Overview of the CLEF 2023 SimpleText Task 2: Difficult Concept Identification and Explanation

Liana Ermakova[1], Hosein Azarbonyad[2], Sarah Bertin[1] and Olivier Augereau[3]

[1]*Université de Bretagne Occidentale, HCTI, Brest, France*
[2]*Elsevier, The Netherlands*
[3]*ENIB, Lab-STICC UMR CNRS 8265, France*

Abstract

In this paper, we present an overview of the *"Task 2: What is unclear? Difficult concept identification and explanation"* within the context of the Automatic Simplification of Scientific Texts (SimpleText) lab, run as part of CLEF 2023. The primary objective of the SimpleText lab is to advance the accessibility of scientific information by facilitating automatic text simplification, thereby promoting a more inclusive approach to scientific knowledge dissemination. *Task 2* focuses on complexity spotting within scientific texts (passage). Thus, the goal is to detect the terms/concepts that require specific background knowledge for understanding the passage, assess their complexity for non-experts, and provide explanations for these detected difficult concepts. A total of 39 submissions were received for this task, originating from 12 distinct teams. In this paper, we describe the data collection process, task configuration, and evaluation methodology employed. Additionally, we provide a brief summary of the various approaches adopted by the participating teams.

Keywords

automatic text simplification, terminology, background knowledge, scientific article, science popularization, contextualization, term difficulty

1. Introduction

Scientific literature has become more accessible to the general public through digitalization. However, there still exists a significant barrier preventing individuals from accessing objective scientific knowledge directly from the original sources. One of the main challenges stems from the high complexity of scientific texts, which poses difficulties for non-experts due to the lack of necessary background knowledge, including the comprehension of specialized terminology. Even for native speakers, understanding terminology outside their area of expertise can be challenging. However, individuals with a basic set of terms acquired through secondary and college education can comprehend popular science publications to a certain extent. *Comprehension of the term* implies grasping the concept it represents without the need for an explicit definition. To understand a concept, it often requires incorporating it into a structured system within our semantic memory, which may necessitate additional knowledge beyond what we have already

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

liana.ermakova@univ-brest.fr (L. Ermakova)
https://simpletext-project.com/ (L. Ermakova)
0000-0002-7598-7474 (L. Ermakova); 0000-0002-9661-3762 (O. Augereau)

learned.

Text simplification techniques can play a crucial role in enabling readers to stay informed about scientific advancements. Traditional methods of simplification aim to eliminate complex terms and structures in order to enhance readability [1]. However, this approach may not always be feasible, particularly when dealing with scientific literature. In such cases, readers relying on popular science publications draw upon their experience in processing new information. They can identify instances where they require definitions or clarifications for unfamiliar terms, as their understanding of the underlying concepts may be limited. This recognition of the need for additional explanations or clarifications reflects readers' awareness of their own comprehension gaps in relation to unfamiliar terminology which is perceived as a difficulty.

We argue that a text simplification method should offer essential information required for understanding complex scientific concepts to address the issue of inadequate background knowledge hindering proper comprehension [2]. This objective is one of the focal points of the CLEF 2023 SimpleText lab. Although there have been notable advancements in automatic text simplification, such as the work by Maddela et al. on controllable simplification [3], there is still an ongoing challenge in automatically enhancing the comprehensibility of scientific texts and adapting them to different target audiences.

The CLEF 2023 SimpleText track[1] is a new evaluation lab that follows up on the CLEF 2021 SimpleText Workshop [4] and CLEF 2022 SimpleText Track [5]. The track offers valuable data and benchmarks to facilitate discussions on the challenges associated with automatic text simplification. It presents an interconnected framework that encompasses various tasks, providing a comprehensive view of the complexities involved:

Task 1: What is in (or out)? Selecting passages to include in a simplified summary.

Task 2: What is unclear? Difficult concept identification and explanation (definitions, abbreviation deciphering, context, applications,..).

Task 3: Rewrite this! Given a query, simplify passages from scientific abstracts.

This paper focuses on the second task of complexity spotting. The goal of this task is to detect difficult terms and provide contextual explanations for them. Identifying and effectively explaining difficult terms is crucial for promoting accessibility and comprehension of scientific texts. Please refer for details of the other tasks to the overview papers of Task 1 [6] and Task 3 [7], as well as the Track overview paper [8].

The rest of this paper is structured in the following way. A comprehensive description of the Task 2 is presented in Section 2. Following that, Section 3 provides an overview of the dataset used, including its composition, size, and relevant characteristics. In Section 4, the paper discusses the evaluation metrics employed to assess the performance of the participants' runs. Section 5 delves into the details of the systems and approaches employed by the participants. In Section 6, we discuss the results of the official submissions. We end with Section 7 discussing the results and findings, and lessons for the future.

[1] https://simpletext-project.com

2. Task description

The objective of Task 2 is twofold, identification of difficult that require contextualization through definition, example, and/or use case, as well as the provision of clear and informative explanations for these concepts. Consequently, the task can be divided into two subtasks:

- to retrieve up to 5 difficult terms in a given passage from a scientific abstract;
- to provide an explanation (one/two sentences) of these difficult terms (e.g. definition, abbreviation deciphering, example, etc.).

In the context of the SimpleText track, difficult terms are defined as words or phrases that present challenges for readers due to their complexity, specialized meanings, or technical nature. These terms require additional explanation or clarification to ensure a better understanding for readers who may not be familiar with them. By providing explanations for such difficult terms, readers can overcome the potential obstacles they pose and enhance their comprehension of the text. Difficult terms often involve scientific jargon, complex theories, mathematical equations, or intricate scientific concepts that may be unfamiliar to the general reader or even to experts in other scientific domains.

Participants in Task 2 are required to submit a ranked list of difficult terms for each passage, along with corresponding difficulty scores on a scale of 0 to 2. A score of 2 indicates the highest level of difficulty, whereas a score of 0 implies that the meaning of the term can be inferred or guessed. Optionally, participants can provide definitions for the identified difficult terms. It is important to note that passages (sentences) are treated as independent entities, meaning that repetition of difficult terms across multiple passages is allowed and evaluated separately. Table 1 serves as a reference, providing examples that illustrate different levels of term difficulty.

3. Data

3.1. Datasets for Task 2.1

As part of the task, participants were supplied with a **training set** consisting of 203 pairs of sentences and their corresponding scientific terms with ground truth annotations of difficulty scores for each term on a scale of 0-2. These sentence and difficult term pairs were extracted from relevant abstracts obtained from Task 1 [9, 10].

