Overview of the CLEF 2024 SimpleText Task 3: Simplify Scientific Text

Liana Ermakova[1], Valentin Laimé[1], Helen McCombie[2] and Jaap Kamps[3]

[1]Université de Bretagne Occidentale, HCTI, France
[2]Université de Bretagne Occidentale, BTU, France
[3]University of Amsterdam, Amsterdam, The Netherlands

Abstract

This article provides a comprehensive summary of the CLEF 2024 SimpleText Task 3, which focuses on simplifying scientific text based on specific queries. We discuss in detail the motivation for lay access to scholarly literature, and provide an overview of the setup of the scientific text simplification task. One of the main innovations of the CLEF 2024 SimpleText Task 3 is to complement sentence-level text simplification with a document-level text simplification task. We describe the resulting sentence-level and document-level text simplification test collection in detail, which consists of a corpus of over 1,500 paired source and reference sentences, and a corpus of over 250 paired source and reference abstracts, both containing the source text from scientific abstracts with direct reference simplifications produced by human annotators. We present the results of the participants submission, with 15 teams submitting 52 sentence-level text simplification runs and 9 teams submitting 31 sentence-level text simplification runs. The article concludes with an in-depth analysis, including information distortion and potential LLM "hallucinations" of the simplified sentences submitted by participants.

Keywords

automatic text simplification, science popularization, information distortion, error analysis, lexical complexity, syntactic complexity, LLMs hallucination

1. Introduction

Becoming science literate is more important than ever before. Objective scientific information helps any user to navigate a world of where misinformation, disinformation, or unfounded generated information is only a single mouse click away. Everyone acknowledges the importance of objective scientific information. However, finding and understanding relevant scientific documents is often challenging due to complex terminology and readers' lack of prior knowledge. The question is can we improve accessibility for everyone?

Text simplification technology holds the promise to remove some of the access barriers [1, 2, 3, 4]. Despite impressive progress, the automatic removal of comprehension barriers between scientific texts and the general public remains an ongoing challenge. The paper highlights that even the most advanced language models currently available face difficulties when it comes to simplifying scientific texts. The described results demonstrate the limitations of these models in effectively tackling the task of simplification in the scientific domain.

The CLEF 2024 SimpleText track brings together researchers and practitioners working on the generation of simplified summaries of scientific texts. It is an evaluation lab that follows up on the CLEF 2021 SimpleText Workshop [5] the CLEF 2022 SimpleText Track [6], and the CLEF 2023 SimpleText Track [7].

The CLEF 2024 SimpleText track is based on four interrelated tasks:

1. Task 1 on *Content Selection*: retrieve passages to include in a simplified summary.
2. Task 2 on *Complexity Spotting*: identify and explain difficult concepts.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 9–12, 2024, Grenoble, France
liana.ermakova@univ-brest.fr (L. Ermakova)
https://simpletext-project.com/ (L. Ermakova)
0000-0002-7598-7474 (L. Ermakova); 0000-0002-6614-0087 (J. Kamps)

Table 1
CLEF 2024 Simpletext Task 3 official run submission statistics

Task	AIIR Lab	AMATU	Arampatzis	Elsevier	L3S	LIA	PiTheory	Sharigans	SINAI	SONAR	AB/DPV	Dajana/Katya	Frane/Andrea	Petra/Regina	Ruby	Tomislav/Rowan	UAmsterdam	UBO	UniPD	UZH Pandas	Total
3.1	4		4	8			11	1		1	1	1	1	1	1	1	4	2		11	52
3.2	4		4	2			10	1							1	1	6	2			31

3. Task 3 on *Text Simplification*: simplify scientific text.

4. Task 4 on *SOTA?*: track the state-of-the-art in scholarly publications.

This paper presents an overview of the CLEF 2024 SimpleText Task 3 on *Content Selection*. For a comprehensive overview of the other tasks, the task overview papers on Task 1 [8], Task 2 [9], and Task 4 [10], as well as the track overview paper [11], provide detailed information and further insights.

The CLEF 2024 SimpleText Task 3 directly addresses the technical and evaluation challenges associated with making scientific information accessible to a wide audience, including students and non-experts. We describes the data and benchmarks provided for scientific text simplification, along with the participants' results and further analysis. This task on simplifying scientific text is a direct continuation of the CLEF 2023 Task 3 [12]. One of the key innovation in 2024 is the introduction of both sentence level and document (abstract) level scientific text simplification subtasks, as Task 3.1 and Task 3.2.

A total of 45 teams registered for our SimpleText track at CLEF 2024. A total of 20 teams submitted 207 runs in total for the Track, of which 15 teams submitted a total of 83 runs for Task 3. The statistics for the Task 3 runs submitted are presented in Table 1. However, some runs had problems that we could not resolve. We do not detail them in the paper as well as the 0-scored runs.

This introduction is followed by Section 2 presenting the text simplification task with the datasets and evaluation metrics used. Section 3 gives an overview of text simplification approaches for scientific text as deployed by the participants. In Section 4, we present and discuss the results of the official submissions. In Section 5, a thorough analysis of the results is carried out, covering several important aspects. This includes examining the relationship between difficult scientific terms and the simplification process, investigating information distortion that may occur during simplification, and exploring instances of language models (LLMs) generating hallucinations and producing inaccurate information. The analysis delves into these topics to provide a comprehensive understanding of the findings and insights derived from the study. We end with Section 6 summarizes the findings and draws perspective for future work.

2. Task 3: Simplify Scientific Text

This section details *Task 3: Text Simplification* on simplify scientific text.

2.1. Description

The goal of this task is to provide a simplified version of the sentences extracted from scientific abstracts. Participants will be provided with popular science articles and queries and matching abstracts of scientific papers, either split into individual sentences or as the entire abstracts. This year will feature both sentence level (Task 3.1) and document or abstract level (Task 3.2) text simplification.

Table 2 shows an example of a human reference simplification, combining the input sentences belonging to the abstract of the document $id = 130055196$ retrieved for query G01.1. Here, we show the deletions and insertions relative to the source input sentences (in this case in the first 4 sentences).

Table 2
Example of SimpleText Task 3 human reference simplifications of the source input: ~~deletions~~ and insertions

Topic	Document	Output
G01.1	130055196	~~As various kinds~~ The rise of output devices ~~emerged , such as highresolution~~ like high-resolution printers ~~or a display of~~ and PDA ~~(Personal Digital Assistant) ,~~ displays has increased the ~~importance of~~ need for high-quality resolution conversion ~~has been increasing~~ . ~~This~~ The paper proposes a new method ~~for enlarging image with~~ to make images bigger while maintaining high quality . ~~One of the largest problems on image enlargement~~ The main issue with enlarging images is ~~the exaggeration of the jaggy~~ that jagged edges can become exaggerated . To ~~remedy~~ solve this problem , we ~~propose~~ suggest a new ~~interpolation~~ method ~~, which uses artificial~~ that helps us to estimate the value of the newly generated pixels using a neural network ~~to determine the optimal values of interpolated pixels~~ . The ~~experimental~~ experiment 's results are ~~shown~~ presented and ~~evaluated~~ analyzed . ~~The~~ We evaluate the effectiveness of our methods ~~is discussed~~ by comparing ~~with the conventional methods~~ them to traditional approaches .

