

# University of Amsterdam at the CLEF 2022 SimpleText Track

Femke Mostert, Ashmita Sampatsing, Mink Spronk, David Rau and Jaap Kamps

University of Amsterdam, Amsterdam, The Netherlands

## Abstract

This paper reports on the University of Amsterdam's participation in the CLEF 2022 SimpleText track. The overall goal of removing barriers that prevent the general public from accessing scientific literature is of great importance to help users make sense of a world of misinformation and shallow opinions. We perform preliminary studies within the track's setup, analyzing the text complexity of searching a large set of academic abstracts in the context of popular science topics emerging in the news, with a specific focus at the relation between the *topical relevance* and the *text complexity* of the retrieved information. Our main findings are the following. First, we analyzed a large corpus of scientific abstracts and confirmed that these are highly complex on average, but that the variation is large and many abstracts with accessible readability levels exist. Second, we ran retrieval experiments and found that standard search ignores readability, yet filtering on the desirable reading level still retains competitive performance while avoiding retrieving relevant but incomprehensible results. Third, we ran complexity spotting experiments and found that straightforward lexical complexity or term frequency measures are strong indicators, but have to be combined with the importance of the concept in the broader context of the information request. Fourth, we ran a GPT-2 based text simplification model in a zero-shot way, resulting in conservative rewriting of abstracts, able to significantly reduce the text complexity. More generally, our results demonstrate that text complexity is an essential aspect to consider for improving non-expert access to scientific information, and opens up new routes to develop effective scientific information access technology tailored to needs of the general public.

## Keywords

Information Storage and Retrieval, Natural Language Processing, Scientific Information Access, Text Simplification

## 1. Introduction

The advent of the internet and social media has been revolutionary in changing every aspect of information creation and information consumption. In the early years, many have lauded all the positives resulting from this, such as breaking down traditional barriers in access to information, as well as providing the means to publish anything by anyone [1]. One of the prime examples of this is Wikipedia. This is leading in principle to ultimate democratization by giving every global citizen a voice, and ultimate equality where quality of arguments rather than the provenance of the author decides the outcome.

In recent years, the emphasis has shifted to the negatives resulting from this, as not only

---


CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

✉ kamps@uva.nl (J. Kamps)

ORCID 0000-0002-6614-0087 (J. Kamps)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

individual citizens have massively joined this open sphere, but also many other commercial or political actors have discovered its potential to influence citizens to suit their financial or political incentives, even leading to social and societal risks. Hence, one of the greatest challenges of today is how users can navigate in a world of misinformation and disinformation. This is a very hard and complex problem, and rather than believing in fairy tales and some magic wand that will make it disappear, we focus on a well-established antidote: objective, scientific information in the academic literature and associated data.

Every citizen agrees on the importance of objective scientific evidence, yet at the same time they predominantly rely on shallow secondary information on the web and in social media. One of the main reasons for not accessing scientific information directly is that they presume the scientific literature is too difficult. The CLEF 2022 SimpleText track investigates the barriers that ordinary citizens face when accessing scientific literature head-on, by making available corpora and tasks to address different aspects of the problem. For details on the exact track setup, we refer to the Track Overview paper CLEF 2022 LNCS proceedings [2], as well as the detailed task overviews in the CEUR proceedings [3, 4, 5].

We conduct an extensive analysis of the corpus of scientific abstracts and the three tasks of the track: Task 1 on content selection and *avoiding* complexity; Task 2 on complexity spotting in extracted sentences from scientific abstracts; and Task 3 on text simplification proper rewriting sentences from these abstracts. The rest of this paper is structured as follows. Next, in Section 2 we discuss our experimental setup and the specific runs submitted. Section 3 discusses the results of our runs and provides a detailed analysis of the corpus and results for each task. We end in Section 4 by discussing our results and outlining the lesson learned.

## 2. Experimental Setup

In this section, we will detail our approach for the three CLEF 2022 SimpleText track tasks.

For details of the exact task setup and results we refer the reader to the detailed overview of the track in Ermakova et al. [2]. The basic ingredients of the track are:

**Corpus** The CLEF 2022 SimpleTrack Corpus consists of 4.9 million bibliographic records, including 4.2 million abstracts, and detailed information about authors/affiliations/citations.