To build the **test set** for Task 2.1, a total of 116,763 sentences were extracted from the DBLP abstracts. Then, a subset of 1,262 unique sentences was manually evaluated to assess the performance of various models in terms of their capability to identify difficult terms and assign appropriate difficulty scores. To obtain a comprehensive evaluation, a pooling mechanism was implemented, resulting in the annotation of 5,142 distinct pairs of sentence-term combinations. Each evaluated source sentence contained the aggregated results from all participating participants. This process ensured a reliable and robust assessment of the performance of different models in detecting difficult terms and estimating their difficulty scores.

To promote a degree of overlap among the partial runs submitted by participants, a set of three test sets was provided: small, medium, and large. It was anticipated that participants

Table 1
Examples of the term difficulty scale used for evaluation: grades 0-2. Difficult terms are highlighted with the green color

Grade	Non-abbreviated (ordinary) term	Abbreviation
2	“We have proven that *transfer learning* is not only applicable in this field, but it requires smaller well-prepared training datasets, trains significantly faster and reaches similar accuracy compared to the original method, even improving it on some aspects.” "The entropy is derived using *singular value decomposition* of the components of stock market indices in financial markets from selected developed economies, i.e., France, Germany, the United Kingdom, and the United States..” “ *Blockchain* as a new technology has created a great amount of hype and hope for different applications.”	“The steering commands from the source and target network are finally merged according to the *LDL* and the merged command is utilized for controlling a car in the target domain.” “Various machine learning techniques like Random Forest, *SVM* as well as deep learning models has been proposed for classifying traffic signs.” "We compared *XCSFHP* to *XCSF* on several problems."
1	“In this paper, we present the development of a *remote server* that provides a user-friendly access to advanced electrocardiographic (ECG) signal processing techniques.” “An attacker can obtain the password, *private-key* , and public-key of the user.” " *Cloud computing* provides an effective business model for the deployment of IT infrastructure, platform, and software services."	*NIST* (The National Institute of Standards and Technology) in “Recently *NIST* has published the second draft document of recommendation for the entropy sources used for random bit generation.” "Applications to increase the functionality of *PDA* are constantly being developed, and occasionally application software must be installed."
0	The *World Wide Web* is a potentially powerful channel for misinformation.” “On the other hand, a 3dimensional (3D) map, which is one of major themes in machine vision research, has been utilized as a simulation tool in city and *landscape planning* , and other engineering fields.”	*2D* (2-dimensional), *3D* (3-dimensional) *maps* as in “The *3D maps* will give more intuitive information compared to conventional 2-dimensional (*2D*) ones.”

would prefer employing Language Models (LLMs), which could lead to the generation of partial runs due to the limitations in efficiency associated with these models. By offering different test sets, we aimed to accommodate diverse computational limitations and facilitate the participation of various approaches, including those leveraging LLMs.

The small dataset was embedded within the medium dataset, and the medium dataset was, in turn, encompassed within the large dataset. By evaluating systems on the small test sets, we ensured some common ground for comparing the partial runs generated by different participants. This approach facilitated the comparison and analysis of system outputs, despite being derived from different test sets.

Inclusion of the train data within the small dataset allowed for a comparison of system performance on both training and testing data. This integration of train data facilitated evaluating how well the systems could generalize to unseen test data by assessing their performance on familiar training examples. Such a comparison provided valuable insights into the effectiveness and robustness of the systems across different datasets.

3.2. Datasets for Task 2.2

For Task 2.2, the **training set** encompasses the same set of 203 difficult terms as in Task 2.1. However, in Task 2.2, the training set is augmented with the addition of corresponding definitions for each of these difficult terms.

To evaluate the performance of the submitted runs for this task, a **test set** comprising approximately 800 terms with ground truth definitions is utilized. This test set serves as the benchmark against which the performance of the participants' systems is measured. The runs were assessed by comparing the outputs of the systems with the ground truth definitions. The utilization of a substantial number of terms in the test set ensures a comprehensive evaluation of the systems' performance in interpreting and providing accurate definitions for the given terms. From this set, $\sim$300 terms are selected for annotation using a pooling mechanism, ensuring that the test set contains a sufficient number of annotated samples for most runs. The test set comprises a total of 15,056 sentences that contain at least one of these terms. These sentences are utilized for evaluating the performance of the submitted runs. For the evaluation of abbreviation expansion, a set of $\sim$1K abbreviations is manually annotated. Additionally, an extra 4,374 abbreviations are extracted using the Schwartz and Hearst [11] algorithm from the sentences in Task 1, resulting in a total of $\sim$5K abbreviations. The final test set consists of 38,416 sentences that contain at least one of these abbreviations, and this set of sentences is employed for the final evaluation of this subtask.

3.3. Input format

The train and the test data are provided in JSON and TSV formats with the following fields:

snt_id a unique passage (sentence) identifier

doc_id a unique source document identifier

query_id a query ID

query_text difficult terms should be extracted from sentences with regard to this query

source_snt passage text

Input example:

```
[{"query_id":"G14.2",
  "query_text":"end to end encryption",
  "doc_id":"2884788726",
  "snt_id":"G14.2_2884788726_2",
  "source_snt":"However, in information-centric networking (ICN) the end-to-end
↪  encryption makes the content caching ineffective since encrypted content stored
↪  in a cache is useless for any consumer except those who know the encryption
↪  key."},

 {"snt_id":"G06.2_2548923997_3",
  "doc_id":2548923997,
  "query_id":"G06.2",
  "query_text":"self driving",
  "source_snt":"These communication systems render self-driving vehicles vulnerable
↪  to many types of malicious attacks, such as Sybil attacks, Denial of Service
↪  (DoS), black hole, grey hole and wormhole attacks."}]
```

3.4. Output format

Results should be provided in a TREC-style JSON or TSV format with the following fields:

run_id Run ID starting with (team_id)_(task_id)_(method_used), e.g. UBO_task_2.1_TFIDF

manual Whether the run is manual $\{0, 1\}$.

snt_id a unique passage (sentence) identifier from the input file.

term Term or another phrase to be explained.

term_rank_snt term difficulty rank within the given sentence.

difficulty difficulty scores of the retrieved term on the scale 0-2 (2 to be the most difficult terms, while the meaning of terms scored 0 can be derived or guessed)

definition (only used for Task 2.2) short (one/two sentence) explanations/definitions for the terms. For the abbreviations, the definition would be the extended abbreviation.