Table 3
CLEF SimpleText Task 3 Scientific Text simplification Corpora

Task	Level	Role	Source	Reference
3.1	Sentence	Train	893 sentences	958 simplified sentences
3.1	Sentence	Test	578 sentences	578 simplified sentences
3.1	Sentence	Combined	1,471 sentences	1,536 simplified sentences
3.2	Document	Train	175 abstracts	175 simplified abstracts
3.2	Document	Test	103 abstracts	103 simplified abstracts
3.2	Document	Combined	278 abstracts	278 simplified abstracts

2.1.1. Data

Task 3 uses a corpus based on the high-ranked abstracts retrieved for the requests of the CLEF 2024 SimpleText Task 1. Our training data is a truly parallel corpus of directly simplified sentences coming from scientific abstracts from the DBLP Citation Network Dataset for *Computer Science* and Google Scholar and PubMed articles on *Health and Medicine*. Other existing text simplification corpora used post-hoc aligned sentences [e.g., 13].

In 2024, we expanded the training and evaluation data. In addition to sentence-level text simplification, we will provide document-level or abstract-level input and reference simplifications. In order to make the sentence-level and document-level tasks fairly comparably, both use the exact same reference simplifications. The scientific sentences from scientific abstracts were simplified either by master students in Technical Writing and Translation or by a domain expert (a computer scientist) and a professional translator (native English speaker) working together.

Table 3 gives an overview of all the SimpleText Task 3 scientific text simplification corpora constructed in 2024. The SimpleText corpus contains 1,536 directly simplified sentences, corresponding to 278 scientific abstracts. This is a useful addition to existing high-quality corpora like Newsela [13], with 2,259 sentences in Newsela-Manual. Our track is the first to focus on the simplification of scientific text with a much higher text complexity than news articles.

Available Task 3 training data is derived from the CLEF 2023 edition [7], and includes 893 source sentences from 175 scientific abstracts paired with the corresponding manual reference simplifications. The new test data created in 2024 consists of 578 sentences paired with reference simplifications for the sentence-level task (Task 3.1), and 103 abstracts paired with reference simplifications for the document-level task (Task 3.2).

2.1.2. Formats

Sources The source data are provided in JSON formats with the following fields:

1. *snt_id* (Task 3.1) or *abs_id* (Task 3.2): a unique sentence (or abstract) identifier
2. *source_snt* (Task 3.1) or *source_abs* (Task 3.2): passage text (sentence or abstract)
3. *doc_id*: a unique source document identifier
4. *query_id*: a query ID
5. *query_text*: difficult terms should be extracted from sentences with regard to this query

An example of the Task 3.1 JSON source input is:

```
{
    "query_id":"G11.1",
    "query_text":"drones",
    "doc_id":2892036907,
    "snt_id":"G11.1_2892036907_2",
    "source_snt":"With the ever increasing number of unmanned aerial vehicles getting
    ↪  involved in activities in the civilian and commercial domain, there is an increased
    ↪  need for autonomy in these systems too."
},
```

Predictions Predictions or submissions of participants were also requested in a JSON format with the following fields:

1. *run_id:* Run ID starting with *<team_id>_<task_id>_<method_used>*, e.g. *UBO_Task**3.1**_BLOOM*
2. *manual:* Whether the run is manual {0,1}
3. *snt_id* (Task 3.1) or *abs_id* (Task 3.2): a unique sentence or abstract identifier from the input file
4. *simplified_snt* (Task 3.1) or *simplified_abs* (Task 3.2): simplified text for the sentence or abstract

An example of the Task 3.1 submission in JSON is:

```
{
    "run_id": "Elsevier@SimpleText_Task3.1_run1",
    "manual": 0,
    "snt_id": "G11.1_2892036907_2",
    "simplified_snt": "As more and more drones are used for civilian and commercial
    ↪  purposes, there is a growing need for them to operate independently."
},
```

References The references are provided in a very similar format as the predictions above. An example of a Task 3.1 reference in JSON is:

```
{
    "snt_id":"G11.1_2892036907_2",
    "simplified_snt":"Drones are increasingly used in the civilian and commercial domain
    ↪  and need to be autonomous."
},
```

2.1.3. Evaluation

In 2024, we emphasize large-scale automatic evaluation measures (SARI, BLEU, compression, readability) that provide a reusable test collection. This automatic evaluation will be supplemented with a detailed human evaluation of other aspects, essential for deeper analysis. Almost all participants used generative models for text simplification, yet existing evaluation measures are blind to potential hallucinations with extra or distorted content [12]. In 2024, we provide further analysis of ways to detect and quantify spurious content in the output, potentially corresponding to what is informally called “hallucinations.”

3. Scientific Text Simplification Approaches

In this section, we discuss a range of text simplification approaches that have been applied to scientific text as provided by the track. A total of 15 teams submitted 83 runs in total.

AB/DPV Varadi and Bartulović [14] submitted one run for Task 3. Their approach is an LSTM model for the sentence-level task.

AIIRLab Largey et al. [15] submitted a total of eight runs for Task 3. Their approach uses LLaMA3 and Mistral models with different prompting and fine-tuning, for both the sentence-level and abstract-level tasks.

Arampatzis (No paper received) submitted a total of eight runs for Task 3. Their approach is a range of models (DistilBERT, T5) for both the sentence-level and abstract-level tasks.

Dajana/Katya (No paper with run details received) submitted one run for Task 3. Their approach which follows standard text simplification approaches is applied to the sentence-level task.

Elsevier Capari et al. [16] submitted a total of ten runs for Task 3. Their approach is based on a GPT-3.5 model experimenting with zero-shot and few-shot prompts for both sentence-level and abstract-level tasks.

Frane/Andrea (No paper with run details received) submitted one run for Task 3. Their approach which follows standard text simplification approaches is applied to the sentence-level task.

Petra/Diana Elagina and Vučić [17] submitted one run for Task 3. Their approach is a LLaMA model for the sentence-level task.

PiTheory (No paper with run details received) submitted a total of twenty runs for Task 3. Their approach uses pre-trained BART and T5 models but contains very few results for both the sentence-level and abstract-level tasks.

Ruby (No paper received) submitted two runs for Task 3. Their approach uses standard models for both sentence-level and abstract-level tasks.

Sharigans Ali et al. [18] submitted a total of two runs for Task 3. Their approach is a GPT-3.5 model for both the sentence-level and abstract-level tasks.

SONAR (No paper received) submitted a single run for Task 3. Their approach is a standard model for the sentence-level task.

