time predictions were dominated by shal-

[1] First-pass skipping rate is the only measure for which higher values are associated with better reading ease. Therefore, for ease of interpretation, we flip the sign of the skipping rates, so that lower numbers can be considered better across all measures.

Dataset	Text genre/content	# words	# readers
EMTeC (Bolliger et al., 2025)	LLM-generated; various genres	50,871	(*)107
OneStopQA (Berzak et al., 2025)	Original and simplified news	35,164	(*)360
MECO-L1 English (Siegelman et al., 2022)	Encyclopedic information	2,107	46
Provo (Luke and Christianson, 2017)	Various genres	2,743	84

Table 1: Datasets used for training the gaze models. (*) means not every text is read by every reader.

Eye-tracking measure	GPT-2	LR	LR + GPT-2 residual
First fixation duration	**0.593**	0.546	0.587
First-pass reading time	**0.636**	0.596	0.633
Go-past time	**0.481**	0.367	0.458
First-pass skipping rate	0.610	0.562	**0.619**
First-pass regression rate	**0.226**	0.140	**0.226**

Table 2: Explained variance (R^2) for each gaze model, averaged across the four validation datasets. LR = linear regression.

low features such as word length and frequency. To avoid this, we trained a second version of each model that does *not* capture the variance associated with these features. We did this by first fitting a linear regression model to predict the eye-tracking measures based word length and word frequency alone.[2] Next, we computed the residuals of the linear regression model on the training data, normalized them to mean 0 and standard deviation 1, and trained the GPT-2-based gaze model to predict these residuals.

3.4. Evaluation

We evaluated our approach on ASSET (Alva-Manchego et al., 2020), an established sentence simplification benchmark for English with multiple crowdsourced references. We report results on the validation set, which consists of 2,000 sentences with ten human reference simplifications each. We consider gaze weights in the range from 0 (gaze model deactivated) to −3 (decrease eye-tracking measure).

We report the reference-based evaluation metrics SARI (Xu et al., 2016) and LENS (Maddela et al., 2023) against all ten references, as well as the reference-free metric LENS-SALSA (Heineman et al., 2023). We also report the effects on lexical and syntactic features such as word frequency and dependency tree depth.[3]

[2]Word length was calculated as the number of characters. Word frequency is measured on the Zipf scale based on the *wordfreq* package (Speer, 2022). Linear regression models were fitted with *scikit-learn* (Pedregosa et al., 2011).

[3]Dependency trees were generated using Stanza (Qi et al., 2020).

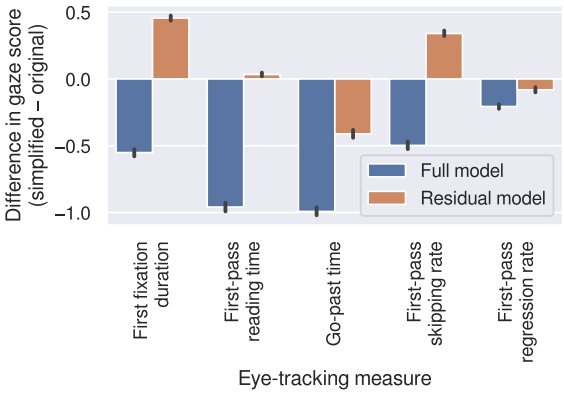

Figure 1: Mean gaze score differences between original and simplified reference texts in the ASSET validation set. A negative value means that the gaze model predicts lower eye-tracking measures for the simplified version (which is the expectation). Error bars show 95% confidence intervals.

4. Results and discussion

4.1. Do the gaze models capture sentence complexity?

The GGTG approach can only work for sentence simplification if the gaze model is able to discriminate simple from complex sentences. While the gaze models described in Section 3.3 are indirectly trained to capture complexity by predicting eye-tracking measures, this does not necessarily translate into a useful model of complexity in general. Moreover, it is unclear which eye-tracking measures are suitable for sentence simplification.