Output example Task 2.1:

```
[{"snt_id":"G14.2_2884788726_2",
  "term":"content caching",
  "difficulty":1.0,
  "term_rank_snt":1,
  "run_id":"team1_task_2.1_TFIDF",
  "manual":0}]
```

Output example Task 2.2:

```
[{"snt_id":"G14.2_2884788726_2",
  "term":"content caching",
  "difficulty":1.0,
  "term_rank_snt":1,
  "definition":"Content caching is a performance optimization mechanism in which data
↪  is delivered from the closest servers for optimal application performance.",
  "run_id":"team1_task_2.2_TFIDF_BLOOM",
  "manual":0}]
```

Table 2
SimpleText Task 2: Examples of the annotation

Sentence	**Term**	**Limits**		**Diffi-**
		OK	**Corrected**	**culty**
This device has two work modes: ‘native’ and ‘remote’.	remote	YES		1
This device has two work modes: ‘native’ and ‘remote’.	work modes	YES		0
This device has two work modes: ‘native’ and ‘remote’.	modes native	NO	work modes	0
This device has two work modes: ‘native’ and ‘remote’.	device work	NO	device	0
This device has two work modes: ‘native’ and ‘remote’.	native remote	NO	native	1

4. Evaluation metrics

In this section, we describe different evaluation metrics used to evaluate the performance of submissions for Task 2.1 and Task 2.2.

4.1. Evaluation metrics for Task 2.1

We have evaluated the performance of different submissions for Task 2.1 based on:

- correctness of detected term span (limits): this metric reflects whether the retrieved difficult terms are well limited or not. This is a binary label assigned to each retrieved term.
- difficulty scores: we used a three-scale terms difficulty score which reflects how difficult the term is in the context for an average user and how necessary it is to provide more context about the term: 0 score corresponds to an easy term (explanation might be given but not required); 1 corresponds to somewhat difficult (explanation could help); 2 corresponds to difficult (explanation is necessary). Table 1 contains examples of terms with different difficulty scores.

If an extracted term is considered to be a scientific term, we then assessed its limits, i.e. that it refers to a scientific concept mentioned in the context sentence, and its difficulty. For difficult scientific terms, after correcting the term limits if necessary, we assessed the difficulty of the scientific term. Finally, if any scientific terms have not been extracted, we added them to the list.

Table 2 provides some examples of the annotation for Task 2. *TERM* refers to the terms retrieved by participants, *Correct limits* is a binary category showing whether the retrieved terms is well limited, *Corrected* is an eventual correction of retrieved term limits, *Difficulty* is a term difficulty score in scale 0-2.

4.2. Evaluation metrics for Task 2.2

For this task, we use the following evaluation metrics:

- **BLEU** score [12] between the reference (ground truth definition) and the predicted definitions.

- **ROUGE L F-measure** [13] which measures the ROUGE F-measure based on the Longest Common Subsequence between the reference and the predicted definitions.
- **Semantic match** between the reference and predicted definitions measured using the *all-mpnet-base-v2* [2] sentence transformer model which is an advanced model for sentence similarity. This measure is the average semantic similarity between reference and predicted definitions for all detected terms.
- **Exact match** is specifically applied to the task of abbreviation extension. In this task, participants are required to provide extensions (full forms) for the detected difficult abbreviations. The exact match metric quantifies the number of cases where the reference and predicted extensions for the abbreviations match exactly. It measures the accuracy of the predicted extensions by considering the extent to which they align perfectly with the provided reference extensions.
- **Partial match** evaluates the similarity between reference and predicted abbreviation extensions by considering non-identical matches with a Levenshtein distance lower than 4 characters. It quantifies the number of cases where the extensions exhibit slight variations, such as differences in plural and non-plural forms, while still being considered as acceptable matches. The partial match metric captures the level of similarity between the reference and predicted extensions, allowing for minor discrepancies within a certain threshold.

Table 3 shows a set of examples of terms and their ground truth defintions used for evaluating the submitted runs.

5. Participants' approaches

12 distinct teams submitted 39 runs in total.

National Polytechnic Institute of Mexico (NLPalma) [14] submitted a total of 2 runs for Task 2, a single run for each of Task 2.1 and Task 2.2. They experimented with BLOOMZ to produce description-style prompts given by text input on a task and a binary classifier based on BERT-multilingual for term difficulty.

University of Amsterdam (UAms) [15] submitted a single run for Task 2 focusing on complexity spotting. Their approach aimed to demonstrate the relative effectiveness of simple and straightforward approaches, and made use of standard TF-IDF based term-weighting using the large test set as a source for within-domain term statistics.

University of Cadiz/Split (Smroltra) [16] submitted a total of 20 runs for Task 2, with both 10 runs for Task 2.1 and 10 runs for Task 2.2. They experimented with a range of keyword extraction approaches (KeyBERT, RAKE, YAKE!, BLOOM, T5, TextRank) for the first task, and a Wikipedia extraction approach, BERT, and BLOOMZ for the second task.

[2] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Table 3
Examples of difficult terms and their definitions. Difficult terms are highlighted with the green color

Grade	Sentence	Definition
1	"In the modern era of automation and robotics, *autonomous vehicles* are currently the focus of academic and industrial research."	" Autonomous vehicles (AVs) use technology to partially or entirely replace the human driver in navigating a vehicle from an origin to a destination while avoiding road hazards and responding to traffic conditions."
1	"They can be used for *information retrieval* and information filtering, in which case they evaluate replies and return only the relevant data."	"Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)."
2	"In this paper, we systematically explore this phenomenon. For this, we propose a 3-phase analysis approach, which enables us to identify mining scripts and conduct a large-scale study on the prevalence of *cryptojacking* in the Alexa 1 million websites."	"Cryptojacking is the act of hijacking a computer to mine cryptocurrencies against the user's will, through websites, or while the user is unaware."

University of Guayaquil/Jaén (SINAI) [17] submitted a total of 6 runs for Task 2, with 4 runs for Task 2.1 and 2 runs for Task 2.2. They investigated zero-shot and few-shot learning strategies over the auto-regressive model GPT-3, and in particular effective prompt engineering.

University of Kiel (TeamCAU) [18] submitted 6 runs for Task 2, based on three different large pre-trained language models (SimpleT5, AI21, and BLOOM). They made three and corresponding submissions to both Task 2.1 and 2.2, and also note the complexities of adapting models with limited train data.

