Overview of the CLEF 2022 SimpleText Lab: Automatic Simplification of Scientific Texts

Ermakova, L.; SanJuan, E.; Kamps, J.; Huet, S.; Ovchinnikova, I.; Nurbakova, D.; Araújo, S.; Hannachi, R.; Mathurin, E.; Bellot, P.

DOI
10.1007/978-3-031-13643-6_28

Publication date
2022

Document Version
Final published version

Published in
Experimental IR Meets Multilinguality, Multimodality, and Interaction

Overview of the CLEF 2022 SimpleText Lab: Automatic Simplification of Scientific Texts

Liana Ermakova[1(✉)], Eric SanJuan[2], Jaap Kamps[3], Stéphane Huet[2], Irina Ovchinnikova[4], Diana Nurbakova[5], Sílvia Araújo[6], Radia Hannachi[7], Elise Mathurin[1], and Patrice Bellot[8]

[1] Université de Bretagne Occidentale, HCTI, Brest, France
liana.ermakova@univ-brest.fr
[2] Avignon Université, LIA, Avignon, France
[3] University of Amsterdam, Amsterdam, The Netherlands
[4] ManPower Language Solution, Tel Aviv, Israel
[5] University of Lyon, INSA Lyon, CNRS, LIRIS, Lyon, France
[6] University of Minho, Braga, Portugal
[7] Université de Bretagne Sud, HCTI, Morbihan, France
[8] Aix Marseille Univ, Université de Toulon, CNRS, LIS, Toulon, France

Abstract. Although citizens agree on the importance of objective scientific information, yet they tend to avoid scientific literature due to access restrictions, its complex language or their lack of prior background knowledge. Instead, they rely on shallow information on the web or social media often published for commercial or political incentives rather than the correctness and informational value. This paper presents an overview of the CLEF 2022 SimpleText track addressing the challenges of text simplification approaches in the context of promoting scientific information access, by providing appropriate data and benchmarks, and creating a community of IR and NLP researchers working together to resolve one of the greatest challenges of today. The track provides a corpus of scientific literature abstracts and popular science requests. It features three tasks. First, *content selection* (what is in, or out?) challenges systems to select passages to include in a simplified summary in response to a query. Second, *complexity spotting* (what is unclear?) given a passage and a query, aims to rank terms/concepts that are required to be explained for understanding this passage (definitions, context, applications). Third, *text simplification* (rewrite this!) given a query, asks to simplify passages from scientific abstracts while preserving the main content.

Keywords: Scientific text simplification · (Multi-document) summarization · Contextualization · Background knowledge · Scientific information distortion

1 Introduction

Scientific literacy is an important ability for people. It is one of the keys for critical thinking, objective decision-making and judgment of the validity and significance of findings and arguments, which allows discerning facts from fiction. Thus, having a basic scientific knowledge may also help maintain one's health, both physiological and

A. Barrón-Cedeño et al. (Eds.): CLEF 2022, LNCS 13390, pp. 470–494, 2022.
https://doi.org/10.1007/978-3-031-13643-6_28

mental. The COVID-19 pandemic provides a good example of such a matter. Understanding the issue itself, choosing to use or avoid particular treatment or prevention procedures can become crucial. However, the recent pandemic has also shown that simplification can be modulated by political needs and the scientific information can be distorted [14]. Thus, the evaluation of the alteration of scientific information during the simplification process is crucial but underrepresented in the state-of-the-art.

Digitization and open access have made scientific literature available to every citizen. While this is an important first step, there are several remaining barriers preventing laypersons to access the objective scientific knowledge in the literature. In particular, scientific texts are often hard to understand as they require solid background knowledge and use tricky terminology. Although there were some recent efforts on text simplification (e.g. [23]), removing such understanding barriers between scientific texts and general public in an automatic manner is still an open challenge. The CLEF 2022 SimpleText track[1] brings together researchers and practitioners working on the generation of simplified summaries of scientific texts. It is a new evaluation lab that follows up the SimpleText-2021 Workshop [11]. All perspectives on automatic science popularisation are welcome, including but not limited to: Natural Language Processing (NLP), Information Retrieval (IR), Linguistics, Scientific Journalism, etc.

SimpleText provides data and benchmarks for discussion of challenges of automatic text simplification by bringing in the following interconnected tasks:

Task 1: What is in (or out)? Select passages to include in a simplified summary, given a query.

Task 2: What is unclear? Given a passage and a query, rank terms/concepts that are required to be explained for understanding this passage (definitions, context, applications, ...).

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

Automatic scientific text simplification is a very ambitious problem which cannot be addressed by a simple solution, but we have isolated three clear challenges that need to be addressed to improve non-expert access to scientific literature. In order to simplify scientific texts, one has to (1) select the information to be included in a simplified summary, (2) decide whether the selected information is sufficient and comprehensible or provide some background knowledge if not, (3) improve the readability of the text [10]. Our lab is organised around this pipeline. Our test data was built accordingly as we asked to rank difficult terms (Task 2) and simplify sentences (Task 3) retrieved for Task 1 and we evaluated the results with regard to the queries from Task 1.

In the CLEF 2022 edition of SimpleText, a total of 62 teams registered for the SimpleText track. A total of 40 users downloaded data from the server. A total of 9 distinct teams submitted 24 runs, of which 10 runs were updated. The details of statistics on runs submitted for shared tasks are presented in Table 1.

This introduction is followed by Section 2 presenting a brief overview of related evaluation initiatives, related tasks and related approaches. The bulk of this paper presents the tasks with the datasets and evaluation metrics used, as well as the results of the participants, in three self-contained sections: Section 3 on the first task about content

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

Table 1. CLEF 2022 SimpleText official run submission statistic

Team	Task 1	Task 2	Task 3	Total runs
UAms	2	1		3
NLP@IISERB	3 (3 updated)			3
SimpleScientificText		1 (1 updated)		1
aaac		1 (1 updated)		1
LEA_T5		1	1	2
PortLinguE			1 (1 updated)	1
CYUT Team2	1		1	2
HULAT-UC3M			10 (4 updated)	10
CLARA-HD			1	1
Total runs	6	4	14	24

selection, Section 4 on the second task about complexity spotting, and Section 5 on the third task about text simplification proper. We end with Section 6 discussing the results and findings, and lessons for the future.

2 Related Work

This section presents a brief overview of related evaluation initiatives, related tasks and related approaches.

In parallel with the CLEF SimpleText track, which was accepted in 2020, there have been a range of related initiatives on scholarly document processing at NLP conference. In 2020, Scholarly Document Processing[2] provided the shared tasks on

- CL-SciSumm: Scientific Document Summarization;
- CL-LaySumm:Lay Summary;
- LongSumm: Generating Long Summaries for Scientific Documents.

CL-SciSumm and LongSumm are focused on summarization task but no adaptation to general public is previewed. The CL-SciSumm'20 LaySummary [6] subtask asked to produce a scientific paper summary without technical jargon. However, terms are not often replaceable due to the risk of information distortion and these complex concepts should be explained to a reader.

In 2022 the Third Workshop on Scholarly Document Processing[3] hosted the following shared tasks which are related to our track although they don't tackle the simplification aspect:

- MSLR22: Multi-Document Summarization for Literature Reviews;
- DAGPap22: Detecting automatically generated scientific papers;
- LongSumm 2022: Generating Long Summaries for Scientific Documents;

[2] https://ornlcda.github.io/SDProc/sharedtasks.html.
[3] https://sdproc.org/2022/sharedtasks.html.