**Context** There are 40 popular science articles, with 20 from *The Guardian*<sup>1</sup> and 20 from *Tech Xplore*.<sup>2</sup>

**Requests** For Task 1, there are 114 requests with 1-4 queries per context article, 47 requests are based on The Guardian and 67 on TechXplore. For Task 2, there are 453 train and 116,764 test sentences. For Task 3, there are 648 train and 116,764 test sentences.

**Assessments** For Task 1, there are qrels for 72 requests (67 requests with at least 1 marginally relevant abstract, 22 requests with at least 5 relevant abstracts).

We created runs for all the three tasks of the track, which we will discuss in order.

---

<sup>1</sup><https://www.theguardian.com/science>

<sup>2</sup><https://techxplore.com/>

**Table 1**

Text complexity: readability in school grade levels

Grade Level	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
School	<i>Elementary</i>					<i>Jr. High</i>			<i>High School</i>				<i>Undergrad.</i>			<i>Grad.</i>		<i>PhD</i>		
	<i>Primary</i>				<i>Secondary</i>						<i>University</i>				<i>PhD</i>					
	<i>Compulsory</i>												<i>Higher Edu.</i>							
Age	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

**Task 1** This task requires ranking scientific abstracts in response to a non-expert, general query prompted by a popular science article. We submitted two runs.

The first run, labeled **UAmS-MF** in [2, 3], is a manual run selecting relevant and accessible results from the top 5 of a vanilla Elastic Search run.

Our second run, labeled **UAmS** in [2, 3], is an automatic runs using a reading level/text complexity score as a filter. Specifically, per request and the top 100 result of a vanilla Elastic Search run, we remove 50% of the abstracts with the highest text complexity based on the popular Flesch readability level score.

**Task 2** What concept needs to be explained or rewritten in a given sentence, extracted from a scientific abstract.

Based on preliminary experiments, our submission also labeled **UAmS** in [2, 4] is using an idf-based term weighting to locate the most rare terms, combined with a simple way to boost particular syntactic categories. Specifically, we used all train and test sentences combined as a reference corpus to calculate document (or rather sentence) frequencies, and use this to rank each term in the source sentence by increasing DF (or decreasing IDF). We include adhoc boost factors for particular part-of-speech, promoting nouns and demoting verbs and adjectives.

**Task 3** Rewrite a sentence from a scientific abstract.

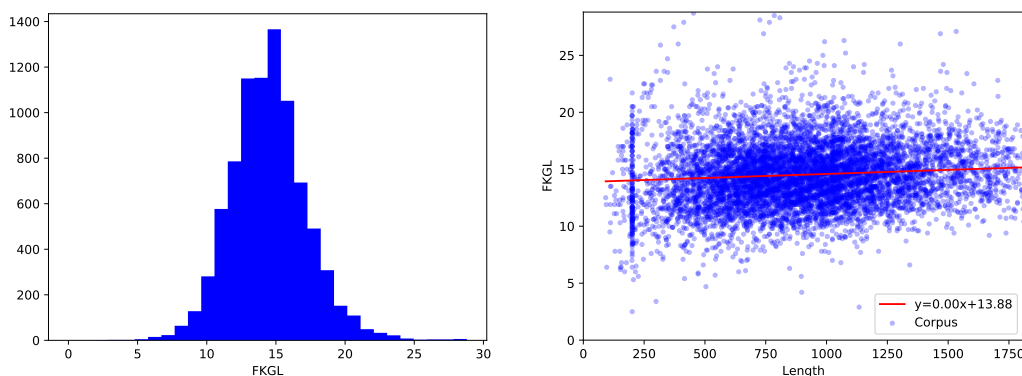
This is a post-submission run, hence not evaluated in the track and task overview papers [2, 5]. We use a standard text simplification model, based on the GPT-2 based keep it simple (KiS) model of Laban et al. [6]. We run a pretrained version of this model available from HuggingFace,<sup>3</sup> in a zero-shot way on both the train and test corpus.