University of Kiel/Split/Malta (MicroGerk) [19] submitted a total of 8 runs for Task 2, with 4 runs for Task 2.1 and 4 runs for Task 2.2. They experimented with a range of models (YAKE!, TextRank, BLOOM, GPT-3) for the first task, and a range of models (Wikipedia, SimpleT5, BLOOMZ, GPT-3) for the second task.

University of Southern Maine (Aiirlab) [20] submitted a total of 6 runs for Task 2, consisting 3 runs for Task 2.1 and 3 runs for Task 2.2. They experimented with keyword extraction approaches (YAKE!, KBIR) and IDF weighting for the first task, and definition detection in top-ranked documents based on a trained classifier.

University of Western Brittany (UBO) [21] submitted a total of 8 runs for Task 2, no less than 7 runs for Task 2.1 and a single run for Task 2.2. They experimented with a range of keyword extraction approaches (FirstPhrase, TF-IDF, YAKE!, TextRank, SingleRank, TopicRank, PositionRank) for the first task and a Wikipedia extraction approach for the second task.

University of Split (Croland) submitted a total of 4 runs for Task 2, specifically 2 runs for Task 2.1 and 2 runs for Task 2.2. They applied GPT-3 and TF-IDF for difficult term detection. They extracted definitions from Wikipedia and applied GPT-3 to generate explanations.

University of Liverpool (UOL-SRIS) submitted a single run for Task 2, specifically for Task 2.1 by applying KeyBERT.

University of Kiel/Cadiz/Gdansk (TheLangVerse) submitted a total of 2 runs for Task 2, a single run for both Task 2.1 and Task 2.2 using GPT-3.

6. Results

We evaluate the performance of the submissions separately for the difficult terms spotting (Task 2.1) and definition extraction/generation (Task 2.2) using separate test sets created per task. In this section, we describe the main results of different submissions per task. The performance of the submissions is evaluated separately for two distinct tasks: difficult terms spotting (Task 2.1) and definition extraction/generation (Task 2.2). Each task is evaluated using its respective test set. In this section, we provide an overview of the key results obtained from different submissions for each task. By examining the results of the submissions, we gain insights into the effectiveness of various approaches employed for difficult terms spotting and definition extraction/generation.

6.1. Results of Task 2.1: difficult term spotting

In this section, we focus on the results of the submissions for Task 2.1. A total of 12 teams participated and submitted a combined total of 39 runs for this task. The outcomes of these runs are presented in Table 4, which contains the following metrics: the total number of evaluated terms, the number of terms with correct term limits, the number of correctly attributed scores (regardless of term limits), and the number of correctly limited terms with correctly attributed scores (+Limits). Among all the runs for Task 2.1, the SINAI_task_2.1_PRM_ZS_TASK2_1_V1 run achieved the highest number of correctly detected terms and scores. The performance of different runs on both train and test sets (only on correctly limited terms) is presented in Table 5. Most approaches achieve comparable performance on both train and test sets.

The estimation of difficulty scores for terms proved to be a challenging task, as the majority of the submitted runs struggled to provide accurate scores for more than half of the detected difficult terms. This indicates the difficulty and subjectivity inherent in determining the level of complexity or difficulty associated with specific terms in scientific texts. The variability in assessing the difficulty of terms highlights the need for further research and improvement in developing robust methods for accurately assigning difficulty scores to terms.

Table 4
SimpleText Task 2.1: Results for the official runs

	Total	**Evaluated**		**Score**	
			+Limits		**+Limits**
SINAI_task_2.1_PRM_ZS_TASK2_1_V1	11081	1322	1185	556	507
UAms_Task_2_RareIDF	675090	1293	1145	309	241
SINAI_task_2.1_PRM_FS_TASK2_1_V1	10768	1235	1122	440	405
Smroltra_task_2.1_keyBERT_FKgrade	11099	1215	1061	379	341
Smroltra_task_2.1_keyBERT_F	11099	1215	1061	223	171
UOL-SRIS_2.1_KeyBERT	23757	1215	1061	0	0
MiCroGerk_task_2.1_TextRank	21516	1275	1002	482	391
Smroltra_task_2.1_TextRank_FKgrade	10056	1275	1002	456	363
SINAI_task_2.1_PRM_ZS_TASK2_1_V2	10952	1075	965	366	330
SINAI_task_2.1_PRM_FS_TASK2_1_V2	8836	1004	915	346	316
Smroltra_task_2.1_YAKE_D	11112	1576	905	627	422
MiCroGerk_task_2.1_YAKE	23790	1576	905	582	362
Smroltra_task_2.1_YAKE_Fscore	11112	1576	905	409	209
MiCroGerk_task_2.1_GPT-3	15892	968	889	487	459
UBO_task_2.1_FirstPhrases	14088	1032	831	210	161
UBO_task_2.1_PositionRank	13881	1071	825	237	181
UBO_task_2.1_SingleRank	14088	981	748	200	151
UBO_task_2.1_TfIdf	14340	1206	740	263	187
UBO_task_2.1_TextRank	14088	960	722	189	139
Smroltra_task_2.1_RAKE_AUI	10660	1016	713	378	288
Smroltra_task_2.1_RAKE_F	10660	1016	713	255	170
UBO_task_2.1_TopicRank	13912	824	663	174	144
UBO_task_2.1_YAKE	14337	1118	576	265	116
MiCroGerk_task_2.1_BLOOM	9600	608	535	235	218
Aiirlab_task_2.2_KBIR	4797	498	429	158	135
TeamCAU_task_2.1_ST5	2234	484	418	222	201
Smroltra_task_2.1_SimpleT5	2234	460	406	259	239
Smroltra_task_2.1_SimpleT5_COLEMAN_LIEAU	2234	460	406	168	152
TheLangVerse_task_2.2_openai-curie-finetuned	2234	445	391	0	0
ThePunDetectives_task_2.1_SimpleT5	152072	428	371	110	91
Aiirlab_task_2.2_YAKEIDF	4790	465	241	154	75
Aiirlab_task_2.2_YAKE	4790	486	234	169	78
TeamCAU_task_2.1_AI21	100	10	6	3	2
Smroltra_task_2.1_Bloom	100	4	2	1	1
TeamCAU_task_2.1_BLOOM	100	1	1	0	0

Participants employed a range of approaches, including Large Language Models (LLMs) and unsupervised methods, to tackle the task. However, several runs were limited or incomplete due to token constraints imposed by LLMs or the time required for their execution. It was observed that the results of the same methods varied significantly depending on the specific implementation, fine-tuning techniques, and prompts utilized during the process. In terms of difficult term detection, LLMs demonstrated comparable performance to other methods such

Table 5
SimpleText Task 2.1: Results for the official runs on train and test sets. The evaluation is done based on the terms with thier limits correctly detected.