– SV-Ident 2022: Survey Variable Identification in Social Science Publications;
– Scholarly Knowledge Graph Generation;
– Multi Perspective Scientific Document Summarization.

As it turns out, the SimpleText tasks and SDProc tasks are complementary, and together build a larger community to work on this important problem.

Popular science articles are generally much shorter than scientific publications. Thus, summarization is a step to text simplification as it reduces the amount of information to be processed. However, information selection is understudied task in document simplification [41] as existing works mainly focus on word/phrase-level [24] or sentence-level simplifications [9]. However, the lack of background knowledge can become a barrier to reading comprehension and there is a knowledge threshold allowing reading comprehension [30]. Scientific text simplification presupposes the facilitation of readers' understanding of complex content by establishing links to basic lexicon while traditional methods of text simplification try to eliminate complex concepts and constructions [24]. SimpleText is not limited to a "Split and Rephrase" task [26] but also aims to provide a sufficient context to a scientific text. Entity linking could mitigate the background knowledge problem, by providing definitions, illustrations, examples, and related entities, but the existing entity linking datasets are focused on people, places, and organisation [19], while a non-expert reader of a scientific article needs assistance with new concepts and methods. INEX/CLEF'11–14 Tweet Contextualization [4] and CLEF'16–17 Cultural Microblog Contextualization [13] tracks aim to provide lacking background knowledge to a tweet. Besides completely different nature of tweets and popular science, this use case differs from the text simplification as this lack of background knowledge is due to the tweet length. In contrast to the Background Linking task at TREC'20 News Track [3], SimpleText focuses on (1) scientific text; (2) selection of notions to be explained; (3) helpfulness of the provided information rather than its relevance.

Large pre-trained AI models, like Jurassic-1 [20], Google's T5 [38], BERT or GPT-3 [5], outperformed other state-of-the-art models on several NLP tasks, including automatic summarization and text simplification [40], but their serious issues are (1) consistency and coherency (coreference errors) [35] and (2) limitation to short texts (<2k tokens) [39]. Simple Wikipedia based datasets could be useful to train AI models but (1) they are not scientific publications; (2) there is no direct correspondence between Wikipedia and Simple Wikipedia articles [14]. Another dataset was introduced at TAC 2014 Biomedical Summarization Track [1] with a goal to retrieve important aspects of a paper from the perspective of the community.

Automatic evaluation metrics have been designed to measure the results of text simplification: SARI [37] targets lexical complexity, while SAMSA estimates sentence structural complexity [32]. Standard evaluation measures (e.g. BLEU, ROUGE) are difficult to apply as one should consider the end user as well as source document content. Since traditional readability indices can be misleading [36], researchers proposed various approaches based on expert judgement [8], readability level [17], relevance judgement [7], crowd-sourcing [2], eye-tracking [18].

In contrast to that, we evaluate simplification in terms of lexical and syntax complexity combining with error analysis. As we demonstrated previously, scientific

information is often distorted accidentally due to misunderstanding of terminology, omission of essential details, insertion of erroneous background etc. [14]. Information distortion analysis is close to scientific claim verification [25,34] but fact checking is limited to search for relevant evidence and decide whether it supports the claim. Another close work is [31], where the TF-IDF cosine similarity between documents is computed on (1) a collection of abstracts of scientific papers from the Citation Network Dataset V1 AMINER [33] and (2) a set of articles from Huffington Post. However, this approach is not robust to lexical changes, which are crucial for text simplification. To the best of our knowledge, no other automatic nor semi-automatic method for information distortion analysis exists.

3 Task 1: What Is in (or Out)?

In this section, we discuss the first task about content selection (and *avoiding* complexity) from a corpus of scientific abstracts, addressing the task:

Select passages to include in a simplified summary, given a query.

The task aims at finding references in computer science that could be inserted as citations in original press articles of general audience for illustration, fact checking or actualization. For each of the selected references, more relevant sentences need to be extracted. These passages can be complex and require further simplification to be carried out in Tasks 2 and 3. Task 1 focuses on content retrieval.

3.1 Evaluation Framework

Corpus. As in 2021, we use the Citation Network Dataset: DBLP+Citation, ACM Citation network (12th version) [33] as source of scientific documents that can be used as reference passages [10]. It contains:

- 4,894,083 bibliographic references published before 2020;
- 4,232,520 abstracts in English;
- 3,058,315 authors with their affiliations;
- 45,565,790 ACM citations.

Textual content together with authorship can be extracted from this corpus. Although we manually preselected abstracts for topics, participants also have access to an Elastic Search index; this index is adequate to passage retrieval using BM25.

Additional datasets have been extracted to generate Latent Dirichlet Allocation models for query expansion or train Graph Neural Networks for citation recommendation as carried out in StellarGraph[4] for example. The shared datasets provide: document abstract content for LDA (Latent Dirichlet Allocation) or Word Embedding (WE); document authors for coauthoring analysis; citation relationship between documents for co-citation analysis; citations by author for author impact factor analysis. These extra datasets are intended to be used to select passages by authors who are experts on the topic (highly cited by the community).

[4] https://stellargraph.readthedocs.io/.

Table 2. SimpleText Task 1: Examples of topics and queries

Topic ID	Query ID	Title or Query
G12		*Patient data from GP surgeries sold to US companies*
	G12.1	`patient data`
G13		*Baffled by digital marketing? Find your way out of the maze*
	G13.1	`digital marketing`
	G13.2	`advertising`

Topics. Topics are a selection of 40 press articles: 20 from *The Guardian*,[5] a major international newspaper for a general audience with a tech section, and 20 from *Tech Xplore*[6] a website taking part in the Science X Network to provide a comprehensive coverage of engineering and technology advances. Each article was selected in the computer science field to be in accordance with the provided corpus. URLs to original articles, the title and textual content of each topic were provided to participants. Articles were enriched with queries manually extracted from their content to provide an indication of the essential technical concepts covered. We manually checked that each query allows participants to retrieve from the corpus at least 5 relevant passages that could be inserted as citations in the press article. The use of these queries were optional. Examples of topics and queries are given in Table 2.

Output Formats. Results had to be provided in a TREC style tabulated format (with a ".csv" extension). The following columns were required (including the first line):

run_id Run ID starting with team ID, followed by "task1" and run name
manual Whether the run is manual {0,1}
topic_id Topic ID
query_id Query ID used to retrieve the document (if one of the queries provided for the topic was used; 0 otherwise)
doc_id ID of the retrieved document (to be extracted from the JSON output)
passage Text of the selected passage (abstract)

For each topic, the maximum number of distinct DBLP references (_id json field) was 100 and the total length of passages was not to exceed 1,000 tokens. Table 3 shows an example of Task 1 output.

Evaluation Metrics. All passages retrieved from DBLP by participants are expected to have some overlap (lexical or semantic) with the article content. Passage relevance were evaluated through:

1. Lexical and semantic overlap of extracted passages with topic article content,
2. Manual assessment of a pool of passages.

[5] https://www.theguardian.com/science.
[6] https://techxplore.com/.

Table 3. SimpleText Task 1: Examples of output

Run	M/A	Topic	Query	Doc	Passage
ST1_task1_1	0	G01	G01.1	1564531496	A CDA is a mobile user device, similar to a Personal Digital Assistant (PDA). It supports the citizen when dealing with public authorities and proves his rights - if desired, even without revealing his identity.
ST1_task1_1	0	G01	G01.1	3000234933	People are becoming increasingly comfortable using Digital Assistants (DAs) to interact with services or connected objects
ST1_task1_1	0	G01	G01.2	1448624402	As extensive experimental research has shown individuals suffer from diverse biases in decision-making.