Tomislav/Rowan Mann and Mikulandric [19] submitted a total of two runs for Task 3. Their approach is the LLama 2 model with a range of prompts and post-processing for both the sentence-level and abstract-level tasks. Their submission only covers a part of the train topics.

UAmsterdam Bakker et al. [20] submitted a total of ten runs for Task 3. They experiment with GPT-2, and Wiki and Cochrane-trained models at the sentence, paragraph, and document-level text simplification, for both sentence-level and document-level tasks.

UBO Vendeville et al. [21] submitted a total of four runs for Task 3. Their approach is to prompt a smaller Phi3 model for lexical and grammatical text simplifications, for both the sentence-level and abstract-level tasks.

UZHPandas Michail et al. [22] submitted a total of ten runs for Task 3. They experiment with a multi-prompt Minimum Bayes Risk (MBR) decoding approach to the sentence-level task. Their approach is a refinement of their CLEF 2023 approach, which was recognized with a prestigious *Best of the Labs* award, and published as part of the CLEF 2024 LNCS proceedings [23].

4. Results

This section details the results of the task, for both sentence-level and abstract-level test simplification subtasks.

4.1. Task 3.1: Sentence-level scientific text simplification

Table 4 shows the Task 3.1 (sentence-level text simplification) results. The table is restricted to submissions covering a sufficient number of input sentences. We show a number of evaluation scores against the human reference simplifications, in particular SARI and BLEU. In addition, we provide additional text statistics on the system output such as FKGL, and a comparison to the source input.

We make a number of observations. First, the table is sorted on SARI, the main automatic text simplification measure used in the track. We observe SARI scores of 30+ % for the majority of systems and 40+ % for the top-scoring systems. This high overlap with the human reference simplifications is encouraging and indicates that the effectiveness of text simplification approaches, traditionally trained on youth news reading corpora like Newsela, also extends to scientific text.

Second, in terms of the level of text complexity, readability measures like FKGL provide a rough indicator of lexical and grammatical complexity. The original sentences have an FKGL of 13-14 corresponding to university-level text, and the majority of systems reduce this to an FKGL of 11-12 corresponding to the exit level of compulsory education. This is an encouraging result, as it indicates that the scientific text simplification approach can be a viable approach to lower the textual complexity of scientific text toward the range acceptable by a layperson. Although this is positive indicator, this approximate measure does not take into account terminological complexities as studied in Task 2, or ways to retrieve all and only more accessible abstracts in Task 1 [24].

Third, the table includes various other scores that indicate that there is still considerable room for improvement in scientific text simplification. Throughout the table the BLEU evaluation measure remains very low, and leads to a different ranking of systems with some of the best systems on BLEU demonstrating superior overlap with the human reference simplifications. The table also reveals some runs with very high "compression" ratios and sentence splits, as well as high proportions of additions. While evaluation measures like SARI are essential for understanding important aspects of text simplification output quality, they are also known to be relative insensitive to content outside the intersection with the manual text simplifications. Hence high levels of insertion of content can still lead to favorable SARI scores, and even improve text statistics like FKGL, without conveying key content of the original text.

4.2. Task 3.2: Abstract-level scientific text simplification

Table 5 shows the Task 3.2 (abstract-level text simplification) results. Again we restrict the table to submissions covering a sufficient number of input abstracts.

We make a number of observations. First, in terms of evaluation measures like SARI we see again similar encouraging performance levels when evaluating against the human reference simplifications. This is partly due to the use of proven sentence-level text simplification models with the output merged back into the entire abstract. Second, there remains room for improvement in capturing the human

Table 4
Results for CLEF 2024 SimpleText Task 3.1 sentence-level text simplification (task number removed from the run_id) on the test set

run_id	count	FKGL	SARI	BLEU	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	578	13.65	12.02	19.76	1.00	1.00	1.00	1.00	0.00	0.00	8.80
Reference	578	8.86	100.00	100.00	0.70	1.06	0.60	0.01	0.27	0.54	8.51
Elsevier_run1	578	10.33	43.63	10.68	0.87	1.06	0.59	0.00	0.45	0.53	8.39
Elsevier_run4	577	11.73	43.14	12.08	0.85	1.00	0.63	0.00	0.37	0.50	8.54
Elsevier_run8	577	12.40	42.95	12.35	0.90	1.02	0.63	0.00	0.35	0.50	8.66
Elsevier_run6	577	12.65	42.88	11.76	0.95	1.00	0.64	0.00	0.38	0.47	8.63
Elsevier_run7	577	12.55	42.87	12.20	0.87	1.00	0.63	0.00	0.35	0.51	8.67
Elsevier_run9	577	12.53	42.61	12.15	0.87	1.00	0.63	0.00	0.35	0.50	8.67
Elsevier_run3	577	11.50	42.58	15.75	0.76	0.98	0.68	0.00	0.23	0.46	8.68
Elsevier_run10	577	12.57	42.49	11.91	0.91	1.02	0.63	0.00	0.34	0.50	8.67
AIIRLab_llama-3-8b_run1	578	8.39	40.58	7.53	0.90	1.37	0.56	0.00	0.48	0.58	8.45
AIIRLab_llama-3-8b_run3	578	9.47	40.36	6.26	1.17	1.52	0.53	0.00	0.53	0.56	8.51
AIIRLab_llama-3-8b_run2	578	10.33	39.76	5.46	1.03	1.19	0.51	0.00	0.60	0.56	8.34
UZHPandas_simple_cot	578	13.74	39.59	3.38	3.44	2.67	0.41	0.00	0.76	0.12	8.61
UZHPandas_simple	578	11.24	39.28	5.67	0.88	0.98	0.52	0.00	0.53	0.62	8.45
Sharingans_finetuned	578	11.39	38.61	18.18	0.83	1.07	0.77	0.11	0.16	0.32	8.70
UZHPandas_selection_sle_cot	578	6.49	38.38	1.03	4.76	6.26	0.30	0.00	0.89	0.14	8.30
UZHPandas_simple_inter_def	578	21.36	38.29	3.13	1.93	0.99	0.46	0.00	0.69	0.33	8.86
UZHPandas_selection_lens_cot	578	6.74	38.16	1.10	4.54	5.88	0.32	0.00	0.87	0.14	8.32
UZHPandas_5Y_target_cot	578	6.39	37.95	0.97	4.73	6.25	0.30	0.00	0.89	0.14	8.30
UZHPandas_selection_lens	578	21.29	37.79	2.71	1.97	1.01	0.44	0.00	0.71	0.34	8.85
UBO_Phi4mini-s	578	8.74	36.78	0.58	18.23	23.48	0.47	0.00	0.66	0.29	8.89
UZHPandas_selection_lens_1	578	7.79	36.72	3.65	0.72	0.98	0.46	0.00	0.54	0.73	8.25
UBO_Phi4mini-sl	578	6.16	36.53	0.61	6.92	9.81	0.38	0.00	0.80	0.42	8.72
UZHPandas_5Y_target_inter_def	578	19.30	36.53	2.27	1.76	1.01	0.45	0.00	0.70	0.41	8.87
UZHPandas_selection_sle	578	6.07	35.30	2.57	0.65	0.98	0.43	0.00	0.56	0.78	8.17
UZHPandas_5Y_target	578	5.94	34.91	2.29	0.66	0.99	0.43	0.00	0.57	0.78	8.17
RubyAiYoungTeam	578	8.76	34.40	15.37	0.60	1.22	0.69	0.03	0.05	0.44	8.71
SONAR_SONARnonlinreg	578	13.14	32.12	18.41	0.97	1.01	0.93	0.13	0.11	0.13	8.73
UAms_GPT2_Check	578	11.47	29.91	15.10	1.02	1.23	0.87	0.14	0.17	0.14	8.68
UAms_GPT2	578	10.91	29.73	13.07	1.30	1.50	0.79	0.06	0.29	0.12	8.63
Arampatzis_T5	578	13.18	28.92	10.66	1.12	1.10	0.72	0.03	0.34	0.37	9.06
UAms_Wiki_BART_Snt	578	12.13	27.45	21.56	0.85	0.99	0.89	0.32	0.02	0.16	8.73
Arampatzis_DistilBERT	578	5.85	19.00	13.56	1.03	3.00	0.95	0.00	0.22	0.11	8.65
UAms_Cochrane_BART_Snt	578	13.22	18.45	19.21	0.95	0.99	0.96	0.59	0.02	0.07	8.77

