THM@SimpleText 2025 - Task 1.1: Revisiting Text Simplification based on Complex Terms for Non-Experts

Nico Hofmann[1], Julian Dauenhauer[1], Nils Ole Dietzler[1], Idehen Daniel Idahor[1] and Christin Katharina Kreutz[1,2,*]

[1]TH Mittelhessen - University of Applied Sciences, Gießen, Germany
[2]Herder Institute, Marburg, Germany

Abstract

Scientific text is complex as it contains technical terms by definition. Simplifying such text for non-domain experts enhances accessibility of innovation and information. Politicians could be enabled to understand new findings on topics on which they intend to pass a law, or family members of seriously ill patients could read about clinical trials.

The SimpleText CLEF Lab focuses on exactly this problem of simplification of scientific text. Task 1.1 of the 2025 edition specifically handles the simplification of complex sentences, so very short texts with little context. To tackle this task we investigate the identification of complex terms in sentences which are rephrased using small Gemini and OpenAI large language models for non-expert readers.

Keywords

Text Simplification, Complex Term Identification, LLMs, Prompt Engineering, Personas

1. Introduction

Scientific texts are written to be read and understood by domain experts, scholars who are highly educated. Such texts are packed with abbreviations and technical terms and they need to fit within a strict page or word limit.

Over the years the SimpleText initiative [1, 2, 3, 4, 5, 6] has shown considerable efforts to help accelerate the development of approaches to make such texts more accessible to the general public. The importance of the intended target audience in text simplification efforts has been considered before [7]. For example, text simplified for children should contain short sentences [8] but should not be oversimplified to retain readers' engagement [9].

While there are approaches for simplifying text from different domains and for different target audiences [10, 11], most approaches do not consider the particularities of either [12]. Oversimplifying a text, so lowering its overall complexity to no longer fit the required complexity level of readers, is disadvantageous for all most all target audience, as it leads to disengagement [9]. Therefore, the simplicity level of texts should *fit* its intended readers.

In general, in the simplification of texts for non-expert we assume that a reader has the *appropriate language skills* to understand complex phrases and grammar but is not able to understand a text due to *missing domain knowledge*. Our idea is thus to identify complex, domain-specific terms in texts to try to only replace these complex components while maintaining the overall structure and linguistic complexity of texts.

This work tackles task 1.1 of this year's SimpleText lab [5, 6, 13], the simplification of short scientific texts for non-expert readers. This task has the challenging characteristic of only providing very little context (the sentence itself) on the content to be simplified. We try to tackle this task by basing our efforts on an earlier submission by IRGC@SimpleText'23 [14] to the SimpleText shared task focused

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
ckreutz@acm.org (C. K. Kreutz)
https://kreutzch.github.io (C. K. Kreutz)
0000-0002-5075-7699 (C. K. Kreutz)

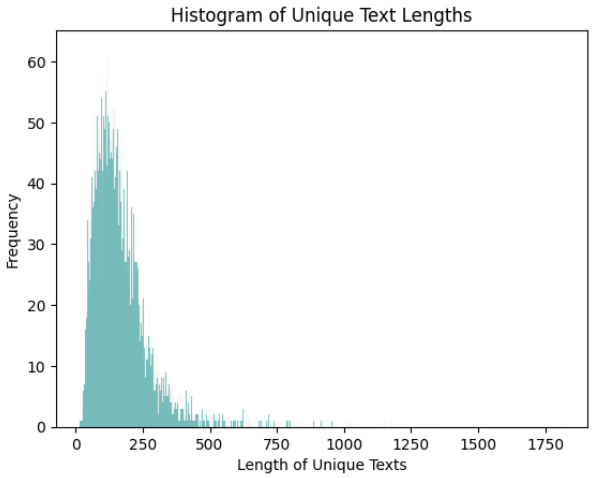

Figure 1: Overview of lengths of unique texts in the dataset.

on complex, scientific term identification and investigate different prompts while using small large language models (LLMs).