To build a pooled test collection, we first extracted all the article IDs ranked by the number of participants who used the article to select passages. From this extraction, we only kept articles chosen by at least two participants and gave a relevance score on a scale of 0 to 5:

0 for irrelevant articles;
1 for marginally relevant articles;
2 when the abstract is relevant with the query;
3 when the abstract and keywords are relevant with the query;
4 when the abstract and keywords are relevant with the query and the topic (title of the original article);
5 when the abstract and keywords are relevant with the query and the extended topic (content of the original article).

In order to speed up the judgment process, for this edition we only evaluated relevance at the article level, and not at the sentence level. The abstract was considered as relevant as soon it has a sentence useful to explain the title or the original article.

Among documents returned by at least three runs we found out:

– 14 Guardian topics with lightly relevant documents;
– 11 Guardian topics with highly relevant documents;
– 10 Tech topics with lightly relevant documents; and
– 9 Tech topics with highly relevant documents.

For the documents returned by two runs, we had a high number of 1 and 2 scores for the Guardian topics. As regards the Tech Xplore topics, which have more technical queries since they deal with more technical and specific areas, queries were less ambivalent and more in keeping with the content of DBLP corpus. This has resulted in usually higher relevant scores, with many articles retrieved by two participants having a score of 3. Globally, whether the query comes from the Guardian or Tech Xplore, human evaluators found abstracts, among the articles retrieved by the participants from DBLP, that really explain the article or have matters which should have been addressed in the

Table 4. SimpleText Task 1: Evaluation scores of official runs

Team	Score	#Docs	Doc Avg	#Queries	Query Avg	NDCG
CYUT	125	44	0.53	77	1.62	0.3322
UAMS-MF*	163	54	0.87	99	1.65	0.2761
UAMS	52	17	0.22	40	1.30	0.1048
NLP@IISERB	26	7	0.35	13	2.00	0.0290

* *Manual run.*

original article. Passages were often issued from publications that are more related to cognitive or information sciences than to technical fields, which shows that the DBLP corpus has expanded beyond computer science.

3.2 Results

A total of 3 teams submitted 6 runs: 4 automatic runs extracted 100 documents or abstracts per subquery, the CYUT automatic run extracted 5 sentences per subquery, and the manual extracted passages for a selection of subqueries.

We consider here the reduced pool of documents returned by at least two runs; there are 72 topics with judgments, with a mean of 6.7 and a median of 4 judged documents per topic. Since we have participants that focused on a short list of documents, we only report results computed at a depth of 5 returned documents. Table 4 shows cumulative (0–5) scores obtained by each run (*Score*), the number of returned documents with a score ≥ 1 (*#Docs*), the number of queries with at least one returned document (*#Queries*) and the average scores per document and query. We also provide NDCG@5 as the metrics used for official ranking on this task. These values show that the automatic run made by CYUT and the manual run significantly outperform other automatic runs in terms of selecting the abstracts with a high relevance.

4 Task 2: What Is Unclear?

In this section, we discuss the second task about complexity spotting in an extracted sentence from a scientific abstract, addressing the task:

> *Given a passage and a query, rank terms/concepts that are required to be explained for understanding this passage (definitions, context, applications etc.).*

The goal of this task is to decide which terms (up to 5) require explanation and contextualization to help a reader to understand a complex scientific text—for example, with regard to a query, terms that need to be contextualized (with a definition, example and/or use-case). For each passage, participants should provide a ranked list of difficult terms with corresponding scores on the scale 1–3 (3 to be the most difficult terms, while the meaning of terms scored 1 can be derived or guessed) and on the scale 1–5 (5 to be the most difficult terms). Passages (sentences) are considered to be independent, i.e. difficult term repetition was allowed.

4.1 Evaluation Framework

Train Dataset. For this task, data is two-fold: *Medicine* and *Computer Science*, as these two domains are the most popular on forums like ELI5 [12,29]. As in 2021, for *Computer Science*, we use scientific abstracts from the Citation Network Dataset: DBLP+Citation, ACM Citation network (12th version)[7] [10]. A master student in Technical Writing and Translation manually annotated each sentence by extracting difficult terms and attributing difficulty scores on a scale of 1–3 (3 to be the most difficult terms, while the meaning of terms scored 1 can be derived or guessed) and on a scale of 1–5 (5 to be the most difficult terms).

In 2022, we introduced new data based on Google Scholar and PubMed articles on muscle hypertrophy and health annotated by a master student in Technical Writing and Translation, specializing in these domains. The selected abstracts included the objectives of the study, the results and sometimes the methodology. The abstracts including only the topic of the study were excluded because of the lack of information. To avoid the curse of knowledge, another master student in Technical Writing and Translation not familiar with the domain was solicited for complexity spotting.

We provided 453 annotated examples in total.

Test Dataset. To construct the test data, we retrieved 116,763 sentences from the DBLP abstracts according to the queries from Task 1. We then manually evaluated 592 distinct sentences for 11 queries. For the query *Digital assistant* we took the first 1,000 sentences retrieved by ElasticSearch. We pool terms submitted by all participants for all these queries, representing a number of 4,167 distinct pairs *sentence-term* in total. We ensured that for each evaluated source sentence the pool contained the results of all participants. Statistics of the number of evaluated sentences per query for Task 2 are given in Table 5.

Input and Output Formats. The input for the train and the test data was provided in JSON and CSV formats with the following fields:

snt_id a unique passage (sentence) identifier.
source_snt passage text.
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.

Input example (JSON format):

```
{"snt_id":"G06.2_2548923997_3", "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.", "doc_id":2548923997, "query_id":"G06.2",
 ↪ "query_text":"self driving"}
```

[7] https://www.aminer.org/citation.

Table 5. SimpleText Task 2: Statistics of the number of evaluated sentences per query

Query		# Sentences	# Sentence-term pairs
1	*guessing attack*	60	389
2	*end to end encryption*	55	390
3	*imbalanced data*	55	381
4	*distributed attack*	54	385
5	*genetic algorithm*	51	374
6	*quantum computing*	51	385
7	*qbit*	50	363
8	*side-channel attack*	49	340
9	*traffic optimization*	47	344
10	*quantum applications*	42	320
11	*cyber-security*	35	244
12	*conspiracy theories*	23	180
13	*crowsourcing*	15	104
14	*digital assistant*	5	32

Participants had to submit a list of terms to be contextualized in a JSON format or a tabulated file TSV (for manual runs) with the following fields:

run_id Run ID starting with (team_id)_(task_id)_(*name*).
manual Whether the run is manual $\{0, 1\}$.
snt_id a unique passage (sentence) identifier from the input file.
term Term or other phrase to be explained.
term_rank_snt term difficulty rank within the given sentence.
score_5 term difficulty score on the scale from 1 to 5 (5 to be the most difficult terms).
score_3 term difficulty score on the scale from 1 to 3 (3 to be the most difficult terms).