### 3. Experimental Results

In this section, we will present the results of our experiments, in four self-contained subsections following the CLEF 2022 SimpleText Track corpus and tasks.

**Table 2**  
CLEF 2022 SimpleText Data: Flesch-Kincaid Grade Level.

Data	Sample Size	Length		FKGL	
		Mean	Median	Mean	Median
Corpus (scientific abstracts)	8,513	951	905	14.55	14.40
News (popular science)	40	5,504	5,540	12.53	12.70
Retrieved results (top 100)	11,400	948	928	13.79	14.40



**Figure 1:** CLEF 2022 SimpleText Corpus: distribution of text complexity in Flesch-Kincaid Grade Levels.

### 3.1. Corpus, Context and Requests

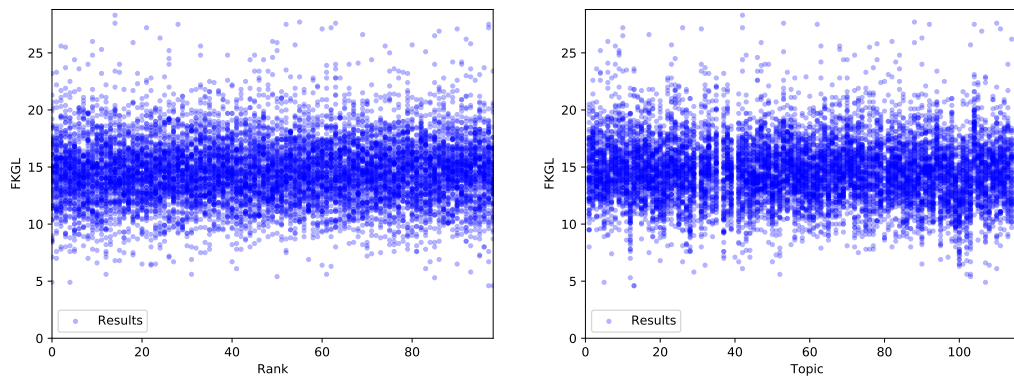
We start with a preliminary analysis of the complexity of the scientific abstracts, in relation to the context and requests. To quantify the complexity, we use the Flesch-Kincaid Grade Level (FKGL) measure based on the lexical and grammatical complexity. This is a simple measure based on word length and sentence length, which may not be the most accurate for a single abstract but a reliable approximation when averaging over larger sets of data. The FKGL score is calibrated to correspond to the readability level suitable for a given school level in the U.S. school system, as shown in Table 1. While literacy levels vary in the population, even among adults, one may assume that an average layperson would have finished compulsory education, corresponding to a high school diploma at a grade level of 12.

#### 3.1.1. Complexity of the Corpus

We down-sampled the corpus by taking every 500th article, resulting in an arbitrary sample of 8,513 non-empty abstracts. As shown in Table 2, the average (median) length of the abstracts is 951 (905) tokens, and the average (median) complexity of the abstracts is 14.55 (14.4) FKGL.

How complex are scientific abstracts? We can immediately confirm that scientific literature is indeed complex: the scale is the U.S. grade levels in years, with 12 being the exit level of compulsory education (high school diploma), hence the observed complexity of 14-15 is

<sup>3</sup>[https://huggingface.co/philippelaban/keep\\_it\\_simple](https://huggingface.co/philippelaban/keep_it_simple)



**Figure 2:** CLEF 2022 SimpleText Top 100 results: distribution of text complexity in Flesch-Kincaid Grade Levels.

translating to students half-way in undergraduate or college education.

What is the target level of complexity? Recall that the track also provides 40 popular science articles from The Guardian and TechXplore, which are written by professional science journalists for a general audience. As also shown in Table 2, the average (median) length of these articles is 5,504 (5,540) tokens, and the average (median) complexity of the articles is 12.53 (12.7) FKGL, confirming that a FKGL around 12, translating to the readability level of a high school diploma, is appropriate for general citizens.

Is every single abstract too complex for an average citizen? Figure 1 (left) shows the distribution of FKGL readability levels, which show a striking variation ranging from 5 (elementary school, 10 year old children) to 25 (graduate school domain expert). Figure 1 (right) visualizes this extreme variation, plotted against the length of the abstracts. There is in fact a weak correlation between text complexity and length ( $r=0.1059$ , highly significant, regression line with slope 0.0007 in red), but for any length we find abstracts on any level of readability.