simplifications more closely, as the BLEU score remains low throughout. Here, the more conservative approaches seem to obtain better scores. Third, we see less extreme values on the other indicators, but still considerable variation in the compression ratio and number of splits, and proportions of addition and deletions. We will investigate how much of the output is grounded in the source sentences and abstracts below.

Many submissions rely on proven sentence-level text simplification approaches, with results closely mirroring those observed for the sentence-level task. It is encouraging to see solid performance for the approaches that perform text simplification at the entire abstracts in one pass. This holds the promise to incorporate the discourse structure, use more complex text simplifications operations such as deletions and merges, and deploy planner-based approaches to the text simplification of long documents.

Table 5
Results for CLEF 2024 SimpleText Task 3.2 abstract-level text simplification (task number removed from the run_id) on the test set

run_id	count	FKGL	SARI	BLEU	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	103	13.64	12.81	21.36	1.00	1.00	1.00	1.00	0.00	0.00	8.88
Reference	103	8.91	100.00	100.00	0.67	1.04	0.60	0.00	0.23	0.53	8.66
AIIRLab_llama-3-8b_run1	103	9.07	43.44	11.73	1.01	1.38	0.51	0.00	0.37	0.56	8.57
AIIRLab_llama-3-8b_run3	103	10.17	43.21	11.03	1.15	1.47	0.52	0.00	0.40	0.51	8.66
Elsevier_run2	103	11.01	42.47	10.54	1.04	1.22	0.51	0.00	0.38	0.55	8.60
AIIRLab_llama-3-8b_run2	103	10.22	42.19	7.99	1.31	1.38	0.48	0.00	0.53	0.52	8.44
Elsevier_run5	103	12.08	42.15	10.96	1.04	1.15	0.52	0.00	0.36	0.53	8.75
Sharingans_finetuned	103	11.53	40.96	18.29	1.20	1.39	0.65	0.00	0.24	0.34	8.80
UBO_Phi4mini-ls	103	8.45	38.79	5.53	1.21	1.75	0.43	0.00	0.40	0.63	8.53
UBO_Phi4mini-l	103	9.96	38.41	10.01	1.29	2.11	0.55	0.00	0.24	0.51	9.03
UAms_GPT2_Check_Abs	103	12.85	36.47	13.12	0.91	0.92	0.59	0.00	0.18	0.45	8.73
UAms_Cochrane_BART_Doc	103	14.46	33.51	9.39	0.65	0.58	0.54	0.04	0.06	0.53	8.80
UAms_Cochrane_BART_Par	103	16.53	31.58	15.40	1.08	0.80	0.67	0.04	0.15	0.32	8.81
UAms_GPT2_Check_Snt	103	11.57	30.71	15.24	1.54	1.70	0.78	0.00	0.27	0.13	8.77
UAms_Wiki_BART_Doc	103	15.68	26.50	15.11	1.51	1.14	0.76	0.01	0.25	0.11	8.79
UAms_Wiki_BART_Par	103	13.11	23.92	19.49	1.39	1.37	0.81	0.01	0.11	0.10	8.86

4.3. Train results

In this section, we show the results over the train data for sentence-level and abstract-level scientific text simplification. This analysis includes those submission retricted to the train data and left out above.

4.3.1. Task 3.1: Sentence-level scientific text simplification

Table 6 shows the sentence-level text simplification results for the train data.

We make the following observations. First, we observed very high performance with SARI scores up to 65% for systems fine-tuned on the train data. Even more striking are very high BLEU scores of over 50%. This is a signal of potential overfitting, although the top performing systems on train still perform reasonably on the new test data. The majority of runs performs similar on train and test, which is according to expectation as most are not particularly trained or fine-tuned on the relatively small set of train sentences and abstracts.

Second, we observe again a clear reduction of FKGL readability, in particular for systems with a high proportion of sentence splits. We make the same proviso that although shorter sentences, and shorter or more common words, is a weak proxy for text complexity, as complex terminology and brief abbreviations may remain and stay opaque for lay users. A very simple grammar is common in youth reading levels, such as target by the popular Newsela-auto [13] data, making FKGL a popular readability score. However, in plain English summaries of scientific text we don't observe such reduction [25].

Third, while we observe higher scores on the train data in Table 4 than on the test data above in Table 4, there seems to be still room for improvement. Throughout the table, we see many low BLEU scores, and very high fractions of additions may risk gratuitous introduction of new content, and hence risk "hallucination."

Table 6
Results for CLEF 2024 SimpleText Task 3.1 sentence-level text simplification (task number removed from the run_id) on the train set