Output example (JSON format):

```
{"run_id":"NP_task_2_run1", "manual":1,
 ↪  "snt_id":"G06.2_2548923997_3", "term":"black hole attack",
 ↪  "term_rank_snt":1, "score_5":5, "score_3":3},
{"run_id":"NP_task_2_run1", "manual":1,
 ↪  "snt_id":"G06.2_2548923997_3", "term":"grey hole attack",
 ↪  "term_rank_snt":2, "score_5":5, "score_3":3},
{"run_id":"NP_task_2_run1", "manual":1,
 ↪  "snt_id":"G06.2_2548923997_3", "term":"Sybil attack",
 ↪  "term_rank_snt":3, "score_5":5, "score_3":3},
{"run_id":"NP_task_2_run1", "manual":1,
 ↪  "snt_id":"G06.2_2548923997_3", "term":"wormhole attack",
 ↪  "term_rank_snt":4, "score_5":5,"score_3":3},
{"run_id":"NP_task_2_run1", "manual":1,
 ↪  "snt_id":"G06.2_2548923997_3", "term":"Denial of service
 ↪  attack", "term_rank_snt":5, "score_5":4, "score_3":3}
```

Table 6. Examples of the term difficulty scale used for evaluation. Difficult terms are highlighted with the green color

Grade	Non-abbreviated (ordinary) term	Abbreviation
7	*external qubit* in "The qubit—qutrit pair acts as a closed system and one *external qubit* serve as the environment for the pair."	*XCSFHP* in "We compared *XCSFHP* to XCSF on several problems."
6	"This paper bring forward based on immune genetic algorithm to solve *man on board automated storage and retrieval system* optimized problem, immune genetic algorithm remains the characteristic which is not ..." "*Tile coding* is a well-known function approximator that has been successfully applied to many reinforcement learning tasks."	"*XCS* with computed prediction, namely XCSF, extends XCS by replacing the classifier prediction with a parametrized prediction function." "Side-channel attack (*SCA*) is a very efficient cryptanalysis technology to attack cryptographic devices."
5	"Experiment simulation result express: the result of *immune genetic algorithm* is better than traditional genetic algorithm in the circumstance of the same clusters and the same evolution generation."	"This paper presents a simple real-coded estimation of distribution algorithm (EDA) design using x-ary extended compact genetic algorithm (*XECGA*) and discretization methods."
4	"Immune genetic algorithm can shorten storage or retrieval distance in application, and enhance storage or *retrieval efficiency*." "*Deep learning* has become increasingly popular in both academic and industrial areas in the past years."	"This paper presents a simple real-coded estimation of distribution algorithm (*EDA*) design using x-ary extended compact genetic algorithm (XECGA) and discretization methods."
3	"The XECGA is then used to build the probabilistic model and to sample a new population based on the *probabilistic model*."	"We evaluate each measure's performance by *AUC* which is usually used for evaluation of imbalanced data classification."
2	"Experiment simulation result express: the result of immune genetic algorithm is better than traditional genetic algorithm in the circumstance of the same *clusters* and the same evolution generation." "Specifically, the real-valued *decision variables* are mapped to discrete symbols of user-specified cardinality using discretization methods."	*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."
1	"video labeling game is a *crowsourcing* tool to collect user-generated metadata for video clips." "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."
0	"This *device* has two work modes: "native" and "remote"." "The proposed rECGA is *simple*, making it amenable for further empirical and theoretical analysis."	*et al.* (from latin "*et alii*" meaning "and others") in "However, Nam *et al.* pointed out..."

Table 7. SimpleText Task 2: Scale conversion rules

Term difficulty scale	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
7 point scale	0	1	2	3	4	5	6	7
⇒ 5 point scale	0	1	2		3	4		5
7 point scale	0	1	2	3	4	5	6	7
⇒ 3 point scale	0	1			2		3	

Table 8. 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

Evaluation Metrics. We evaluated terms according to:

- correctness of term limits;
- term difficulty score on the scale 1–3;
- term difficulty score on the scale 1–5.

For both scales of term difficulty, we used a converted scale 1–7. This scale 1–7 was chosen following the psycho-linguistic research of the perception and evaluation of lexical meanings performed by Osgood and his colleagues [27], in contrast to the psychometric Likert scale (1–5, Strongly disagree/Disagree/Neither agree nor disagree/Agree/Strongly agree), commonly used in the research that employs questionnaires [21]. In the classical version of the semantic differential technique, the scale shows the variety of the human perception of semantic nuances from negative (-3) to positive (+3) polarity where 0 marks the "norm" [27]. The scale 1–7 matches the Osgood's scale and seems more suitable to evaluate concepts and features avoiding associations with negative/positive assessment. Since the 1970s, the scale has been employed in various studies as an evaluation tool for qualitative features.

Table 6 provides examples of the used term difficulty scale. We separate the examples of abbreviations from non-abbreviated phrases/words.

We added 0 for terms that should not be explained at all and we converted the original scale 1–7 as presented in Table 7.

Table 8 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 1–7.

Table 9. SimpleText Task 2: Results for the official runs

	Total	Evaluated		Score_3		Score_5	
			+Limits		+Limits		+Limits
aaac	581,285	2,951	1,388	702	318	415	175
SimpleScientificText	63,027	298	262	48	44	47	42
UAms	263,022	1,315	1,175	105	69	60	49
lea_t5	23,331	5	4	0	0	0	0

Table 10. SimpleText Task 2: Results on a subset of 167 common sentences

	Total	Evaluated		Score_3		Score_5	
			+Limits		+Limits		+Limits
aaac	581,285	833	414	200	104	127	67
UAms	263,022	574	514	46	28	25	21
SimpleScientificText	63,027	208	188	33	32	32	29

4.2 Results

A total of 4 teams submitted runs, of which 2 runs were updated. The results are given in Tables 9 and 10. In both tables, we present results for correctly attributed scores regardless the correctness of term limits (*Score_3* and *Score_5*) and the number of correctly limited terms with correctly attributed scores (+ *Limits*). Table 9 provides the results on all sentences we evaluated. However, to have comparable results for partial runs we also report scores on a subset 167 common sentences in Table 10, although we were constrained to exclude the run *lea_t5* due to a very low number of evaluated sentences.

5 Task 3: Rewrite This!

In this section, we discuss the third task about text simplification proper, rewriting an extracted sentence from a scientific abstract, addressing the task:

Given a query, simplify passages from scientific abstracts.

The goal of this task is to provide a simplified version of text passages (sentences) with regard to a query. Participants were provided with queries and abstracts of scientific papers. The abstracts could be split into sentences. The simplified passages were evaluated manually in terms of the produced errors as follows.

5.1 Evaluation Framework

Train Dataset. As for *Task 2: What is unclear?*, we provided a parallel corpus of simplified sentences from two domains: *Medicine* and *Computer Science* (see Sect. 4.1). As previously, we use scientific abstracts from the DBLP Citation Network Dataset for

Table 11. SimpleText Task 3: Statistics of the number of evaluated sentences per query

Query		# Distinct source sentences	# Distinct simplified sentences
1	*digital assistant*	370	1,280
2	*conspiracy theories*	195	398
3	*end to end encryption*	55	102
4	*imbalanced data*	55	87
5	*genetic algorithm*	51	85
6	*quantum computing*	51	85
7	*qbit*	50	76
8	*quantum applications*	42	73
9	*cyber-security*	28	47
10	*fairness*	18	22
11	*crowsourcing*	14	21

Computer Science and Google Scholar and PubMed articles on muscle hypertrophy and health *Medicine* [10,12].

Text passages issued from abstracts on computer science were simplified by either a master student in Technical Writing and Translation or a pair of experts: (1) a computer scientist and (2) a professional translator, English native speaker but not specialist in computer science [12]. Each passage was discussed and rewritten multiple times until it became clear for non-computer scientists. Medicine articles were annotated by a master student in Technical Writing and Translation specializing in this domain. Sentences were shortened, excluding every detail that was irrelevant or unnecessary to the comprehension of the study, and rephrased, using simpler vocabulary. If necessary, concepts were explained.