	Total	**Test**		**Train**	
		Evaluated	**Score**	**Evaluated**	**Score**
SINAI_task_2.1_PRM_ZS_TASK2_1_V1	11081	1185	507	94	56
UAms_Task_2_RareIDF	675090	1145	241	40	21
SINAI_task_2.1_PRM_FS_TASK2_1_V1	10768	1122	405	81	40
Smroltra_task_2.1_keyBERT_FKgrade	11099	1061	341	41	4
Smroltra_task_2.1_keyBERT_F	11099	1061	171	41	7
UOL-SRIS_2.1_KeyBERT	23757	1061	0	42	0
MiCroGerk_task_2.1_TextRank	21516	1002	391	87	61
Smroltra_task_2.1_TextRank_FKgrade	10056	1002	363	87	29
SINAI_task_2.1_PRM_ZS_TASK2_1_V2	10952	965	330	94	53
SINAI_task_2.1_PRM_FS_TASK2_1_V2	8836	915	316	76	41
Smroltra_task_2.1_YAKE_D	11112	905	422	71	21
MiCroGerk_task_2.1_YAKE	23790	905	362	71	51
Smroltra_task_2.1_YAKE_Fscore	11112	905	209	71	32
MiCroGerk_task_2.1_GPT-3	15892	889	459	79	43
UBO_task_2.1_FirstPhrases	14088	831	161	49	19
UBO_task_2.1_PositionRank	13881	825	181	71	29
UBO_task_2.1_SingleRank	14088	748	151	67	19
UBO_task_2.1_TfIdf	14340	740	187	50	13
UBO_task_2.1_TextRank	14088	722	139	67	16
Smroltra_task_2.1_RAKE_AUI	10660	713	288	48	25
Smroltra_task_2.1_RAKE_F	10660	713	170	48	21
UBO_task_2.1_TopicRank	13912	663	144	61	21
UBO_task_2.1_YAKE	14337	576	116	44	11
MiCroGerk_task_2.1_BLOOM	9600	535	218	64	34
Aiirlab_task_2.2_KBIR	4797	429	135	38	11
TeamCAU_task_2.1_ST5	2234	418	201	90	79
Smroltra_task_2.1_SimpleT5	2234	406	239	82	74
Smroltra_task_2.1_SimpleT5_COLEMAN_LIEAU	2234	406	152	82	21
TheLangVerse_task_2.2_openai-curie-finetuned	2234	391	0	165	0
ThePunDetectives_task_2.1_SimpleT5	152072	371	91	118	49
Aiirlab_task_2.2_YAKEIDF	4790	241	75	17	4
Aiirlab_task_2.2_YAKE	4790	234	78	19	3
TeamCAU_task_2.1_AI21	100	6	2	0	0
Smroltra_task_2.1_Bloom	100	2	1	0	0
TeamCAU_task_2.1_BLOOM	100	1	0	0	0

as RareIDF, TextRank, and YAKE!. However, it is worth noting that the term difficulty scores assigned by the models differed considerably from the lay annotations.

Table 6
SimpleText Task 2.2: Results for the official runs on the test set

Run	Evaluated	BLEU	ROUGE	Semantic
UBO_task_2.1_FirstPhrases_Wikipedia	393	29.73	0.41	0.80
Croland_task_2_PKE_Wiki	43	33.68	0.46	0.70
MiCroGerk_task_2.2_GPT-3_Wikipedia	932	26.38	0.41	0.75
Smroltra_task_2.2_Text_Wiki	547	17.59	0.33	0.75
Smroltra_task_2.2_RAKE_Wiki	337	16.95	0.32	0.74
Smroltra_task_2.2_YAKE_Wiki	436	16.94	0.32	0.73
TeamCAU_task_2.1_BLOOM	10	10.46	0.27	0.48
MiCroGerk_task_2.2_GPT-3_BLOOMZ	1,108	9.07	0.40	0.83
Smroltra_task_2.2_keyBERT_Wiki	302	8.60	0.23	0.69
MiCroGerk_task_2.2_GPT-3_GPT-3	1,108	7.73	0.38	0.83
NLPalma_task_2.2_BERT_BLOOMZ	537	7.22	0.39	0.76
Smroltra_task_2.2_Bloomz	23	7.15	0.30	0.69
TeamCAU_task_2.1_AI21	22	6.38	0.31	0.78
TheLangVerse_task_2.2_openai-curie-finetuned	444	5.03	0.25	0.74
Croland_task_2_GPT3	69	4.83	0.27	0.77
SINAI_task_2.1_PRM_FS_TASK2_2_V1	649	4.23	0.21	0.78
MiCroGerk_task_2.2_GPT-3_simpleT5	1,108	4.22	0.28	0.77
TeamCAU_task_2.1_ST5	379	3.33	0.20	0.60
Smroltra_task_2.2_SimpleT5	392	3.09	0.22	0.72
SINAI_task_2.1_PRM_ZS_TASK2_2_V1	649	3.08	0.19	0.69
Smroltra_task_2.2_keyBERT_dict	120	2.07	0.14	0.51
Smroltra_task_2.2_YAKE_WN	48	1.88	0.15	0.44
Aiirlab_task_2.2_KBIR	556	1.62	0.15	0.50
Smroltra_task_2.2_keyBERT_WN	328	1.33	0.14	0.45
Aiirlab_task_2.2_YAKEIDF	179	1.13	0.14	0.41
Aiirlab_task_2.2_YAKE	165	1.10	0.15	0.43
Smroltra_task_2.2_RAKE_WN	70	0.00	0.14	0.46

6.2. Reslts of Task 2.2: difficult term explanation.

For Task 2.2, a total of 10 teams submitted 29 runs. The main results for this task can be found in Table 6. It is important to note that the low number of evaluated sentences for most runs is due to the fact that they were conducted on a smaller subset of sentences from the test set. Nevertheless, the remaining runs demonstrated strong performance in terms of the semantic similarity between their provided definitions and the ground truth definitions. Notably, the runs UBO_task_2.1_-FirstPhrases_Wikipedia, Croland_task_2_PKE_Wiki, and MiCroGerk_task_2.2_GPT-3_-Wikipedia achieved high scores in terms of BLEU metric. This indicates that even though these runs used different sets of words compared to the ground truth definitions, they were able to provide explanations for the difficult terms that were semantically similar to the reference definitions. Moreover, it is worth mentioning that the runs based on

Table 7
SimpleText Task 2.2: Results for the official runs on the train set. Runs with less than 5 evaluated sentences are excluded from this table.