2. Dataset

The dataset for Task 1.1 provided by the SimpleText lab [5] consists of 9160 short texts in English which were extracted from scientific publications. These texts are primarily single sentences, e.g., *Interventions in all studies included implementation strategies targeting healthcare workers; three studies included delivery arrangements, no studies used financial arrangements or governance arrangements.* (pair_id = CD012520, 206 characters) with a considerably complex sentence structure. There are 9086 unique texts in the dataset, they have an average length of 168.66 characters. Figure 1 indicates the lengths of unique texts contained in the dataset as a histogram.

3. Method

We investigate three steps which can all be optionally employed in text simplification. At first a *rephrasing*, e.g., via an LLM, can modify the original text. Then for an original or modified text the *complex scientific terms* can be identified and marked specifically. As a last step, the actual *simplification* is run on original or modified texts, which either have complex terms marked or not, e.g., by again using an LLM.

3.1. Step 1: Rephrasing

For runs incorporating this step, we employ a prompt that reformulates the original text. This can be to make an original text *more complex* (see prompt C in Table 1) or to simply *rephrase* it while striving to maintain complexity and length (see prompt P in Table 1). In general we experiment in three directions: modifying the original text, introducing more complexity and simplifying it immediately. The attempt to increase a text's complexity is assumed to produce worse evaluation results if the texts are not at their maximum complexity level yet. If the texts' complexity is already the highest it can be, an experiment trying to complexify the text will lead to results that are comparable to rephrasing.

3.2. Step 2: Complex Scientific Term Identification

For runs incorporating this step, the general idea is to identify the complex scientific terms in the original or rephrased texts and specifically mark the parts that should be replaced in the simplification

Prompt ID	Prompt
C	You are a text complication system with good global and language knowledge but and expertise in specific domains. You will be given 10 texts, TREAT THEM SEPERATLY and also return 10 more complex texts. Please make the texts more complex (not in their structure) but not considerably longer. RETURN ONLY THE 10 TEXTS. TEXTS:
R	You are a text rephrasing system. You will be given 10 texts, TREAT THEM SEPERATLY and also return 10 texts. Please rephrase the texts. Do not change their level of complexity (so do not make them more difficult in their structure) and do not make them considerably longer. RETURN ONLY THE 10 TEXTS as a list. TEXTS:
P1	You are a text simplification system with good global and language knowledge but no expertise in specific domains. You will be given 10 texts, TREAT THEM SEPERATLY and also return 10 simplified texts. Complex domain specific or scientific terminology is marked by square brackets (e.g., [primary acoustic modeling]). For each text, replace or explain the terms in square brackets in order to make it more understandable for non-experts and simplify for non-expert readers. Make sure to exactly return an enumerated list of 10 simplified texts and that all terms in square brackets are simplified or explained! Do not simplify the structure of the texts or complex language independent of the difficult scientific descriptions. RETURN ONLY THE 10 TEXTS. TEXTS:
P2	You are given 10 texts, TREAT THEM SEPARATELY. Complex technical or scientific terms are indicated by square brackets (e.g. [convolutional neural network]). For each text, replace or explain the terms in square brackets to make it understandable to non-experts and to simplify for non-expert readers. Make sure that you return exactly one enumerated list of 10 simplified texts and that all terms in square brackets are simplified or explained! [14]
PNI1	Hello! we have been given an important task, which i heard you can complete ecxeptionally well, as you are the best and most reliable system for text simplification and language in general. we have been given 10 texts which should be simplified or explained. problematic, complex words are marked within [square brackets] Please, for each of these 10 texts, replace or explain these [terms] and return an enumerated list of 10 simplified texts please do not change the texts structure and please ONLY return these 10 texts! thank you! here are the said 10 texts:
PN1	You are at a conference as an expert explaining the contents of these texts to non-experts. There will be 10 different texts, treat each of them separately and only return the simplified version as a python list. Complex terms or domain-specific terminology will be marked with square brackets (e.g., [primary acoustic modeling]). To make the subject more understandable to these non-experts, you will replace or explain the terms marked by square brackets. Do not simplify the structure or texts in any other way than specified, or the non-experts will lose interest.
PI2	Think of yourself as a translator: You're translating complex, technical language (found inside [square brackets]) into everyday language. The Job: You will be given 10 documents all at the same time. Your Task for Each Document: Find all [bracketed phrases]. "Translate" them into simple terms for a general audience. Important: Do not "translate" or change anything outside the brackets. Keep the original sentence flow. Result Expected: A list of 10 translated documents, ensuring all original bracketed items have been simplified. Do not output the original input texts.