We provided 648 parallel sentences in total.

Test Dataset. We used the same 116,763 sentences retrieved by the ElasticSearch engine from the DBLP dataset according to the queries as for Task 2 (see Sect. 4.1). We manually evaluated 2,276 pairs of sentences for 11 queries. For the query *Digital assistant* we took the first 1,000 sentences retrieved by ElasticSearch. We pool source sentences coupled with their simplified versions submitted by all participants for all these queries. We ensured that for each evaluated source sentence the pool contained the results of all participants. The detailed statistics of the number of evaluated sentences per query for Task 3 are given in Table 11.

Input and Output Format. The input train and the test data were provided in JSON and CSV formats with the following fields:

snt_id a unique passage (sentence) identifier.
source_snt passage text.
doc_id a unique source document identifier.
query_id a query ID.
query_text simplification should be done with regard to this query.

Input example (JSON format):

```
{"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.", "doc_id":2892036907, "query_id":"G11.1",
↪ "query_text":"drones"}
```

Participants were asked to provide a list of terms to be contextualized in a JSON format or a tabulated file TSV (for manual runs) with the following fields:

run_id Run ID starting with (team_id)_(task_3)_(*name*).
manual Whether the run is manual $\{0, 1\}$.
snt_id a unique passage (sentence) identifier from the input file.
simplified_snt Text of the simplified passage.

Output example (JSON format):

```
{"run_id":"BTU_task_3_run1", "manual":1,
↪ "snt_id":"G11.1_2892036907_2", "simplified_snt":"Drones are
↪ increasingly used in the civilian and commercial domain and
↪ need to be autonomous."}
```

Evaluation Metrics. We filtered out the simplified sentences identical to the source ones and the truncated simplified sentences by keeping only passages matching the regular expression (valid snippets) `.+[?.!"]'*$'`.

Professional linguists manually annotated simplifications provided with regard to a query according to the following criteria. We evaluated binary errors:

- Incorrect syntax;
- Unresolved anaphora due to simplification;
- Unnecessary repetition/iteration (lexical overlap);
- Spelling, typographic or punctuation errors.

The lexical and syntax complexity of the produced simplifications were assessed on an absolute scale, value 1 referring to a simple output sentence regardless of the complexity of the source one, 7 corresponding to a complex one. Lexical complexity is mostly identical to that presented in Section 4.1.

We consider **syntax complexity** based on syntactic dependencies, their length and depth. The dependency trees reveal latent complications for reading and understanding text; thus, psycholinguists consider the syntactic dependencies to be a relevant tool to evaluate text readability [16]. The depth and length of the syntactic chains we interpret according to [16].

We evaluate **syntax complexity** as follows:

1. Simple sentence (without negation/passive voice): *Over Facebook, we find many interactions.*

2. Simple sentence with negation/passive voice (e.g. *Many interactions were found over Facebook*) or Simple sentences with syntactic constructions that show chains of dependency and shallow embedding depth (e.g. *Over Facebook, we find many interactions between public pages and both political wings.*)
3. Simple sentences with long chains of dependency and shallow embedding depth, with syntactic constructions like complex object, gerund construction, etc. (e.g. *Despite the enthusiastic rhetoric about the so-called collective intelligence, conspiracy theories have emerged.*) or Short complex or compound sentence (e.g. *We propose a novel approach that was used in terms of information theory.*)
4. Simple sentences with long chains of dependency and deep embedding depth, with syntactic constructions like complex object, gerund construction, etc. (e.g. *Over Facebook, we find many interactions between public pages for military and veterans, and both sides of the political spectrum*) or Complex or compound sentence that contains long chains of dependency and deep embedding depth;
5. Simple sentences with long chains of dependency and deep embedding depth, with several syntactic constructions like complex object, gerund construction, etc. or & Complex or compound sentence that contains long chains of dependency and deep embedding depth;
6. Complex or compound sentences that contain long chains of dependency and deep embedding depth along with complex object, gerund construction, etc. or Simple sentence that contains modifications, topicalization, parenthetical constructions: *Moreover, we measure the effect of 4709 evidently false information (satirical version of conspiracist stories) and 4502 debunking memes (information aiming at contrasting unsubstantiated rumors) on polarized users of conspiracy claims.*
7. Long complex or compound sentences that contain several clauses of different types, long chains of dependency and deep embedding depth along with complex object, gerund construction, etc.

We evaluate the information quality of the simplified snippet based on its content and readability. Transformation of information from the source snippet brings in omission of details, insertion of basic terms to explain particular terminology and complex concepts, reference to resources. Due to necessary insertions and references, the simplified snippets often contain more words and syntactic constructions as compared to their source. Nevertheless, the goal is to reduce lexical and syntax complexity in the extended simplified snippets. In case the simplified snippet lacks information mentioned in the source, we evaluate the degree of the information loss. Irrelevant insertions, iterations and wordy statements in the extended simplified snippet we consider as a misrepresentation or distortion of source information when a reader experiences difficulties in processing source content due to wordiness of the loosely structured simplified snippet.

We assessed the information loss severity during the simplification with regard to a given query on the scale from 1 to 7, where 1 corresponds to unimportant information loss while 7 refers to the most severe information distortion. We consider the information loss as a kind of information damage even if the information in the simplified text contains the information of the source passage but has some insertions, which impedes perception of the content.

We distinguish the following 11 types of misrepresentation of source information. Our classification leans on the error typology in machine translation [22,28].

The simplified snippet often combines several types of distortion, e.g. omission and ambiguity. Nevertheless, we observed many instances of small distortions that severely diminish the quality of the simplification; therefore, we need to explain each type providing the clear and transparent instances. Our evaluation of the value of the information distortion leans on the calculation of the information loss and assessment of the diminished readability of the simplified snippet that generates difficulties in text semantic processing by readers.

We distinguish the following types of information distortion:

1. **Style** (distortion severity 1)
   Source snippet: *In order to facilitate knowledge transfer between specialists and generalists and between experts and novices, and to promote interdisciplinary communication, there is a need to provide methods and tools for doing so.*
   Simplified snippet: *There is a need to provide methods and tools for doing so. In order to facilitate knowledge transfer between specialists and generalists and between experts and novices, we need to promote interdisciplinary communication. We need to make it easier for people to share their knowledge with each other.*
   Comment: Deviations from the style norms do not lead to information loss; however, they diminish the quality of text structure and affect readers' assessment of the text and its content.
2. **Insertion of unnecessary details with regard to a query** (distortion severity 1)
   Source snippet: *In the clinical setting, availability of needed information can be crucial during the decision-making process.*
   Simplified snippet: *availability of needed information can be crucial during the decision-making process. In the clinical setting, needed information is often difficult to come by. For confidential support call the Samaritans on 08457 90 90 90 or visit a local Samaritans branch, see* www.samaritans.org *for details.*
   Comment: The simplified snippet often contains more information than the source since the terminology is needed to be explained. An irrelevant insertion does not lead to the loss of information; however, it may bring in diminishing of the text readability and generate discomfort during text perception. The irrelevant reference to the support in the simplified snippet does not clarify the source. The source does not need any simplification.
3. **Redundancy** (without lexical overlap) (distortion severity 2)
   Source snippet: *The capability to get updated information and news is an important and decisive factor in business and finance.*
   Simplified snippet: *The capability to get updated information and news is an important and decisive factor in business and finance. The ability to get updates on the latest news is also an important factor in the success of a business or finance company. For more information, visit CNN.com/News.*
   Comment: Irrelevant insertions, iterations and wordy statements in the extended simplified snippet we consider as a misrepresentation or distortion of source information when a reader may misunderstand source content due to wordiness of the loosely structured simplified snippet.
4. **Insertion of false or unsupported information** (distortion severity 3)
   Source snippet: *The proposed method leads to not only faster running time but also efficient text localization.*

Simplified snippet: *The proposed method leads to not only faster running time but also efficient text localization. The proposed method is based on the idea that text should be written in a single block of text, rather than a series of blocks of text. The method is being developed by the University of California, San Diego.*
Comment: False and unsupported information is inserted in the simplified snippet because of external links of the source snippet to the open Web resources. False information confuses the readership, thus it is inappropriate in simplified texts.