Run	Evaluated	BLEU	ROUGE	Semantic
UBO_task_2.1_FirstPhrases_Wikipedia	14	17.35	0.31	0.73
MiCroGerk_task_2.2_GPT-3_Wikipedia	70	12.6	0.28	0.57
TeamCAU_task_2.1_ST5	90	10.5	0.3	0.67
Smroltra_task_2.2_YAKE_Wiki	52	8.49	0.22	0.62
Smroltra_task_2.2_Text_Wiki	69	7.77	0.22	0.61
Smroltra_task_2.2_keyBERT_dict	6	7.48	0.16	0.37
Smroltra_task_2.2_SimpleT5	82	6.13	0.3	0.71
MiCroGerk_task_2.2_GPT-3_GPT-3	89	5.63	0.31	0.75
NLPalma_task_2.2_BERT_BLOOMZ	22	5.14	0.36	0.74
TheLangVerse_task_2.2_openai-curie-finetuned	165	4.99	0.23	0.67
Smroltra_task_2.2_RAKE_Wiki	36	4.72	0.2	0.58
MiCroGerk_task_2.2_GPT-3_simpleT5	89	4.31	0.28	0.69
SINAI_task_2.1_PRM_FS_TASK2_2_V1	53	4.11	0.23	0.77
Aiirlab_task_2.2_KBIR	38	3.73	0.18	0.5
MiCroGerk_task_2.2_GPT-3_BLOOMZ	89	3.7	0.33	0.73
Smroltra_task_2.2_keyBERT_WN	17	3.06	0.12	0.27
Aiirlab_task_2.2_YAKE	19	2.09	0.19	0.56
Aiirlab_task_2.2_YAKEIDF	17	2.08	0.2	0.56
SINAI_task_2.1_PRM_ZS_TASK2_2_V1	51	2.07	0.2	0.68
Smroltra_task_2.2_keyBERT_Wiki	20	0.0	0.1	0.44

Wikipedia as a resource displayed the highest similarity with the ground truth definitions. This suggests that leveraging Wikipedia as a knowledge source yielded favorable results in terms of aligning the provided definitions with the reference definitions. We observe a similar ranking of the runs on the train set presented in Table 7. An interesting observation is made regarding the scores on the training set being lower than those on the test set. This discrepancy can potentially be explained by the inclusion of a significant number of abbreviations in the training set. The presence of these abbreviations may have introduced additional complexity and challenges during the training process, resulting in lower scores on the training set. To mitigate this issue and ensure a fair evaluation, the test set was designed to consider abbreviations separately. This separate consideration of abbreviations in the test set allows for a more accurate assessment of the models' performance in handling and interpreting these specific elements.

Table 8 presents the performance of the runs on the abbreviation expansion task. The MiCroGerk_task_2.2_GPT-3_BLOOMZ run achieved the highest performance among all the runs for this task. This top-performing model successfully provided 326 identical expansions to the true expansions and 185 partially correct expansions. Overall, LLMs (such as BLOOMz and GPT-3) demonstrated the best performance in terms of abbreviation expansion. It is important to note that the scores provided in the table are averaged over the number of evaluated instances,

Table 8
SimpleText Task 2.2: Results for the official runs on the abbreviation expansion task

Run	Evaluated	BLEU	ROUGE	Semantic	Exact	Partial
MiCroGerk_task_2.2_GPT-3_BLOOMZ	854	13.87	0.68	0.76	326	185
MiCroGerk_task_2.2_GPT-3_GPT-3	855	11.86	0.64	0.73	294	166
MiCroGerk_task_2.2_GPT-3_Wikipedia	855	4.68	0.43	0.60	205	109
MiCroGerk_task_2.2_GPT-3_Wikipedia	618	5.01	0.56	0.64	198	109
NLPalma_task_2.2_BERT_BLOOMZ	345	6.83	0.39	0.52	50	47
Smroltra_task_2.2_SimpleT5	185	0.00	0.12	0.39	8	7
TeamCAU_task_2.1_ST5	141	1.48	0.14	0.40	6	3
TheLangVerse_task_2.2_openai-curie-finetuned	204	1.60	0.14	0.42	1	2
SINAI_task_2.1_PRM_ZS_TASK2_2_V1	228	1.61	0.13	0.55	1	0
TeamCAU_task_2.1_AI21	10	1.87	0.14	0.38	0	0
SINAI_task_2.1_PRM_FS_TASK2_2_V1	228	1.35	0.10	0.53	0	0
UBO_task_2.1_FirstPhrases_Wikipedia	116	5.09	0.19	0.47	0	0
Aiirlab_task_2.2_KBIR	202	1.17	0.07	0.44	0	0
Smroltra_task_2.2_RAKE_Wiki	27	0.54	0.04	0.14	0	0
Smroltra_task_2.2_Bloomz	4	0	0.22	0.61	0	0
Aiirlab_task_2.2_YAKEIDF	19	0	0.10	0.40	0	0
Smroltra_task_2.2_keyBERT_WN	188	0	0.04	0.27	0	0
Smroltra_task_2.2_keyBERT_Wiki	163	0.21	0.02	0.13	0	0
Smroltra_task_2.2_keyBERT_dict	46	0	0.04	0.34	0	0
Smroltra_task_2.2_RAKE_WN	21	0	0.04	0.24	0	0
Smroltra_task_2.2_YAKE_WN	32	0	0.02	0.21	0	0
Smroltra_task_2.2_YAKE_Wiki	31	0	0.03	0.11	0	0
Smroltra_task_2.2_Text_Wiki	50	0	0.02	0.10	0	0
Aiirlab_task_2.2_YAKE	9	0	0.13	0.36	0	0
TeamCAU_task_2.1_BLOOM	3	0	0	0.14	0	0

giving preference to smaller runs. The presence of many partial runs can be attributed to the token or time constraints imposed by LLMs. Furthermore, it should be acknowledged that the evaluation results are dependent on the terms extracted in Task 2.1. The performance of the abbreviation expansion task is influenced by the quality and accuracy of the detected difficult terms in the previous task.