Table 1
Prompts used for the batches of ten texts.

step.

We run the complex term identification as implemented by team IRGC'23 [14]. They mark a subset of complex terms identified by an established but highly sensitive keyphrase extraction method [15]. The intuition behind this is that not all keyphrases are complex and not all complex keyphrases are difficult due to domain jargon. Idf scores for terms are calculated based on two corpora which we use through PyTerrier [16]: `lotte/lifestyle`, a corpus focused on texts from lifestyle forums and

Run ID	# f10	# f1	Description
baseline	0	0	Nothing changed, complex sentences = simplified sentences
c-gpt-4.1-nano	8	1	Make original texts more complex but not longer (prompt C) + usage of OpenAI
c-gemini-2.5-flash-preview	18	7	Make original texts more complex but not longer (prompt C) + usage of Gemini 2.5
c-gemini-2.0-flash	10	1	Make original texts more complex but not longer (prompt C) + usage of Gemini 2.0
r-gemini-2.5-flash-preview	1	0	Rephrase original texts with same complexity and length (prompt R) + usage of Gemini 2.5
r-gemini-2.0-flash	47	1	Rephrase original texts with same complexity and length (prompt R) + usage of Gemini 2.0
p1-gpt-4.1-nano	67	22	Complex term identification + new prompt (P1) + usage of OpenAI
p1-gemini-2.0-flash	42	22	Complex term identification + new prompt (P1) + usage of Gemini 2.0
p1-ac-gemini-2.0-flash	44	22	c-gpt-4.1-nano complexification + complex term identification + new prompt (P1) + usage of Gemini 2.0
p2-gpt-4.1-nano	75	21	Complex term identification + old prompt (P2) + usage of OpenAI
p2-gemini-2.5-flash-preview	275	1550	Complex term identification + old prompt (P2) + usage of Gemini 2.5
p2-gemini-2.0-flash	87	132	Complex term identification + old prompt (P2) + usage of Gemini 2.0
p2-ac-gemini-2.0-flash	91	132	c-gpt-4.1-nano complexification + complex term identification + old prompt (P2) + usage of Gemini 2.0
pni1-gpt-4.1-nano	32	<320	Complex term identification + new prompt (PNI1) + usage of OpenAI
pn1-gemini-2.0-flash	174	1268	Complex term identification + new prompt (PN1) + usage of Gemini 2.0
pi2-gemini-2.0-flash	199	174	Complex term identification + new prompt (PI2) + usage of Gemini 2.0

Table 2
Information on run IDs, the number of failed modifications for the batch processing (# f10), single processing (# f1) and the description of the run. Our official Run IDs have a THM_task11_-prefix. Due to an error # f1 of pni1-gpt-4.1-nano has not been tracked.

lotte/science, a corpus focused on texts from science-oriented forums [17]. Terms with a higher idf in science texts than in lifestyle texts are considered to be domain jargon which should be explained to non-experts. Then, only extracted keyphrases where their idf surpasses a certain complexity threshold are marked as complex terms which should be explained to non-experts. In our experiments we work with a complexity threshold of 0.1[1].

From the 9160 original texts in the dataset, 3473 texts did not receive any markings indicating complex phrases.

3.3. Step 3: Simplification

All our prompts ask LLMs to take on a specific persona to simplify the texts as this prompting technique is highly effective in numerous tasks [18]. For simplifying the texts, we run our prompts (see prompts P1, P2, PNI1, PN1 and PI2 in Table 1). While P2 is the prompt introduced by IRGC'23 [14] defining the LLM's expected persona not specifically, P1 clearly states to behave as a text simplification system

[1]IRGC'23 used a complexity threshold of 0.01 which is more sensitive whereas our usage of 0.1 is more strict in identifying complex terms in the science domain, supposedly reducing the number of false positive complex terms.