Therefore, as a first step, we assess whether the gaze scores predicted by our models differ between original and simplified sentences in ASSET. These differences are visualized in Figure 1. For the full models trained to predict eye-tracking measures, we observe lower gaze scores in the simplified versions on average, with the largest differences for first-pass reading time and go-past time. In contrast, the models that were trained on the linear regression residuals consistently predict smaller differences, or even differences in the opposite direction. This is expected, as these models do not have access to some of the most salient predictors

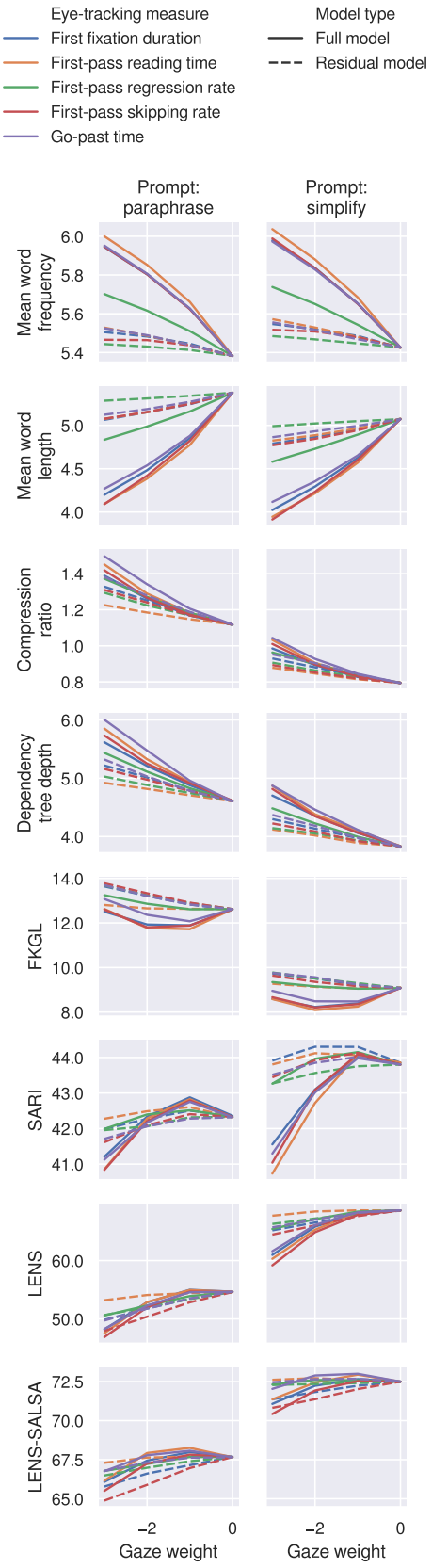

Figure 2: Effect of gaze models on generated texts and evaluation metrics. A gaze weight of 0 indicates generation without a gaze model, a negative gaze weight means steering the language model to decrease the corresponding eye-tracking measure.

of simplicity – word length and frequency. However, go-past time and first-pass regression rate – both associated with later cognitive processing – are still predicted to be lower in the simplified texts. This suggests that these two residual models capture some aspects of simplicity in the ASSET dataset that go beyond word length and frequency, possibly features at the syntactic or semantic level.

4.2. How does GGTG affect the generated texts?

Figure 2 shows how changing the gaze weight for the different gaze models affects the output texts, comparing full vs residual models (see Section 3.3.3) and the two prompt types (see Section **??**). As expected, word length and frequency (plots in first two rows) are strongly affected by all full models, but less so by the residual models. Compression ratio increases with stronger gaze weights, indicating that output sentences tend to become longer. This is due to the fact that during beam search, appending tokens that decrease the overall gaze score is preferred over stopping the generation (which would mean that the gaze score remains unchanged). Dependency trees also tend to increase in depth, likely due to the increased sentence length. This suggests that syntactic nesting is not reduced in the simplified sentences.

Flesch-Kincaid Grade Level (FKGL; Kincaid et al., 1975) estimates a text's readability using the number of sentences, words, and syllables as surface-level proxies. We observe (plots in fifth row) that almost all full models decrease FKGL (i.e., improve readability) with gaze weights -1 and -2, but this effect is negated by the increase in sentence length around gaze weight -3. SARI and LENS-SALSA slightly improve with gaze weight -1, but quickly decrease with lower weights.

Overall, there is no clear pattern regarding the different eye-tracking measures. First-pass reading time appears to have the most consistent positive effect on evaluation metrics, which may be explained by this gaze model's relatively strong performance (see Table 2). Residual models generally have weaker effects, but in the case of first-fixation duration and first-pass reading time, improvements in SARI can still be observed. As expected, the *simplify* prompt yields better readability and evaluation metrics than the *paraphrase* prompt, but even here, GGTG can further improve evaluation metrics. See Table 3 for example outputs.