6.3. Analysis of terms and definitions' difficulty.

A master student in translation and technical writing manually assigned difficulty scores on a scale of 1-7 to the syntax and vocabulary of 319 simplified sentences from the participants' runs corresponding to 17 distinct source sentences. Table 9 provides evidence that automatic simplification is effective in terms of reducing syntax difficulty. However, lexical difficulty, i.e. the presence of difficult scientific terms, is much higher, remaining the main barrier to understanding a scientific text. More details can be found in [7].

Table 9
Statistics on the levels of the difficulty of simplified sentences on the scale of 1-7

	1	**2**	**3**	**4**	**5**	**6**	**7**
syntax complexity	259	51	9				
lexical complexity	93	119	62	26	19		

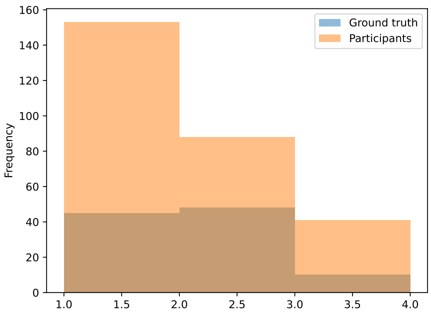

Figure 1: Histogram of the difficulties of the definitions on a scale of 1-3 (1 - easy; 2 - difficult; 3 - very difficult)

To assess the effectiveness of the provided definitions, a master's student in translation and technical writing conducted a manual evaluation. The student assigned difficulty scores on a scale of 1 to 3 (1 being easy, 2 being difficult, and 3 being very difficult) to a total of 744[3] definitions and corresponding terms pooled from the participants' runs and the ground truth. Out of the initial 744 instances, a total of 386 instances were retained for further analysis and evaluation. The decision to exclude the remaining instances was based on the fact that they were deemed incorrect in the given context. The analysis covered 135 unique terms, and the student evaluated the level of difficulty associated with each definition to gauge its helpfulness in understanding the terms.

Most of the difficulty level 3 scientific terms we have identified were abbreviations or highly specific terms within a particular domain. An abbreviation that is not explained within a given excerpt is often incomprehensible and impossible to guess for a general reader. The other difficulty level 3 terms we have identified are also highly specific to a domain and cannot be understood unless one is specialized in that field. They are also terms that are difficult to define in a simple manner for the most part.

Figure 1 presents the distribution of easy, difficult, and very difficult definitions in both the participants' runs and the ground truth. The figure offers compelling evidence that in almost half of the cases for both the runs and the ground truth, the definitions are perceived as easy by a non-expert in computer science. Notably, in our ground truth, there is a higher proportion of difficult definitions and a lower proportion of very difficult definitions compared to the participants' runs. This discrepancy suggests that the ground truth definitions tend to be

[3] Note that new instances were treated compared to the results reported in [8]

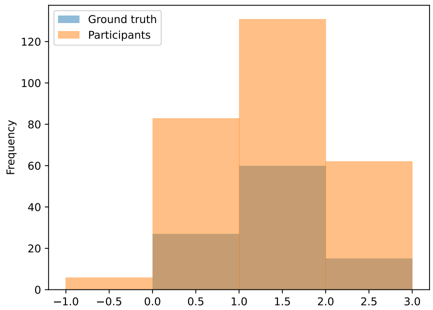

Figure 2: Difference between term difficulty and definition difficulty on a scale of 1-3 (1 - easy; 2 - difficult; 3 - very difficult). Positive values on X axis show helpful definitions. 0 refers to unhelpful definitions. Negative values increase the difficulty.

relatively more challenging, with fewer easy definitions, as compared to the definitions provided by the participants. It highlights the importance of considering the difficulty level of definitions and striving for a balance that caters to a diverse audience, including non-experts in the field.

Although the majority of definitions are considered to be easy, this evidence is not enough to make a conclusion about their helpfulness. Therefore, we decided to compare the term difficulty and the corresponding definitions' difficulty. Figure 2 illustrates the histogram of the differences between the difficulty of terms and the difficulty of their corresponding definitions. The X-axis represents the values of these differences. Positive values indicate helpful definitions, where the term difficulty is higher than the difficulty of the corresponding definition. A value of 0 represents an unhelpful definition, as it shares the same difficulty as the term it aims to explain. Negative values on the X-axis indicate increased difficulty, meaning that the definition is more challenging than the corresponding term.

The results depicted in Figure 2 reveal that approximately 30%-40% of definitions can be classified as either unhelpful or even more difficult than the terms they are meant to clarify. This suggests that a significant portion of the definitions may not effectively assist readers in understanding the associated terms. The ground truth definitions, in contrast to the participants' runs, do not exhibit such harmful patterns, implying that they maintain a higher level of coherence and clarity.

An interesting observation made during the evaluation is that for highly complex terms, the corresponding definitions often turned out to be equally difficult to understand. Explaining a complicated term may require introducing additional complex terms, leading to longer and more intricate definitions. This complexity stems from the necessity of explaining not just one term, but multiple related terms in order to provide a comprehensive understanding. Additionally, it was noticed that certain terms like "spins" or "chips" had generated definitions that were incorrect in the context of the given text. Although these terms are commonly used in everyday language, they hold distinct meanings in specific domains such as quantum physics or computer science. Assigning the correct definitions to these terms within their relevant scientific or technical contexts is of paramount importance to ensure accurate comprehension and avoid potential misunderstandings.

7. Conclusion and future work

For Task 2 focused on difficult concept identification and explanation, we created a corpus of sentences extracted from the abstracts of scientific publications, with manual annotations of scientific term difficulty and their definitions. 12 distinct teams participated in the task and submitted 39 runs demonstrating a diverse range of approaches in addressing the task's objectives and challenges varying from traditional statistic methods to LLMs.

Several noteworthy observations emerge from the study. Firstly, it is evident that even when employing similar models, the results achieved by the same methods can vary significantly. This variance can be attributed to factors such as the specific implementation approach, the fine-tuning techniques employed, and the choice of prompts utilized during the simplification process. These implementation-related aspects play a crucial role in determining the effectiveness and outcomes of the methods applied. Hence, careful consideration and optimization of these factors are essential to achieving desirable results in text simplification tasks.

Another significant observation is that efficiency plays a vital role alongside effectiveness in text simplification tasks, including difficulty spotting and explanation. Many partial runs were received due to the limitations of tokens or time constraints imposed by LLMs. The results of difficult term detection achieved by LLMs were found to be comparable to those obtained through unsupervised methods both for difficult term spotting and providing definitions (e.g. definitions retrieved from Wikipedia). This observation highlights the trade-off between efficiency and effectiveness in text simplification, where the choice of approach must consider both aspects to strike the right balance.