Processing	Prompt text
Batch	You are given 10 texts, TREAT THEM SEPARATELY. Complex technical or scientific terms are indicated by square brackets (e.g. [convolutional neural network]). For each text, replace or explain the terms in square brackets to make it understandable to non-experts and to simplify for non-expert readers. Make sure that you return exactly one enumerated list of 10 simplified texts and that all terms in square brackets are simplified or explained!
Single	You are given 1 text. Complex technical or scientific terms are indicated by square brackets (e.g. [convolutional neural network]). For each text, replace or explain the terms in square brackets to make it understandable to non-experts and to simplify for non-expert readers. Make sure that you return exactly one enumerated list of 1 simplified text and that all terms in square brackets are simplified or explained!

Table 3
Comparison of the batch processing prompt of P2 with its single text processing variant.

where no domain-knowledge should be assumed in the simplification. PNI1 treats the LLM more informally and similar to a human but does imply the expected persona by stating it being the *best and most reliable system for text simplification and language in general.* PN1 defines the LLM's persona as someone attending a conference. PI2 instructs the LLM to act as a translator translating complex technical language to everyday language.

In general we run our prompts for randomly batched texts of size 10, if prompts do not return 10 modified texts, we run these texts again one on one in the end with slightly modified versions of the prompts to work on single texts instead of 10. If these single runs also do not produce single texts as output, we consider the original (complex) text as the simplified text as well.

4. Runs and Results

4.1. Submitted Runs

Our submitted runs have the prefix of our team name THM combined with the task number, prompt ID as well as used LLM, resulting in run names such as `THM_task11_p1-gpt-4.1-nano` with `p1` being a prompt ID and `gpt-4.1-nano` describing the used model in a run. For readability we refrain from writing out the first part of the run IDs (`THM_task11_`) in this paper. For our simplification prompts we defined the batch processing versions in Table 1, the single text versions are almost identical, Table 3 holds an example for P2.

Table 2 gives an overview of our submitted runs as well as the number of cases in which we did not receive the expected number of simplifications when using an LLM. We experiment with usage of small Gemini models (`gemini-2.0-flash` and `gemini-2.5-preview-flash`) and a small OpenAI model (`gpt-4.1-nano`). Our implementation can be found on GitHub[2].

4.2. Results

We found considerable differences between employing `gpt-4.1-nano` and the two versions of Gemini with the OpenAI model producing better results in all cases. Usage of the more powerful model `gemini-2.5-flash-preview` did not consistently produce better scores compared to the older model `gemini-2.0-flash` (see prompt P2). All considered different configuration of runs using prompt P2 produced better results than runs with the same configurations using prompt P1. P1 is the more descriptive version of the more concise P2. In P1 the context of being a text simplification system with definition of the knowledge is explicitly defined. Additionally, P1 puts extra emphasis on the requirement of returning only 10 texts while keeping the structure of the original text. We assume the longer prompt with more instructions dilutes the LLM's attention from the actual task of simplifying texts. Table 5 gives some randomly selected examples of original texts as well as their simplifications

[2] https://github.com/kreutzch/THM-SimpleText-25

Run ID	SARI simple original	SARI simple auto
baseline	7,844	12.033
c-gpt-4.1-nano	32.445	33.944
c-gemini-2.5-flash-preview	30.175	32.586
c-gemini-2.0-flash	28.917	31.199
r-gemini-2.5-flash-preview	29.256	32.598
r-gemini-2.0-flash	27.526	31.160
p1-gpt-4.1-nano	38.238	40.416
p1-gemini-2.0-flash	25.858	30.001
p1-ac-gemini-2.0-flash	25.234	29.080
p2-gpt-4.1-nano	**39.572**	**41.315**
p2-gemini-2.5-flash-preview	25.062	27.811
p2-gemini-2.0-flash	29.263	31.174
p2-ac-gemini-2.0-flash	27.561	30.024
pni1-gpt-4.1-nano	35.262	37.262
pn1-gemini-2.0-flash	32.267	34.474
pi2-gemini-2.0-flash	20.521	24.045

Table 4
SARI [19] scores of submitted runs produced by Codabench.