5. **Omission of essential details with regard to a query** (distortion severity 4)
Source snippet: *In short, this thesis aims to repatriate young people's web practices from the sterile, positivist methods space of questionnaires and tests of digital literacy to social contexts of everyday life.*
Simplified snippet: *In short, this thesis aims to repatriate young people's web practices from the sterile, and tests of digital literacy.*
Comment: Simplified texts regularly omit details and minor concepts since the objective of the simplification is to facilitate perception and processing of the main idea and novel information of the source. Therefore, distinguishing the essential details and concepts from the minor ones plays a crucial role in the text simplification. Omission of the essential details leads to the lack of information and novelty of the message. The simplified snippet lacks the essential information about the direction where the young people's web practice needs to be repatriated (*to social contexts of everyday life*).
6. **Overgeneralization** (distortion severity 5)
Source snippet: *Online Social Networks explode with activity whenever a crisis event takes place.*
Simplified snippet: *It explodes when a crisis event takes place.*
Comment: In the simplified snippet the subject of the source snippet *Online Social Networks* is omitted being substituted by the pronoun *It*; also the essential detail *with activity* is omitted that brings in overgeneralization: the statement refers to all cases of the dysfunctions (server, browser, users' activity, moderators' restrictions, etc.)
7. **Oversimplification** (distortion severity 5)
Source snippet: *If we accept the current orthodoxy and then blame the Web we offer a technological determinist explanation of reality: technology produces misinformed populations.*
Simplified snippet: *If we accept the current orthodoxy and then blame the Web we have an explanation of reality.*
Comment: Oversimplification appears in the shortened simplified snippets when source utterance is transformed into a trivial statement or even a claim. The simplified snippet claims that *we have an explanation of reality* thanks to acceptance of the current orthodoxy and accusation of the Web. Meanwhile, the source snippet discusses the technological determinist explanation of reality. The omission of the essential details leads to the oversimplified statement that cannot explain the reality of the technological epoch.
8. **Topic shift** (distortion severity 5)
Source snippet: *global warming induced by chemtrails or the link between vaccines and autism - find on the Web a natural medium for their dissemination.*

Simplified snippet: *The link between vaccines and autism - can be found on the Web a natural medium for changing.*
Comment: Topic shift is revealed in a substitution of the source topic by omitting its part or selecting a wrong basic word to replace the peculiar term in the source. The source snippet lost the essential part of its topic (*global warming induced by chemtrails*) during the simplification process; moreover, the simplification resulted in the inappropriate syntactic structure of the snippet.

9. **Contra sense/contradiction** (distortion severity 6)
Source snippet: *In this paper we discuss architectural design issues and trade-offs in connection with our experiences porting our agent-based platform, Opal, to the Sharp Zaurus personal digital assistant (PDA).*
Simplified snippet: *The Sharp Zaurus is a personal digital assistant (PDA) developed by Sharp. It is based on the Opal agent-based platform. We discuss architectural design issues and trade-offs in connection with our experiences porting Opal to the Zaurus PDA.*
Comment: Contradictions in simplified snippets appear due to elimination of essential concepts or interrelations among concepts, omission of significant details, and transformation of the semantic structure of the source snippet. The simplified snippet mentions agent-based platform Opal as the basis for the Sharp Zaurus, but at the same time claims that Opal was ported to the Sharp Zaurus. The source snippet *But the new phenomena, the non-agenda ownership, overcome any ideological influence, especially under the conditions of punishment mechanism applied to old politicians* lost its semantic structure since the concepts *ideological influence* and *punishment mechanism* were eliminated in the process of its simplification. Thus, the simplified snippet *But the new phenomena, the ownership of the non-agenda, had a lot of influence on old politicians* lacks any explanation how *the non-agenda ownership* is related to *old politicians* and why they are influence by *the new phenomena*.
10. **Ambiguity** (distortion severity 6)
Source snippet: *The experimental results show that 3D maps with texture on mobile phone display size, and 3D maps without texture on PDA display size are superior to 2D maps in search time and error rate.*
Simplified snippet: *3D maps with texture on mobile phone display size are superior to 2D maps in search time and error rate. The experimental results show that 3D maps without texture on PDA display size were superior to those with texture. The results were published in the journal 3D Maps.*
Comment: Ambiguity presupposes that a statement has several equiprobable interpretations. The instance of the ambiguous simplified snippet above lacks a key to understand whether the 3D maps without texture outperform those with texture or not. Ambiguity often appears due to syntactic simplification of the source. In the source, the clause *changes in the strength of competition also reveal key asymmetrical differences* is replaced by shorter clause *but they do not have any biases* that produces ambiguity: whether *evidence* corresponds to reality or not. The source clarifies the differences between two political parties: *Though both Republicans and Democrats show evidence of implicit biases, changes in the strength of competition also reveal key asymmetrical differences* however, the simplified snippet

doubts the reliability of the evidence: *Both Republicans and Democrats show evidence of biases, but they do not have any biases*. Readers of the simplified snippet are unable to resolve the ambiguity.

11. **Nonsense** (distortion severity 7)
    Source snippet: *The large availability of user provided contents on online social media facilitates people aggregation around shared beliefs, interests, worldviews and narratives.*
    Simplified snippet: *The large amount of user provided contents on online social media is called aggregation.*
    Comment: The source snippet was transformed into a simple sentence. The transformation brings in erroneous usage of the word aggregation that leads to the loss of meaning of the whole sentence. Instead of the original statement about accessibility of the social or public media on the Web, which facilitates dissemination of fake news and rumors, the simplified snippet claims that there is an opportunity to find a resource to read about fake news and rumors.

The final ranking for Task 3 was done by the average harmonic mean of normalized opposite values of *Lexical Complexity (LC), Syntactic Complexity (SC)* and *Distortion Level (DL)* as follows:

$$s_i = \frac{3}{\frac{7}{7-LC} + \frac{7}{7-SC} + \frac{7}{7-DL}} \tag{1}$$

$$Score = \frac{\Sigma_i \begin{cases} s_i, & \text{if } No\ Error \\ 0, & \text{otherwise} \end{cases}}{n} \tag{2}$$

In Eq. 2, variable n refers to the total number of judged snippets and *No Error* means that the snippet i does not have any of *Uncorrect syntax, Unresolved anaphora*, nor *Unnecessary repetition/iteration* error.

5.2 Results

A total of 5 different teams submitted 14 runs (5 runs were updated). Absolute number of errors and average *Lexical Complexity, Syntax Complexity* and *Information Loss* are provided in Tables 12 and 13. The final ranking for Task 3 is given in Table 14. We removed all runs with the 0 score.