run_id	count	FKGL	SARI	BLEU	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	893	14,30	19,18	38,95	1,00	1,00	1,00	1,00	0,00	0,00	8,72
Reference References	893	11,70	100,00	100,00	0,84	1,07	0,72	0,04	0,21	0,37	8,63
Sharingans_finetuned	714	11,69	64,75	52,53	0,82	1,07	0,73	0,05	0,19	0,37	8,61
Elsevier@SimpleText_run3	714	11,78	46,78	25,55	0,76	0,99	0,68	0,00	0,23	0,47	8,62
Elsevier@SimpleText_run6	714	12,58	44,36	20,64	0,90	1,02	0,64	0,00	0,37	0,47	8,56
Elsevier@SimpleText_run7	714	12,67	43,76	20,51	0,85	1,00	0,63	0,00	0,35	0,50	8,61
Elsevier@SimpleText_run8	714	12,54	43,64	20,69	0,85	1,02	0,63	0,00	0,34	0,50	8,60
Elsevier@SimpleText_run9	714	12,66	43,59	20,33	0,86	1,00	0,63	0,00	0,35	0,51	8,63
Elsevier@SimpleText_run10	714	12,57	43,37	20,29	0,86	1,02	0,63	0,00	0,34	0,50	8,61
Elsevier@SimpleText_run4	714	11,79	43,30	20,05	0,84	1,01	0,62	0,00	0,38	0,52	8,49
Elsevier@SimpleText_run1	714	10,52	41,05	15,56	0,86	1,07	0,59	0,00	0,45	0,53	8,35
Tomislav&Rowan_LLAMA	25	11,84	40,67	4,27	3,94	2,86	0,41	0,00	0,73	0,28	8,36
AIIRLab_Mistral_7B_Instruct_V0.2	893	10,64	39,36	14,07	0,74	1,05	0,58	0,00	0,32	0,58	8,62
UBO_Phi4mini-s	714	8,60	39,27	1,15	17,05	22,28	0,48	0,00	0,65	0,30	8,85
UZH_Pandas_simple_with_cot	714	13,81	38,73	4,62	3,42	2,74	0,41	0,00	0,77	0,12	8,57
AIIRLab_llama-3-8b_run1	714	8,32	38,53	11,75	0,89	1,39	0,56	0,00	0,46	0,59	8,39
AIIRLab_llama-3-8b_run3	714	9,28	37,89	9,35	1,12	1,51	0,54	0,00	0,52	0,58	8,45
UZH_Pandas_simple_with_intermediate_definitions	714	21,60	36,71	5,10	1,91	0,99	0,46	0,00	0,70	0,34	8,83
PiTheory_T5	97	9,94	36,53	11,02	1,37	1,53	0,63	0,00	0,48	0,30	8,51
team1_Petra_and_Regina_task3_ST	893	8,42	36,19	19,72	0,58	1,29	0,66	0,03	0,05	0,47	8,66
UBO_RubyAiYoungTeam	893	8,42	36,19	19,72	0,58	1,29	0,66	0,03	0,05	0,47	8,66
SONAR_SONARnonlinreg	714	13,61	36,01	29,89	0,96	1,02	0,92	0,12	0,10	0,13	8,65
UBO_RubyAiYoungTeam	714	8,67	35,97	19,73	0,59	1,27	0,68	0,04	0,05	0,45	8,67
UZH_Pandas_simple	714	10,91	35,56	8,27	0,84	0,99	0,52	0,00	0,52	0,64	8,37
UZH_Pandas_selection_with_lens	714	21,45	35,56	4,26	1,91	1,00	0,44	0,00	0,71	0,35	8,84
AIIRLab_llama-3-8b_run2	714	10,43	35,47	6,87	1,00	1,18	0,52	0,00	0,59	0,58	8,29
UAms_GPT2_Check	714	11,87	35,21	27,35	1,02	1,22	0,87	0,11	0,17	0,14	8,59
UAms_GPT2	714	11,21	34,73	23,69	1,28	1,47	0,79	0,05	0,28	0,12	8,56
UZH_Pandas_selection_with_lens_cot	714	6,41	34,32	1,34	4,44	6,16	0,32	0,00	0,88	0,14	8,28
FRANE_AND_ANDREA_t5	893	8,57	34,20	33,58	0,87	1,72	0,82	0,17	0,11	0,24	8,73
Dajana&Kathy_t5	893	8,57	34,20	33,58	0,87	1,72	0,82	0,17	0,11	0,24	8,73
UZH_Pandas_5Y_target_with_intermediate_definitions	714	19,83	34,20	3,40	1,74	0,99	0,45	0,00	0,71	0,41	8,86
UAms_Wiki_BART_Snt	714	12,34	34,19	37,18	0,83	0,99	0,88	0,29	0,02	0,19	8,64
UZH_Pandas_selection_with_sle_cot	714	6,23	34,07	1,15	4,66	6,51	0,31	0,00	0,89	0,14	8,28
UZH_Pandas_5Y_target_with_cot	714	6,16	33,98	1,13	4,66	6,53	0,30	0,00	0,89	0,14	8,26
Arampatzis_T5	893	12,15	33,12	21,85	1,09	1,25	0,72	0,03	0,35	0,38	9,07
UBO_Phi4mini-sl	714	7,02	32,94	1,02	5,49	7,03	0,39	0,00	0,79	0,44	8,69
UZH_Pandas_selection_with_lens	714	7,85	32,31	4,96	0,72	0,99	0,46	0,00	0,54	0,73	8,21
UZH_Pandas_selection_with_sle	714	6,22	30,25	2,45	0,66	0,99	0,43	0,00	0,56	0,78	8,18
UZH_Pandas_5Y_target	714	6,02	29,88	2,03	0,66	1,00	0,42	0,00	0,58	0,79	8,19
UAms_Cochrane_BART_Snt	714	13,74	26,70	36,69	0,94	0,99	0,95	0,56	0,03	0,08	8,67
Arampatzis_DistilBERT	893	6,07	26,42	29,20	1,03	2,94	0,95	0,00	0,21	0,10	8,63

4.3.2. Task 3.2: Abstract-level scientific text simplification

Table 7 shows the abstract-level text simplification results for the train data.

We make the following observations. First, we observe higher scores for systems who deploy

Table 7
Results for CLEF 2024 SimpleText Task 3.2 abstract-level text simplification (task number removed from the run_id) on the train set

run_id	**count**	**FKGL**	**SARI**	**BLEU**	**Compression ratio**	**Sentence splits**	**Levenshtein similarity**	**Exact copies**	**Additions proportion**	**Deletions proportion**	**Lexical complexity score**
Source	175	14,30	19,53	39,95	1,00	1,00	1,00	1,00	0,00	0,00	8,88
Reference References	175	11,80	100,00	100,00	0,80	1,04	0,70	0,00	0,20	0,40	8,75
Sharingans_finetuned	119	11,36	60,65	45,74	0,78	1,07	0,68	0,00	0,20	0,41	8,71
Mistral-7B-Instruct-V0.2	175	12,85	40,66	16,52	0,79	0,92	0,60	0,00	0,29	0,51	8,83
AIIRLab_llama-3-8b_run3	119	9,77	40,62	15,04	0,70	1,03	0,55	0,00	0,31	0,57	8,59
Elsevier@SimpleText_run5	119	12,16	40,30	14,23	0,71	0,84	0,55	0,00	0,30	0,57	8,62
UBO_Phi4mini-l	119	9,39	39,95	14,41	1,87	3,23	0,56	0,00	0,18	0,56	8,95
AIIRLab_llama-3-8b_run1	119	8,49	39,51	13,00	0,65	1,03	0,54	0,00	0,31	0,61	8,47
Elsevier@SimpleText_run2	119	11,09	39,32	12,43	0,68	0,86	0,53	0,00	0,31	0,60	8,56
Tomislav&Rowan_LLAMA	20	10,48	37,61	15,26	1,13	1,70	0,53	0,00	0,45	0,48	8,73
AIIRLab_llama-3-8b_run2	119	10,42	37,13	9,95	0,82	1,01	0,51	0,00	0,47	0,57	8,37
UAms_GPT2_Check_Abs	119	12,75	36,68	16,48	0,59	0,66	0,60	0,01	0,11	0,50	8,61
UAms_GPT2_Check_Snt	119	11,88	35,97	28,86	1,00	1,22	0,85	0,01	0,18	0,15	8,71
UAms_Cochrane_BART_Par	119	16,15	35,12	26,23	0,70	0,59	0,70	0,04	0,08	0,36	8,72
UBO_Phi4mini-ls	119	8,71	34,81	7,23	0,89	1,50	0,44	0,00	0,34	0,68	8,57
Arampatzis_T5	175	11,39	33,94	9,61	0,48	0,60	0,53	0,00	0,07	0,59	8,90
UAms_Wiki_BART_Doc	119	16,45	33,36	28,35	1,01	0,83	0,81	0,00	0,18	0,15	8,73
UAms_Cochrane_BART_Doc	119	14,78	33,23	9,55	0,40	0,40	0,52	0,03	0,01	0,61	8,76
UAms_Wiki_BART_Par	119	13,26	30,31	36,76	0,89	1,00	0,88	0,01	0,03	0,13	8,81
Arampatzis_DistilBERT	175	11,24	25,17	30,75	1,02	1,67	0,96	0,00	0,16	0,09	8,78