5. Conclusion

The appeal of using eye-tracking data for text simplification is that, as a cognitive signal, it reflects reading ease more directly than human judgments

Version	Text
Source	A Georgian inscription around the drum attests his name.
LM-only	An inscription on the drum confirms his name.
Reading time -1	There is an inscription on the drum with his name on it.
Reading time -2	The name of the person is written on a drum.
Regression rate -1	There is an inscription on the drum that confirms his name.
Regression rate -2	The name of the person is mentioned in an inscription on a drum.

Table 3: Example outputs with the *simplify* prompt. Reading time -1 indicates that the gaze model predicts first-pass reading time and a gaze weight of -1 is used.

or manually simplified texts. Our results show that some simplification operations can be achieved by applying GGTG, and that small improvements in automatic evaluation metrics can be achieved, even if the LM is already explicitly prompted to simplify.

However, more complex operations beyond the lexical surface level remain a challenge, even for the residual models, which are trained to focus on less superficial features. A reason for this challenge could be the training data for the gaze models. There is a growing amount of available eye-tracking data, but extracting more subtle effects and patterns from naturalistic reading corpora is difficult. In contrast, psycholinguistic research commonly uses minimal pairs to measure these effects. Leveraging these resources could also be helpful for text simplification.

In sum, while GGTG can help at least at a superficial, lexical level, applying it on its own is not yet sufficient in reality. However, thanks to the modularity of the approach, it is easily possible to combine it with other techniques, including prompting and fine-tuning.

Limitations

Automatic evaluation. Automatic evaluation metrics can only measure the adequacy and difficulty of simplified texts to a very limited degree, and human evaluation is usually recommended (Alva-Manchego et al., 2021; Grabar and Saggion, 2022; Carrer et al., 2024). Since our work is exploratory and the number of investigated parameters would have made a comprehensive human evaluation unfeasible, we decided to prioritize automatic evaluation metrics.

English only. Our evaluation is limited to English texts, limiting the generalizability of our results to other languages. The main reason for this is the scarcity of eye-tracking data in other languages.

Number of tested models. Finally, we only considered a single base model for both the language and gaze model, limiting generalizability to other models. While we initially experimented with several base models, the results were not substantially different, so our results only include one set of relatively small models for simplicity and reproducibility.

Ethical considerations

Trustworthiness of model outputs. Given the use of large language models and the nature of our approach, there is little control over the content and semantic accuracy of the generated texts. Therefore, our method should not be used without additional safeguards and manual inspection or post-editing. This is particularly important in accessibility scenarios with potentially vulnerable end users, which is among the most common use cases of text simplification.

Reproducibility. All datasets and code libraries used in this project are open-source and received due citations. The code and data for reproducing the results and figures in this paper is available from the accompanying repository: https://github.com/mainlp/gaze-guided-sentence-simplification/

Use of generative models. We used GitHub Copilot to accelerate programming tasks. All generated code was thoroughly checked and tested. We did not use generative models for ideation, results interpretation, or paper writing.

Acknowledgements

We thank the three anonymous reviewers for their valuable feedback. This research is in parts supported by the ERC Consolidator Grant DIALECT 101043235.

References

Sweta Agrawal and Marine Carpuat. 2023. Controlling pre-trained language models for grade-specific text simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819. Association for Computational Linguistics.

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, pages 1–12. ACM.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679. Association for Computational Linguistics.

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.

Yevgeni Berzak, Jonathan Malmaud, Omer Shubi, Yoav Meiri, Ella Lion, and Roger Levy. 2025. OneStop: A 360-participant English eye tracking dataset with different reading regimes. *Scientific Data*, 12(1).

Lena S. Bolliger, Patrick Haller, Isabelle C. R. Cretton, David R. Reich, Tannon Kew, and Lena A. Jäger. 2025. EMTeC: A corpus of eye movements on machine-generated texts. *Behavior Research Methods*, 57(7).

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.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Aline Godfroid. 2019. *Eye Tracking in Second Language Acquisition and Bilingualism: A Research Synthesis and Methodological Guide*. Routledge.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, et al. 2024. The Llama 3 herd of models. *arXiv*.

Keren Gruteke Klein, Shachar Frenkel, Omer Shubi, and Yevgeni Berzak. 2025a. Surprisal takes it all: Eye tracking based cognitive evaluation of text readability measures. *arXiv*.

Keren Gruteke Klein, Omer Shubi, Shachar Frenkel, and Yevgeni Berzak. 2025b. The effect of text simplification on reading fluency and reading comprehension in L1 English speakers. *OSF*.