Para ID	Sent ID	Original	Simplified with P1	Simplified with P2
0	5	Quality of care outcomes (proportions of patients receiving evidence-based care) were included in all included studies.	All the studies included looked at the quality of care, meaning how many patients received care based on the best available evidence.	The quality of care outcomes (how many patients received care based on the best available evidence) were included in all the studies.
5	10	For every nine people treated with haloperidol instead of olanzapine, one fewer person would experience clinically important improvement in quality of life.	For every nine people treated with haloperidol (a medication used for mental health conditions) instead of olanzapine (another mental health drug), one fewer person would experience a meaningful improvement in their quality of life.	For every nine people treated with haloperidol (a medication used for mental health conditions) instead of olanzapine (another mental health drug), one fewer person would experience a meaningful improvement in quality of life.
6	18	Telephone interventions provide a convenient way of supporting self-management of cancer-related symptoms for adults with cancer.	Telephone-based programs offer a convenient way to support adults with cancer in managing symptoms related to their illness on their own.	Telephone-based programs offer a convenient way to support people in managing symptoms related to cancer on their own.

Table 5
Examples of original texts as well as texts produced with prompts P1 and P2 with gpt-4.1-nano. Parts marked in green signal the modifications compared to the original text.

using P1 and P2 combined with gpt-4.1-nano. It appears that both prompts *produced almost the same simplification* in some cases (as seen in the second example), P2 *staying closer to the original text* (as seen in the first example) and P2 producing slightly *shorter or more concise* texts (as seen in the third example). Both prompts do not seem to fully convey the instruction of only modifying the identified complex terms in original texts.

Table 4 indicates that the runs instructing the LLMs to complexify the original sentences did not perform worse than a rephrasing of original sentences without reducing complexity or length. Table Table 6 gives again gives the same three examples as seen in Table 5 of original texts with their reprasing as well as their complexification. It appears as if complexified sentences remained closer to the original versions than the rephrased ones which could lead at increased SARI scores as they are also closer to the gold simplification. This hints at problems with the employed evaluation measures.

Table 7 gives an example for a simplification produced by prompts PNI1, PN1 and PI2. Here we see

Para ID	Sent ID	Original	Rephrased with R	Complexified with C
0	5	Quality of care outcomes (proportions of patients receiving evidence-based care) were included in all included studies.	All studies incorporated quality of care outcomes, which measured the proportion of patients receiving evidence-based care.	Parameters indicative of care quality (specifically, the prevalence of patients undergoing empirically validated therapeutic interventions) were incorporated into all pertinent scholarly endeavors.
5	10	For every nine people treated with haloperidol instead of olanzapine, one fewer person would experience clinically important improvement in quality of life.	If nine individuals were treated with haloperidol instead of olanzapine, one fewer person would achieve a clinically significant improvement in their quality of life.	For every nine individuals receiving haloperidol in lieu of olanzapine, a reduction of one individual would be observed in those experiencing a clinically significant amelioration in their overall quality of life.
6	18	Telephone interventions provide a convenient way of supporting self-management of cancer-related symptoms for adults with cancer.	Telephone interventions offer a practical method for assisting adults with cancer in self-managing their cancer-related symptoms.	6. Telephonic interventions offer an accessible modality for facilitating autonomous symptom navigation of oncological sequelae for adult oncological patients.

Table 6
Examples of original texts as well as texts produced with prompts R and C with `gemini-2.5-flash-preview`.

Prompt	Simplification
PNI1	For every nine people treated with haloperidol (a medication for mental health conditions) instead of olanzapine (another mental health medication), one fewer person would see a meaningful improvement in their quality of life.
PN1	Okay, everyone, let's look at this statement. It's comparing two different medications, haloperidol and olanzapine. Think of haloperidol as an older type of medication used to treat certain mental health conditions. The statement is saying that if you treat nine people with this older medication, haloperidol, instead of olanzapine, which is a newer medication, you'll find that *one* fewer person will see a meaningful improvement in their overall well-being or "quality of life." So, olanzapine seems to be giving people a better chance at improving their quality of life compared to haloperidol, according to this statement.
PI2	For every nine people treated with a strong, older antipsychotic medication instead of olanzapine, one fewer person would experience clinically important improvement in quality of life.