Very interesting partial runs were provided by the HULAT-UC3M team as the generated simplifications provided the explanations of difficult terms. However, HULAT-UC3M's 8 runs over 10 were not in the pool with selected topics. Thus, we provided only automatic evaluation results. The HULAT-UC3M's runs provide clear evidence of the interconnection of tasks 2 and 3.

Table 12. SimpleText Task 3: General results of official runs

Run	Total	Unchanged	Truncated	Valid	Longer	Length Ratio	Evaluated	Uncorrect Syntax	Unresolved Anaphora	Minors	Syntax Complexity	Lexical Complexity	Information Loss
CLARA-HD	116,763	128	2,292	111,627	201	0.61	851	28	3	68	2.10	2.42	3.84
CYUT Team2	116,763	549	101,104	111,818	49	0.81	126	1		32	2.25	2.30	2.26
PortLinguE_full	116,763	42,189	852	111,589	3,217	0.92	564	7		5	2.94	3.06	1.50
PortLinguE_run1	1,000	359	7	970	30	0.93	80	1			3.63	3.57	2.27
lea_task3_t5	23,360	52	23,201	22,062	24	0.35	.	.	.	.	.	.	.
HULAT-UC3M01	1,000	.	13	973	968	2.46	95	10	1	20	4.69	3.69	2.20
HULAT-UC3M02	2,001	3	58	1,960	1,920	2.53	205	10	1	37	3.60	3.53	2.34
HULAT-UC3M03	1,000	2	13	958	966	2.53	.	.	.	.	.	.	.
HULAT-UC3M04	2,000	.	33	1,827	1,957	37	.	.	.	.	.	.	.
HULAT-UC3M05	2,000	.	56	1,921	1,918	2.38	.	.	.	.	.	.	.
HULAT-UC3M06	2,000	.	47	1,976	1,921	2.45	.	.	.	.	.	.	.
HULAT-UC3M07	1,000	.	56	970	972	2.43	.	.	.	.	.	.	.
HULAT-UC3M08	2,000	.	62	1,964	1,919	2.59	.	.	.	.	.	.	.
HULAT-UC3M09	2,000	.	170	1,964	1,904	2.15	.	.	.	.	.	.	.
HULAT-UC3M10	2,000	.	215	1,963	1,910	2.13	.	.	.	.	.	.	.

Table 13. SimpleText Task 3: Information distortion in evaluated runs

Run	Evaluated	Non-Sense	Contresens	Topic Shift	Wrong Synonym	Ambiguity	Omission Of Essential Details	Overgeneralization	Oversimplification	Unsupported Information	Unnecessary Details	Redundancy	Style
CLARA-HD	851	162	68	37	20	80	314	59	203	26	10	29	13
CYUT Team2	126	2	1	.	.	4	42	4	5	.	.	.	4
PortLinguE_full	564	9	3	4	3	19	94	9	13	2	2	5	1
PortLinguE_run1	80	.	.	1	.	.	27	5	2	.	.	.	.
lea_task3_t5	.	.	.	.	.	.	.	.	.	.	.	.	.
HULAT-UC3M01	95	1	7	2	.	5	2	.	1	5	38	36	.
HULAT-UC3M02	205	4	9	4	.	9	4	.	.	12	72	61	1

Table 14. SimpleText Task 3: Ranking of official submissions on combined score

Run	Score
PortLinguE_full	0.149
CYUT Team2	0.122
CLARA-HD	0.119

6 Conclusion and Future Work

We introduced the CLEF 2022 SimpleText track, containing three interconnected shared tasks on scientific text simplification. We pipelined the passages retrieved for Task 1 in order to rank difficult terms (Task 2) and simplify sentences (Task 3). We evaluated term difficulty and simplifications with regard to the queries from Task 1.

For Task 1, we created a large corpus of scientific abstracts, a set of popular science requests with detailed relevance judgments on the level of relevance of scientific abstracts to the request and broader context of a newspaper article on this topic. The abstracts of scientific papers retrieved for these requests were used in the follow up tasks. For Task 2 and 3, we created a corpus of sentences extracted from the abstracts of scientific publications, with manual annotations of term complexity (Task 2). In contrast to previous work, we evaluate simplification in terms of lexical and syntax complexity combining with error analysis. We introduced a new classification of information distortion types for automatic simplification and we annotated the collected simplifications according to this error classification (Task 3). Recent pandemics have shown that simplification can be modulated by political needs and the scientific information can be distorted. Thus, in contrast to previous work, we evaluated the simplifications in terms of information distortion.

For next year, we plan continue the Task 1 setup, but also refine the relevance judgments to sentence level, and provide additional evaluation measures of readability levels. We will extend Task 2 to provide a context to difficult terms and we will work on automatic metrics based on the insights we obtained this year. In particular, for Task 2, participants will be asked to provide context for difficult terms. This context should provide a definition and take into account ordinary readers' needs to associate their particular problems with the opportunities that science provides them to solve the problems [29]. This year, the HULAT-UC3M team submitted runs which combine tasks 2 and 3 which demonstrates strong interconnection of the tasks as often the terminology cannot be removed nor simplified but it needs to be explained to a reader. Finally, we plan to continue the Task 3 setup, continuing the detailed manual annotations of samples, but also working on automatic metrics that best reflect the insights of this year's analysis.

For details about this year's track and the approaches of individual teams we refer to the CLEF CEUR proceedings [15]. 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!

Acknowledgment. We like to acknowledge the support of the Lab Chairs of CLEF 2022, Allan Hanbury and Martin Potthast, for their help and patience.Special thanks to the University Translation Office of the Université de Bretagne Occidentale, and to Nicolas Poinsu and Ludivine Grégoire for their major impact in the train data construction and Léa Talec-Bernard and Julien Boccou for their help in evaluation of participants' runs. We thank Josiane Mothe for reviewing papers. We also thank Alain Kerhervé, and the MaDICS (https://www.madics.fr/ateliers/simpletext/ research group.

References

1. Text Analysis Conference (TAC) 2014 Biomedical Summarization Track (2014). https://tac.nist.gov/2014/BiomedSumm/
2. Alva-Manchego, F., Martin, L., Bordes, A., Scarton, C., Sagot, B., Specia, L.: Asset: a dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations (2020). https://arxiv.org/abs/2005.00481
3. Anand Deshmukh, A., Sethi, U.: IR-BERT: leveraging BERT for semantic search in background linking for news articles 2007, July 2020. http://adsabs.harvard.edu/abs/2020arXiv200712603A
4. Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., Tannier, X.: INEX tweet contextualization task: evaluation, results and lesson learned. Inf. Process. Manage. **52**(5), 801–819 (2016). https://doi.org/10.1016/j.ipm.2016.03.002
5. Brown, T.B., et al.: Language models are few-shot learners, July 2020. http://arxiv.org/abs/2005.14165
6. Chandrasekaran, M.K., Feigenblat, G., Hovy, E., Ravichander, A., Shmueli-Scheuer, M., de Waard, A.: Overview and insights from the shared tasks at scholarly document processing 2020: Cl-scisumm, laysumm and longsumm. In: Proceedings of the First Workshop on Scholarly Document Processing, pp. 214–224 (2020)
7. Cohan, A., Goharian, N.: Revisiting summarization evaluation for scientific articles, April 2016. http://arxiv.org/abs/1604.00400
8. De Clercq, O., Hoste, V., Desmet, B., van Oosten, P., De Cock, M., Macken, L.: Using the crowd for readability prediction. Nat. Lang. Eng. **20**(3), 293–325 (2014). http://dx.doi.org/10.1017/S1351324912000344. ISSN 1469–8110
9. Dong, Y., Li, Z., Rezagholizadeh, M., Cheung, J.C.K.: EditNTS: an neural programmer-interpreter model for sentence simplification through explicit editing. In: Proceedings of the 57th Annual Meeting of the ACL, Florence, Italy, pp. 3393–3402. ACL, July 2019. https://www.aclweb.org/anthology/P19-1331
10. Ermakova, L., et al.: Overview of SimpleText 2021 - CLEF workshop on text simplification for scientific information access. In: Candan, K.S., et al. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 432–449. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_27
11. Ermakova, L., et al.: Text simplification for scientific information access: CLEF 2021 SimpleText workshop. In: Proceedings of Advances in Information Retrieval - 43nd European Conference on IR Research, ECIR 2021, Lucca, Italy, 28 March–1 April 2021 (2021)
12. Ermakova, L., et al.: Automatic simplification of scientific texts: SimpleText lab at CLEF-2022. In: Hagen, M., et al. (eds.) Advances in Information Retrieval, vol. 13186, pp. 364–373. Springer, Cham (2022). ISBN 978-3-030-99738-0 978-3-030-99739-7
13. Ermakova, L., Goeuriot, L., Mothe, J., Mulhem, P., Nie, J.-Y., SanJuan, E.: CLEF 2017 microblog cultural contextualization lab overview. In: Jones, G.J.F., et al. (eds.) CLEF 2017. LNCS, vol. 10456, pp. 304–314. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65813-1_27