finetuning, which doesn't seem to generalize to the unseen test evaluation before. Most systems, however, wer not particularly trained or finetuned on the train data and show similar performance on both train and test.

Second, we observe solid performance for the more complex document-level scientific text simplification task, but this is due to many systems deploying proving sentence-level text simplification technology with merging the sentence-level output back into complete abstracts.

Third, while a sentence-level approach to document-level text simplification is a pragmatic choice and viable strategy, several model perform direct abstract-level or paragraph-level taking the discourse structure and more complex sentences reordering and deletion into account. These document-level text simplification approach tend to lead to far greater compression, including whole sentence deletions, making their output far more succinct than sentence-level approaches to document-level text simplification. Giving their succinct output, and in light of the sentence-level constructed human reference simplifications, the scores of direct abstract-level or paragraph-level approaches is impressive. Further research in such document-level text simplification approaches would be important in the future of the CLEF SimpleText track.

5. Analysis

This section provides further analysis of the submitted runs, and the task as whole.

Table 8
Example of SimpleText Task 3 output versus input: ~~deletions~~, insertions, and whole sentence insertions

Topic	Document	Output
G01.1	130055196	As various kinds of output devices emerged , such as highresolution printers or a display of PDA (Personal Digital Assistant) ~~, the~~ . The importance of high-quality resolution conversion has been increasing . This paper proposes a new method for enlarging an image with high quality . It will involve using a combination of high-speed imaging and high-resolution video . One of the ~~largest~~ biggest problems on image enlargement is the exaggeration of the jaggy edges . This is especially true when the image is enlarged , as in this case . To remedy this problem , we propose a new interpolation method ~~, which~~ . This method uses artificial neural network to determine the optimal values of interpolated pixels . The experimental results are shown and evaluated . The results are compared to other studies and found to be inconclusive . The effectiveness of our methods is discussed by comparing with the conventional methods . Our methods are designed to help people with mental health problems , not just as a way to cure them .

5.1. Human Evaluation

Due to the delayed submission deadline, as well as, follow-up correspondence with teams on partial or incorrect output, the manual annotation of system output has been limited to a small sample, and is still ongoing. We report here only initial observations from the translation professionals conducting this analysis, based on the expectation of what a professional editor would provide as reference output. We looked in particular at the novel document-level simplifications of the entire abstract, and it's coherence and discourse structure.

First, and foremost, something is working. The automatic text simplifications are generally of impressive quality despite the remaining limitations that are the focus of this section. The fluency and language variation is impressive, and far exceeds earlier language generation technology often reflecting the protocol, and template or rule-based system underlying it.

Second, changes can be unnecessary nor helpful. Frequently, as we observed in our work on the project last year [12], the information is written in another way but does not offer simplification. Sometimes the vocabulary does no change but is simply rearranged.

Third, discourse structure matters. In other examples the resulting text is not shaped as a whole, with a proper beginning middle and end, but is reorder to the detriment of clarity. For example, the first sentence of the "simplified" abstract can contain a reference back to information already given. Another example: start of a first sentence with "*However, ...*" in the simplification when source text started "*It is the purpose of this study, ...*" or with "*For example, ...*" when the original first sentences presented the subject.

Fourth, brevity is not always clearer. Although some examples shorten the sentences within an abstract, thus technically simplifying, their interrelation is not necessarily maintained, producing a choppy style. Better results were produced when the new text was split into subsections dedicated to particular subtopics, including their explanation.

Fifth, gratuitous additions are problematic. Another type of problem is illustrated by the creation of a cumbersome nominal group "*the 21st Century managed care needs of patients, ...*" which does not exist in the original, where we instead had an evocative example: "*the emergency room at home.*" Here though, both things belong in the same domain. Elsewhere, seeming hallucinations appeared, for example, through the addition of an off-topic sentence. For example, to an abstract about digital tools to aid Parkinson's sufferers, we found the following last sentence added during simplification: "*It includes advice on how to manage consultant work, such as research and development .*" Although, in terms of meaning, this has no equivalent in the source text, the source text starting sentence was: "*The paper also discusses how a practitioner can accomplish UCSD in the context of product development and consultant work.*", which mentions the topic in a different context.

Table 9
Analysis of SimpleText Task 3.1: Spurious generation

Run	# Input Sentences	Spurious Content	
		Number	Fraction
AB/DVP_SequentialLSTM	4797	4788	1.00
AIIRLab_Mistral_7B_Instruct_V0	779	23	0.03
AIIRLab_llama-3-8b_run3	4797	129	0.03
AIIRLab_llama-3-8b_run3	4797	381	0.08
AIIRLab_llama-3-8b_run3	4797	489	0.10
Dajana/Kathy_t5	779	80	0.10
Elsevier@SimpleText_run1	4797	50	0.01
Elsevier@SimpleText_run10	4796	49	0.01
Elsevier@SimpleText_run3	4795	36	0.01
Elsevier@SimpleText_run4	4795	32	0.01
Elsevier@SimpleText_run6	4796	46	0.01
Elsevier@SimpleText_run7	4796	41	0.01
Elsevier@SimpleText_run8	4796	46	0.01
Elsevier@SimpleText_run9	4796	43	0.01
FRANE_AND_ANDREA_t5	779	80	0.10
SONAR_SONARnonlinreg	4797	15	0.00
Sharingans_finetuned	4797	51	0.01
UAms-1_Cochrane_BART_Snt	4797	25	0.01
UAms-1_GPT2	4797	1390	0.29
UAms-1_GPT2_Check	4797	3	0.00
UAms-1_Wiki_BART_Snt	4797	14	0.00
UBO_Phi4mini-s	4797	2055	0.43
UBO_Phi4mini-sl	4797	1822	0.38
UBO_RubyAiYoungTeam	779	169	0.22
UBO_RubyAiYoungTeam	4797	1051	0.22
UZHPandas_5Y_target	4797	2607	0.54
UZHPandas_5Y_target_cot	4797	3383	0.71
UZHPandas_5Y_target_intermediate_defs	4797	365	0.08
UZHPandas_selection_lens	4797	283	0.06
UZHPandas_selection_lens_cot	4797	3265	0.68
UZHPandas_selection_sle	4797	2311	0.48
UZHPandas_selection_sle_cot	4797	3362	0.70
UZHPandas_simple	4797	166	0.03
UZHPandas_simple_cot	4797	2915	0.61
UZHPandas_simple_intermediate_defs	4797	79	0.02
Arampatzis_DistilBERT	5576	5575	1.00
Arampatzis_T5	5576	336	0.06
team1_Petra_and_Regina_ST	779	169	0.22