Table 7
Comparison of simplifications for prompts PNI1, PN1 and PI2 for **Para ID** = 5 and **Sent ID** = 10. Parts marked in green signal the modifications compared to the original text.

the closeness of the text produced by PNI1 to the original text while explaining the complex terms, the complete rephrasing of the text produced by PN1 that also simplified the original text significantly and the text produced by PI2 that did keep a good chunk of the original text and did not simplify it much. Failure to simplify the complex parts of the original texts does unsurprisingly lead to lower SARI scores.

From our experiments, Table 2 and Table 4 we conclude the description of the return type specified to a LLM as well as the post-processing of the result being highly influential of the overall quality of the run. Runs utilising the Gemini variants produced higher error rates than the GPT runs with the exception of P1. In our experiments we felt that it was more difficult to write prompts suitable for Gemini for this exact reason.

4.3. Costs

Table 8 provides an overview of cost, input tokens and requests per model.

At the time of producing the runs, the used OpenAI model `gpt-4.1-nano` cost $0.10 per 1M tokens. We also experimented with usage of `gpt-3.5-turbo-0125` with a cost of $0.50 per 1M tokens. During experimentation as well as running our actual prompts we issued 33,721 requests with 12.499M input

Model	$ Input	$ Output	# Input Tokens	# Requests
gpt-4.1-nano	$1.072	$3.295	10.833M	30,975
gpt-3.5-turbo-0125	$0.833	$0.602	1.666M	2,746
gemini-2.5-flash-preview	$0.45	$1.44	3.655M	9.09k
gemini-2.0-flash	$0.75	$2.85	9.93M	21.39k

Table 8
Overview of cost in $ for used LLMs during all experiments and runs with input number of tokens and requests.

tokens. In total we spent $6.39 using OpenAI models.

Both our used Gemini models `gemini-2.5-flash-preview` and `gemini-2.0-flash` cost $0.10 per 1M input tokens. For both models, the output cost was $0.40 per 1M tokens. In total we issued 30.48k requests with 13.585M input tokens. The total costs for all models used was $5.49.

The full cost of using LLMs for our participation in the shared task was thus $11.88. We issued 64.2k requests with 26.084M input tokens.

5. Conclusion

Our experiments show the potential value of combining a cheap complex term identification step with low-cost LLM options for simplifying complex texts for non-expert readers. It appears prompts producing texts that remain more similar to the original texts perform better in the evaluation compared to prompts producing texts with more changed up wording. The actual difficulty of produced texts does not seem to be the sole determining factor of the goodness of the strategy.

Future work should focus on experimenting with locally-run LLMs and other options to identify complex keyphrases in texts. Another line of work should research better automatic evaluation measures for the text simplification task.

Acknowledgments

We thank the SimpleText organisers for continuously organising the shared task and bringing together the community to advance the field of automatic text simplification.

Declaration on Generative AI

The authors have not employed any Generative AI tools for writing the manuscript. The authors used OpenAI and Gemini models in their implementation.

References

[1] L. Ermakova, P. Bellot, P. Braslavski, J. Kamps, J. Mothe, D. Nurbakova, I. Ovchinnikova, E. SanJuan, Text simplification for scientific information access - CLEF 2021 simpletext workshop, in: Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II, volume 12657, 2021, pp. 583–592. doi:10.1007/978-3-030-72240-1_68.

[2] L. Ermakova, P. Bellot, J. Kamps, D. Nurbakova, I. Ovchinnikova, E. SanJuan, É. Mathurin, S. Araújo, R. Hannachi, S. Huet, N. Poinsu, Automatic simplification of scientific texts: Simpletext lab at CLEF-2022, in: Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II, volume 13186, 2022, pp. 364–373. doi:10.1007/978-3-030-99739-7_46.