The third observation highlights the ongoing challenge of achieving robustness in approaches. Specifically, it is noted that a significant percentage, ranging from 30% to 40%, of the generated definitions are either unhelpful or even more difficult than the corresponding terms themselves. The complexity of highly complex terms often translates into equally intricate definitions, as it may be necessary to introduce additional complex terms to provide a comprehensive explanation. This can result in longer and more convoluted definitions, which may still pose difficulties for readers. This indicates that there is room for improvement in ensuring that the generated definitions effectively simplify and enhance the understandability of the terms for the target audience. Overcoming this challenge requires developing more robust and accurate approaches that can consistently provide helpful and simplified explanations for difficult terms.

Furthermore, the issue of incorrect definitions for terms like "spins" or "chips" highlights the importance of context and domain-specific knowledge. Although these terms have everyday usage, their meanings can differ significantly in scientific or technical domains such as quantum physics or computer science. A significant portion of the provided definitions, specifically nearly half of them, were deemed incorrect. Providing accurate definitions that align with the specific context is crucial for enabling accurate comprehension and avoiding potential misunderstandings.

These observations underscore the need for precise and contextually appropriate definitions that strike a balance between simplicity and accuracy. Addressing these challenges will require developing advanced techniques that can accurately capture the nuances of complex terms while providing simplified and accurate explanations tailored to the target audience and context.

So the general upshot of the CLEF 2023 SimpleText track is both that we observed great progress, but at the same time that there is also still a lot of room for improvement.

The simplification techniques are successful in simplifying the structure and syntax of the text, making it more accessible and easier to comprehend. However, despite the improvements in syntax, the challenge of lexical difficulty remains a significant barrier to understanding scientific texts.

In future work, there are plans to focus on classifying difficult term explanations, including definitions, examples, and abbreviation deciphering. We consider conducting the systems' evaluation based on the usefulness and complexity of the explanations they provide for scientific terms.

Further details about the lab can be found at the SimpleText website: http://simpletext-project.com. Please join us and help to make scientific results understandable!

Acknowledgments

This research was funded, in whole or in part, by the French National Research Agency (ANR) under the project ANR-22-CE23-0019-01. We would like to thank Jaap Kamps, Radia Hannachi, Silvia Araújo, Pierre De Loor, Olga Popova, Diana Nurbakova, Quentin Dubreuil, and all the other colleagues and participants who helped run this track.

References

[1] M. Maddela, W. Xu, A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification, in: Proc. of EMNLP 2018, ACL, Brussels, Belgium, 2018, pp. 3749–3760. URL: https://www.aclweb.org/anthology/D18-1410.

[2] T. O'Reilly, Z. Wang, J. Sabatini, How Much Knowledge Is Too Little? When a Lack of Knowledge Becomes a Barrier to Comprehension:, Psychological Science (2019). URL: https://journals.sagepub.com/doi/10.1177/0956797619862276.

[3] M. Maddela, F. Alva-Manchego, W. Xu, Controllable Text Simplification with Explicit Paraphrasing (2021). URL: http://arxiv.org/abs/2010.11004.

[4] 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: CLEF'21: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, 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.

[5] 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: CLEF'22: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, 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.

[6] E. SanJuan, S. Huet, J. Kamps, L. Ermakova, Overview of the CLEF 2023 SimpleText Task 1: Passage selection for a simplified summary, in: [22], 2023.

[7] L. Ermakova, S. Bertin, H. McCombie, J. Kamps, Overview of the CLEF 2023 SimpleText Task 3: Scientific text simplification, in: [22], 2023.

[8] 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.), CLEF'23: Proceedings of the Fourrteenth International Conference of the CLEF Association, Lecture Notes in Computer Science, Springer, 2023.

[9] E. SanJuan, S. Huet, J. Kamps, L. Ermakova, Overview of the CLEF 2022 SimpleText Task 1: Passage selection for a simplified summary, volume 3180 of *CEUR Workshop Proceedings*, 2022. URL: https://ceur-ws.org/Vol-3180/.

[10] 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.), CLEF'22: Proceedings of the Thirteenth International Conference of the CLEF Association, Lecture Notes in Computer Science, Springer, 2022.

[11] A. S. Schwartz, M. A. Hearst, A simple algorithm for identifying abbreviation definitions in biomedical text, in: Biocomputing 2003, 2002, pp. 451–462.

[12] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proc. of the 40th annual meeting on ACL, ACL, 2002, pp. 311–318.

[13] C.-Y. Lin, E. Hovy, Automatic evaluation of summaries using n-gram co-occurrence statistics, in: Proc. of the 2003 Conference of the North American Chapter of the ACL on Human Language Technology-Volume 1, ACL, 2003, pp. 71–78.

[14] C. P. P. Victor Manuel Palma, G. Sidorov, NLPalma @ CLEF 2023 SimpleText: BLOOMZ and BERT for complexity and simplification task, in: [22], 2023.

[15] M. Adib, R. Hutter, J. Sutmullera, D. Rau, J. Kamps, University of Amsterdam at the CLEF 2023 SimpleText Track, in: [22], 2023.

[16] O. Popova, P. Dadić, CLEF 2023 SimpletText Tasks 2 and 3: Enhancing Language Comprehension: Addressing Difficult Concepts and Simplifying Scientific Texts Using GPT, BLOOM, KeyBert, Simple T5 and More, in: [22], 2023.

[17] J. A. Ortiz-Zambrano, C. Espin-Riofrio, A. Montejo-Ráez, SINAI participation in SimpleText Task 2 at CLEF 2023: GPT-3 in Lexical Complexity Prediction for general audience, in: [22], 2023.

[18] A. Anjum, N. Lieberum, Automatic Simplification of Scientific Texts using Pre-trained Language Models: A Comparative Study at CLEF Symposium 2023, in: [22], 2023.

[19] D. R. Davari, A. Prnjak, K. Schmitt, CLEF2023 SimpleText Task 2, 3: Identification and Simplification of Difficult Terms, in: [22], 2023.

[20] B. Mansouri, S. Durgin, S. Franklin, S. Fletcher, R. Campos, AIIR and LIAAD Labs Systems for CLEF 2023 SimpleText, in: [22], 2023.

[21] Q. Dubreuil, UBO Team @ CLEF SimpleText 2023 Track for Task 2 and 3 - Using IA models to simplify Scientific Texts, in: [22], 2023.

[22] Proceedings of the Working Notes of CLEF 2023: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2023.