14. Ermakova, L.N., Nurbakova, D., Ovchinnikova, I.: Covid or not Covid? Topic shift in information cascades on Twitter. In: Association for Computational Linguistics (ed.) 3rd International Workshop on Rumours and Deception in Social Media (RDSM) Collocated with COLING 2020, pp. 32–37. Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM), Barcelona, Spain, December 2020. https://hal.archives-ouvertes.fr/hal-03066857
15. Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (eds.): Proc. of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings (2022)
16. Futrell, R., et al.: The natural stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. Lang. Resour. Eval. **55**(1), 63–77 (2021). https://doi.org/10.1007/s10579-020-09503-7. ISSN 1574-0218
17. Gala, N., François, T., Fairon, C.: Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. In: eLex-Electronic Lexicography (2013)
18. Grabar, N., Farce, E., Sparrow, L.: Study of readability of health documents with eye-tracking approaches. In: 1st Workshop on Automatic Text Adaptation (ATA) (2018)
19. Hoffart, J., et al.: Robust disambiguation of named entities in text. In: Proceedings of EMNLP 2011, pp. 782–792 (2011)
20. Lieber, O., Sharir, O., Lentz, B., Shoham, Y.: Jurassic-1: technical details and evaluation, p. 9 (2021)
21. Likert, R.: A technique for the measurement of attitudes. Arch. Psychol. **22**(140), 55 (1932)
22. Lommel, A., Görög, A., Melby, A., Uszkoreit, H., Burchardt, A., Popović, M.: Harmonised metric. Qual. Transl. **21**(QT21) (2015). https://www.qt21.eu/wp-content/uploads/2015/11/QT21-D3-1.pdf
23. Maddela, M., Alva-Manchego, F., Xu, W.: Controllable text simplification with explicit paraphrasing, April 2021. http://arxiv.org/abs/2010.11004
24. Maddela, M., Xu, W.: A word-complexity lexicon and a neural readability ranking model for lexical simplification. In: Proceedings of EMNLP 2018, Brussels, Belgium, pp. 3749–3760. ACL (2018). https://www.aclweb.org/anthology/D18-1410
25. Nakov, P., et al.: Automated fact-checking for assisting human fact-checkers, May 2021. http://arxiv.org/abs/2103.07769
26. Narayan, S., Gardent, C., Cohen, S.B., Shimorina, A.: Split and rephrase. In: Proceedings of EMNLP 2017, Copenhagen, Denmark, pp. 606–616. ACL, September 2017. https://www.aclweb.org/anthology/D17-1064
27. Osgood, C.E.: Semantic differential technique in the comparative study of cultures. Am. Anthropol. **66**(3), 171–200 (1964). https://onlinelibrary.wiley.com/doi/abs/10.1525/aa.1964.66.3.02a00880. ISSN 1548-1433
28. Ovchinnikova, I.: Impact of new technologies on the types of translation errors. In: CEUR Workshop Proceedings (2020)
29. Ovchinnikova, I., Nurbakova, D., Ermakova, L.: What science-related topics need to be popularized? A comparative study. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, 21–24 September 2021, vol. 2936, pp. 2242–2255. CEUR Workshop Proceedings (2021). http://ceur-ws.org/Vol-2936/paper-203.pdf
30. O'Reilly, T., Wang, Z., Sabatini, J.: How much knowledge is too little? When a lack of knowledge becomes a barrier to comprehension. Psychol. Sci., July 2019. https://journals.sagepub.com/doi/10.1177/0956797619862276
31. Pradeep, R., Ma, X., Nogueira, R., Lin, J.: Scientific claim verification with VerT5erini, October 2020. http://arxiv.org/abs/2010.11930

32. Sulem, E., Abend, O., Rappoport, A.: Simple and effective text simplification using semantic and neural methods. In: Proceedings of the 56th Annual Meeting of the ACL (Volume 1: Long Papers), Melbourne, Australia, pp. 162–173. ACL, July 2018. https://www.aclweb.org/anthology/P18-1016
33. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: extraction and mining of academic social networks. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 2008, Las Vegas, Nevada, USA, p. 990. ACM Press (2008). http://dl.acm.org/citation.cfm?doid=1401890.1402008. ISBN 978-1-60558-193-4
34. Wadden, D., et al.: Fact or fiction: verifying scientific claims, October 2020. http://arxiv.org/abs/2004.14974
35. Wang, W., Li, P., Zheng, H.T.: Consistency and coherency enhanced story generation, October 2020. http://arxiv.org/abs/2010.08822
36. Wubben, S., van den Bosch, A., Krahmer, E.: Sentence simplification by monolingual machine translation. In: Proceedings of the 50th Annual Meeting of the ACL (Volume 1: Long Papers), pp. 1015–1024 (2012)
37. Xu, W., Callison-Burch, C., Napoles, C.: Problems in current text simplification research: new data can help. Trans. ACL **3**, 283–297 (2015). https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_00139. ISSN 2307-387X
38. Xue, L., et al.: mT5: a massively multilingual pre-trained text-to-text transformer. In: Proceedings of the 2021 Conference of the North American Chapter of the ACL: Human Language Technologies, pp. 483–498. ACL, June 2021. https://aclanthology.org/2021.naacl-main.41
39. Yang, L., Zhang, M., Li, C., Bendersky, M., Najork, M.: Beyond 512 tokens: siamese multi-depth transformer-based hierarchical encoder for long-form document matching, April 2020. arXiv:2004.12297
40. Zhao, S., Meng, R., He, D., Saptono, A., Parmanto, B.: Integrating transformer and paraphrase rules for sentence simplification. In: Proceedings of EMNLP 2018, Brussels, Belgium, pp. 3164–3173. ACL, October 2018. https://www.aclweb.org/anthology/D18-1355
41. Zhong, Y., Jiang, C., Xu, W., Li, J.J.: Discourse level factors for sentence deletion in text simplification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 05, pp. 9709–9716, April 2020. https://ojs.aaai.org/index.php/AAAI/article/view/6520. ISSN 2374-3468