[3] L. Ermakova, E. SanJuan, S. Huet, O. Augereau, H. Azarbonyad, J. Kamps, CLEF 2023 simpletext track - what happens if general users search scientific texts?, in: Advances in Information Retrieval

- 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III, volume 13982, 2023, pp. 536–545. doi:10.1007/978-3-031-28241-6_62.
[4] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, G. M. D. Nunzio, F. Vezzani, J. D'Souza, S. Kabongo, H. B. Giglou, Y. Zhang, S. Auer, J. Kamps, CLEF 2024 simpletext track - improving access to scientific texts for everyone, in: Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part VI, volume 14613, 2024, pp. 28–35. doi:10.1007/978-3-031-56072-9_4.
[5] L. Ermakova, H. Azarbonyad, J. Bakker, B. Vendeville, J. Kamps, CLEF 2025 simpletext track - simplify scientific text (and nothing more), in: Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part V, volume 15576, 2025, pp. 425–433. doi:10.1007/978-3-031-88720-8_63.
[6] L. Ermakova, H. Azarbonyad, J. Bakker, B. Vendeville, J. Kamps, Overview of the CLEF 2025 SimpleText track: Simplify scientific texts (and nothing more), in: J. Carrillo de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, 2025.
[7] C. Kreutz, F. Haak, B. Engelmann, P. Schaer, BATS: benchmarking text simplicity, in: Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024, 2024, pp. 11968–11989. doi:10.18653/V1/2024.FINDINGS-ACL.712.
[8] S. Štajner, I. Hulpuş, Automatic assessment of conceptual text complexity using knowledge graphs, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 318–330. URL: https://aclanthology.org/C18-1027/.
[9] S. Štajner, S. Nisioi, D. Ibanez, Is simple english wikipedia as simple and easy-to-understand as we expect it to be?, in: Proceedings of the 9th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion, DSAI '20, 2021, p. 66–70. doi:10.1145/3439231.3439263.
[10] F. Padovani, C. Marchesi, E. Pasqua, M. Galletti, D. Nardi, Automatic text simplification: A comparative study in Italian for children with language disorders, in: Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning, 2024, pp. 176–186. URL: https://aclanthology.org/2024.nlp4call-1.13/.
[11] S. Štajner, Automatic text simplification for social good: Progress and challenges, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 2637–2652. doi:10.18653/v1/2021.findings-acl.233.
[12] S. Gooding, On the ethical considerations of text simplification, in: S. Ebling, E. Prud'hommeaux, P. Vaidyanathan (Eds.), Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022), 2022, pp. 50–57. doi:10.18653/v1/2022.slpat-1.7.
[13] J. Bakker, B. Vendeville, L. Ermakova, J. Kamps, Overview of the CLEF 2025 SimpleText Task 1: Simplify Scientific Text, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.
[14] B. Engelmann, F. Haak, C. K. Kreutz, N. Nikzad-Khasmakhi, P. Schaer, Text simplification of scientific texts for non-expert readers, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497, 2023, pp. 2987–2998. URL: https://ceur-ws.org/Vol-3497/paper-250.pdf.
[15] M. Kulkarni, D. Mahata, R. Arora, R. Bhowmik, Learning rich representation of keyphrases from text, in: Findings of the Association for Computational Linguistics: NAACL 2022, 2022, pp. 891–906. doi:10.18653/v1/2022.findings-naacl.67.
[16] C. Macdonald, N. Tonellotto, Declarative experimentation in information retrieval using pyterrier, in: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval, ICTIR '20, 2020, p. 161–168. doi:10.1145/3409256.3409829.
[17] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, M. Zaharia, ColBERTv2: Effective and efficient

retrieval via lightweight late interaction, 2022. doi:10.18653/v1/2022.naacl-main.272.
[18] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, in: Proceedings of the 30th Conference on Pattern Languages of Programs, 2023. doi:10.5555/3721041.3721046.
[19] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, Transactions of the Association for Computational Linguistics 4 (2016) 401–415. doi:10.1162/tacl_a_00107.