5.2. Spurious or overgeneration

We conduct a deeper analysis of how much of the generated simplified output sentences and abstracts can be traced to the source input. In particular, we look at spurious generated content and it's prevalence in the submitted generated text simplifications. This content is at risk of being introduced gratuitously by the generative model, and what is informally referred to as "hallucinations."

Earlier in Table 2, we showed an example of a human reference simplification, combining the input sentences belonging to the abstract of the document $id = 130055196$ retrieved for query G01.1. We can do the same for the automatically generated scientific text simplifications. We show again the deletions and insertions relative to the source input sentences. Table 8 shows an example output simplification of one of the participating teams, for the same input sentences as in Table 2 above. Most simplifications are revisions of the input, but we also observe that sometimes an entire sentence is inserted (shown as xxx in Table 8). The example in Table 8 is an extreme case picked to illustrate both the importance and complexity of detecting such spurious content.

We provide a detailed analysis quantifying the prevalence of spurious content in the CLEF 2024

Table 10
Results for SimpleText Task 3.2: Spurious generation

Run	# Input Abstracts	Spurious Content	
		Number	Fraction
AIIRLab_llama-3-8b_run1	782	56	0.07
AIIRLab_llama-3-8b_run2	782	121	0.15
AIIRLab_llama-3-8b_run3	782	98	0.13
Elsevier@SimpleText_run2	782	28	0.04
Elsevier@SimpleText_run5	782	30	0.04
Mistral-7B-Instruct-V0	119	6	0.05
Sharingans_finetuned	782	59	0.08
UAms-2_Cochrane_BART_Doc	782	2	0.00
UAms-2_Cochrane_BART_Par	782	28	0.04
UAms-2_GPT2_Check_Abs	782	1	0.00
UAms-2_GPT2_Check_Snt	782	111	0.14
UAms-2_Wiki_BART_Doc	782	74	0.09
UAms-2_Wiki_BART_Par	782	46	0.06
UBO_Phi4mini-s	782	102	0.13
UBO_Phi4mini-sl	782	104	0.13
Arampatzis_DistilBERT	901	118	0.13
Arampatzis_T5	901	5	0.01

SimpleText Task 3 submissions. Table 9 quantifies how often such spurious generation occurs. We re-aligned the generated output with the original source sentences, and flag here only entire output sentences that do not share a single token with the input. Our analysis reveals that the amount of spurious content is varying but far from infrequent. A total of 17 out 36 submissions (47%) have spurious whole sentences in at least 10% of the input sentences. In fact, 14 (39%) submissions in at least 20% of the input, and 7 (19%) submissions in at least 50% of the input sentences. The detection of non-aligned output sentences is indicative but imperfect. For example, a significant reordering of content may lead to false positives in rare cases, and unusual tokenization or formatting may affect the alignment with the source even systematically. Note also that the detected additions may introduce helpful background knowledge or other useful information to contextualize the information in the source sentences.

Table 10 quantifies how often such spurious generation occurs for the abstract-level output. Here we look again at the spurious output at the end of the input abstract, rather than conducting a sentence-level analysis as done above. Aligning longer text is more complex than sentences. For those generating true paragraph or document level simplifications, we observe more variation involving content of multiple input sentences leading to a more complex alignment. Hence we focus on detecting spurious content at the end of the generated abstract. As a result, for those aggregating sentence-level output merged into the abstracts, we are only able to detect spurious content for the final sentence.

We make a number of observations based on our analysis in this section. First, the fraction of sentences with spurious content is very low for some submissions, however, for other submissions, the fraction is very substantial. Second, the standard evaluation measures used for text simplification, and in fact for any text generation task in NLP, do not take this aspect into account. A submission with significant spurious content can still obtain very high text overlap with the reference, and hence obtain a very high performance score. Third, and more generally, human evaluation and this type of analysis feel crucial to accurately evaluate generative models for the NLP and IR challenges addressed in our Track and in CLEF in general.

6. Conclusions

The paper provides an overview of the CLEF 2024 SimpleText Task 3: Text Simplification, which focuses on the simplification of scientific text. The objective of the task is to simplify either the separate sentences or the entire scientific abstracts in order to enhance their accessibility and comprehensibility for a general audience. We highlighted the key aspects and goals of the task within the broader context of the CLEF 2024 SimpleText track [11].

Our main findings are the following: First, we observe competitive performance for scientific text simplification, both on evaluation against the human reference simplifications and on text statistics such as FKGL readability score. Second, the abstract-level text simplification results is a mixture of sentence-level and passage-level text simplification approaches. Third, our analysis reveals a very high and varying range of spurious text generation, not detected by standard evaluation measures, and a major concern in the use of these model in a real-world setting. More generally, almost all participants use generative models (for the task, the track, and CLEF in general), and the track offers a unique setting to study some of the inherent limitations of generative models.

The main aim of our task, the track, and the CLEF evaluation forum as a whole, is i) to foster a community of IR, NLP, and AI researchers working together on the important task of making science more accessible for everyone, and ii) to construct corpora and evaluation resources to stimulate research on scientific text summarization and simplification. In terms of a building a community researching scientific text summarization and simplification, the task saw a record attendance in 2024: due to the additional abstract level task we received 83 runs from 15 teams, the largest number of participating teams ever. In fact, the community is broadening beyond CLEF and raising general interest in generative scientific text summarization and simplification [1].

Within the CLEF 2024 SimpleText Task 3, we have constructed extensive corpora and manually labeled evaluation data for scientific text simplification. Specifically, we added in 2024 a a parallel corpus of manually simplified sentences and abstracts from the scientific literature:

- Train, sentence level: 958 source sentences from scientific abstracts paired with corresponding human reference simplifications.
- Test, sentence level: 578 source sentences from scientific abstracts paired with corresponding human reference simplifications.
- Train, abstract level: 175 source scientific abstracts paired with corresponding human reference simplifications.
- Test, abstract level: 103 source scientific abstracts paired with corresponding human reference simplifications.