David Heineman, Yao Dou, Mounica Maddela, and Wei Xu. 2023. Dancing between success and failure: Edit-level simplification evaluation using SALSA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3466–3495. Association for Computational Linguistics.

Nora Hollenstein, Maria Barrett, and Marina Björnsdóttir. 2022. The copenhagen corpus of eye tracking recordings from natural reading of Danish texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1712–1720, Marseille, France. European Language Resources Association.

Oksana Ivchenko and Natalia Grabar. 2024. Study of medical text reading and comprehension through eye-tracking fixations. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 84–92, Torino, Italia. ELRA and ICCL.

Deborah Noemie Jakobi, Maja Stegenwallner-Schütz, Nora Hollenstein, Cui Ding, Ramune Kaspere, Ana Matić Škorić, Eva Pavlinusic Vilus, Stefan Frank, Marie-Luise Müller, Kristine M Jensen de López, Nik Kharlamov, Hanne B. Søndergaard Knudsen, Yevgeni Berzak, Ella Lion,

Irina A. Sekerina, Cengiz Acarturk, Mohd Faizan Ansari, Katarzyna Harezlak, Pawel Kasprowski, Ana Bautista, et al. 2025. MultiplEYE: Creating a multilingual eye-tracking-while-reading corpus. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, ETRA '25, pages 1–11. ACM.

Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

Tannon Kew and Sarah Ebling. 2022. Target-level sentence simplification as controlled paraphrasing. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 28–42. Association for Computational Linguistics.

J. Peter Kincaid, Robert P. Fishburne Jr, Richard L Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel. Technical report.

Victor Kuperman, Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Maria Da Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost, Carolina A. Gattei, Areti Kalaitzi, Kaidi Lõo, Marco Marelli, Kelly Nisbet, et al. 2023. Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus. *Studies in Second Language Acquisition*, 45(1):3–37.

Steven G. Luke and Kiel Christianson. 2017. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408. Association for Computational Linguistics.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

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.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable text simplification with lexical constraint loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2017. Lexical simplification with neural ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 34–40, Valencia, Spain. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

David R. Reich, Shuwen Deng, Marina Björnsdóttir, Lena Jäger, and Nora Hollenstein. 2024. Reading does not equal reading: Comparing, simulating and exploiting reading behavior across populations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13586–13594, Torino, Italia. ELRA and ICCL.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013. Simplify or help? Text simplification strategies for people with dyslexia.

In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, W4A '13, pages 1–10. ACM.

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

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Maria Da Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost, Carolina A. Gattei, Areti Kalaitzi, Nayoung Kwon, Kaidi Lõo, Marco Marelli, et al. 2022. Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-Movement Corpus (MECO). *Behavior Research Methods*, 54(6):2843–2863.

Noam Siegelman, Sascha Schroeder, Yaqian Borogjoon Bao, Cengiz Acartürk, Niket Agrawal, Lena S. Bolliger, Jan Brasser, César Campos-Rojas, Denis Drieghe, Dušica Filipović Đurđević, Sofya Goldina, Romualdo Ibáñez Orellana, Lena A. Jäger, Ómar I. Jóhannesson, Anurag Khare, Nik Kharlamov, Hanne B. S. Knudsen, Árni Kristjánsson, Charlotte E. Lee, Jun Ren Lee, et al. 2025. Wave 2 of the Multilingual Eye-Movement Corpus (meco): New text reading data across languages. *Scientific Data*, 12(1).

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan. 2016. Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 202–212, Osaka, Japan. The COLING 2016 Organizing Committee.

Robyn Speer. 2022. wordfreq. Zenodo.

Andreas Säuberli, Darja Jepifanova, Diego Frassinelli, and Barbara Plank. 2026. Controlling reading ease with gaze-guided text generation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2383–2397. Association for Computational Linguistics.

Sowmya Vajjala, Detmar Meurers, Alexander Eitel, and Katharina Scheiter. 2016. Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 38–48, Osaka, Japan. The COLING 2016 Organizing Committee.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

A. Prompts

***paraphrase* prompt:**

```
Paraphrase the following text.  You
may change the wording and structure
of the text, but not its meaning.
Only respond with the paraphrased
text.  Here is the text:  [...]
```

***simplify* prompt:**

```
Simplify the following text.  You
may change the wording and structure
of the text, but not its meaning.
Only respond with the simplified
text.  Here is the text:
```