These reusable corpora and evaluation resources are available to participants and other researchers who want to work on the important problem of making scientific information open and easily accessible for everyone.

Acknowledgments

This track would not have been possible without the great support of numerous individuals. We want to thank in particular the colleagues and the students who participated in data construction and evaluation. Please visit the SimpleText website for more details on the track.[1]

Liana Ermakova is funded by the French National Research Agency (ANR) *Automatic Simplification of Scientific Texts* project (ANR-22-CE23-0019-01),[2] and the MaDICS research group.[3] Jaap Kamps is partly funded by the Netherlands Organization for Scientific Research (NWO CI # CISC.CC.016, NWO NWA # 1518.22.105), the University of Amsterdam (AI4FinTech program), and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

[1] https://simpletext-project.com/
[2] https://anr.fr/Project-ANR-22-CE23-0019
[3] https://www.madics.fr/ateliers/simpletext/

References

[1] G. M. D. Nunzio, F. Vezzani, L. Ermakova, H. Azarbonyad, J. Kamps (Eds.), Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024. URL: https://aclanthology.org/2024.determit-1.0.

[2] S. Štajner, H. Saggio, M. Shardlow, F. Alva-Manchego (Eds.), Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023. URL: https://aclanthology.org/2023.tsar-1.0.

[3] S. Štajner, H. Saggion, D. Ferrés, M. Shardlow, K. C. Sheang, K. North, M. Zampieri, W. Xu (Eds.), Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Virtual), 2022. URL: https://aclanthology.org/2022.tsar-1.0.

[4] H. Saggion, S. Stajner, D. Ferrés, K. C. Sheang (Eds.), 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 (SEPLN2021), Online (initially located in Málaga, Spain), September 21st, 2021, volume 2944 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: https://ceur-ws.org/Vol-2944.

[5] L. Ermakova, P. Bellot, P. Braslavski, J. Kamps, J. Mothe, D. Nurbakova, I. Ovchinnikova, E. SanJuan, Overview of simpletext 2021 - CLEF workshop on text simplification for scientific information access, in: K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings, volume 12880 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 432–449. URL: https://doi.org/10.1007/978-3-030-85251-1_27. doi:10.1007/978-3-030-85251-1_27.

[6] L. Ermakova, E. SanJuan, J. Kamps, S. Huet, I. Ovchinnikova, D. Nurbakova, S. Araújo, R. Hannachi, É. Mathurin, P. Bellot, Overview of the CLEF 2022 simpletext lab: Automatic simplification of scientific texts, in: A. Barrón-Cedeño, G. D. S. Martino, M. D. Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings, volume 13390 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 470–494. URL: https://doi.org/10.1007/978-3-031-13643-6_28. doi:10.1007/978-3-031-13643-6_28.

[7] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, O. Augereau, J. Kamps, Overview of the CLEF 2023 simpletext lab: Automatic simplification of scientific texts, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings, volume 14163 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 482–506. URL: https://doi.org/10.1007/978-3-031-42448-9_30. doi:10.1007/978-3-031-42448-9_30.

[8] E. SanJuan, S. Huet, J. Kamps, L. Ermakova, Overview of the CLEF 2024 SimpleText task 1: Retrieve passages to include in a simplified summary, in: [26], 2024.

[9] G. M. Di Nunzio, F. Vezzani, V. Bonato, H. Azarbonyad, J. Kamps, L. Ermakova, Overview of the CLEF 2024 SimpleText task 2: Identify and explain difficult concepts, in: [26], 2024.

[10] J. D'Souza, S. Kabongo, H. B. Giglou, Y. Zhang, Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the State-of-the-Art in Scholarly Publications, in: [26], 2024.

[11] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, G. M. Di Nunzio, F. Vezzani, J. D'Souza, J. Kamps, Overview of the CLEF 2024 SimpleText track: Improving access to scientific texts for everyone, in: [27], 2024.

[12] L. Ermakova, S. Bertin, H. McCombie, J. Kamps, Overview of the CLEF 2023 simpletext task 3: Simplification of scientific texts, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org,

2023, pp. 2855–2875. URL: https://ceur-ws.org/Vol-3497/paper-240.pdf.
[13] W. Xu, C. Callison-Burch, C. Napoles, Problems in current text simplification research: New data can help, Transactions of the Association for Computational Linguistics 3 (2015) 283–297. URL: https://aclanthology.org/Q15-1021. doi:10.1162/tacl_a_00139.
[14] D. P. Varadi, A. Bartulović, SimpleText 2024: Scientific Text Made Simpler Through the Use of AI, in: [26], 2024.
[15] N. Largey, R. Maarefdoust, S. Durgin, B. Mansouri, AIIR Lab Systems for CLEF 2024 SimpleText: Large Language Models for Text Simplification, in: [26], 2024.
[16] A. Capari, H. Azarbonyad, G. Tsatsaronis, Z. Afzal, Enhancing Scientific Document Simplification through Adaptive Retrieval and Generative Models, in: [26], 2024.
[17] R. Elagina, P. Vučić, AI Contributions to Simplifying Scientific Discourse in SimpleText 2024, in: [26], 2024.
[18] S. M. Ali, H. Sajid, O. Aijaz, O. Waheed, F. Alvi, A. Samad, Improving Scientific Text Comprehension: A Multi-Task Approach with GPT-3.5 Turbo and Neural Ranking, in: [26], 2024.
[19] R. Mann, T. Mikulandric, CLEF 2024 SimpleText Tasks 1-3: Use of LLaMA-2 for text simplification, in: [26], 2024.
[20] J. Bakker, G. Yüksel, J. Kamps, University of Amsterdam at the CLEF 2024 SimpleText Track, in: [26], 2024.
[21] B. Vendeville, L. Ermakova, P. De Loor, UBO NLP report on the SimpleText track at CLEF 2024, in: [26], 2024.
[22] A. Michail, P. S. Andermatt, T. Fankhauser, Scientific Text Simplification Using Multi-Prompt Minimum Bayes Risk Decoding: Examining MBR's Decisions, in: [26], 2024.
[23] A. Michail, P. S. Andermatt, T. Fankhauser, Scientific text simplification using multi-prompt minimum bayes risk decoding: Simpletext best of labs in CLEF 2023, in: [27], 2024.
[24] L. Ermakova, J. Kamps, Complexity-aware scientific literature search: Searching for relevant and accessible scientific text, in: G. M. D. Nunzio, F. Vezzani, L. Ermakova, H. Azarbonyad, J. Kamps (Eds.), Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 16–26. URL: https://aclanthology.org/2024.determit-1.2.
[25] J. Bakker, J. Kamps, Plan-guided simplification of biomedical documents, in: Under Submission, 2024.
[26] G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
[27] L. Goeuriot, G. Q. Philippe Mulhem, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, 2024.