Our analysis confirms the presumption that scientific literature is complex, and a large fraction of abstracts would be very challenging for a layperson. However, our analysis also reveals that a large fraction of abstracts is within the readability levels of most adult citizens.

### 3.1.2. Complexity of the Requests

What subset of abstracts is selected by a general query based on the popular science newspaper articles? We use the default elastic search engine, and retrieve the top 100 scientific articles for each request, and analyze the text complexity of each retrieved abstract. Over the 114 queries, this results in a sample of 11,400 abstracts. As shown also in Table 2, the average (median) length of the retrieved abstracts is 948 (928) tokens, and the average (median) complexity of the abstracts is 13.79 (14.4) FKGL. Hence, the retrieved abstracts are comparable to the corpus statistics, both in terms of length and text complexity, and also the distribution of FKGL (not shown) is very similar.

Figure 2 shows the distribution of FKGL readability levels over rank of retrieval (left-hand side), and over each individual query (right-hand side). In both cases we see that the standard

**Table 3**

Results for the SimpleText Task 1: Manual run and filter on median readability score per query.

Run	Score@5	#Docs@5	Doc Avg	#Queries	Query Avg
Manual selection of top 5 Elastic Search	163	54	0.87	99	1.65
Keep only 50% most readable abstracts	52	17	0.22	40	1.30

retrieval engine is completely blind to the text complexity, and exclusively focusing on the topical relevance of the abstract. As a result, for any rank and any topic, we see again a striking variation in FKGL, ranging from 10 (starting high school, 15 year old children) to 20 (doctoral/PhD candidate).

There are three conclusions based on our analysis: First, a negative result is that we can confirm that scientific abstracts on average are complex, confirming and validating the presumption of laypersons avoiding scientific information. Second, a positive result is that we also found that the variation of complexity is dramatic, and a large fraction of abstracts is within the readability level of an average educated adult citizen. Third, standard search engines are optimized for topic relevance and are completely ignoring secondary aspects such as the reading level of the text.

Generally, this last finding can immediately explain why laypersons have a disappointing experience when searching academic literature, and explain why they avoid academic sources even when they care about objective evidence. However, note that this finding is also actionable to dramatically improve existing search technology, by explicitly factoring readability into the ranking model, and retrieving all *and only* relevant abstracts that are not prohibitively complex for an interested outsider.

### 3.2. Task 1: Content Selection

We continue with Task 1, asking to retrieve scientific articles in response to a query based on a popular science article. We submitted two runs, one based on a manual selection of the top 5 results of the default elastic search engine, and an automatic run filtering the top 100 results for the 50% of abstracts with the lowest text complexity based on the readability measure.

Table 3 shows the results of our official submissions for Task 1. All scores are based on the top 5 abstracts retrieved per query, based on a pool of abstracts retrieved by at least 2 submissions, and evaluated on a scale ranging from 0 (irrelevant articles) to 5 (when the abstract and keywords are relevant with the query and the content of the original article). We see that the set of documents in the top 5 of the manual run has a higher level of relevance than the readability level filtered automatic run, and also finds more relevant abstracts per query.

Using the final qrels, we can also evaluate using familiar search measures, with the resulting evaluation being shown in Table 4. We also include here the evaluation of the standard Elastic Search for the designated queries (regular queries without quotes). The qrels include 72 topics, but our manual run does not include results for topics in which the top 5 abstracts as returned by Elastic search API were deemed non-relevant for the context of the article. Hence we also evaluate the manual run over the intersection of 52 topics, for which the manual run includes

**Table 4**  
Evaluation of SimpleText Task 1 (graded measures).

Run	Top.	NDCG			FKGL	
		5	10	20	Mean	Median
Elastic	72	0.4053	0.4334	0.4438	13.79	14.40
Automatic	72	0.3531	0.3776	0.4073	11.70	12.80
Manual	72	0.3494	0.3328	0.3270	14.80	14.80
Manual	52	0.4837	0.4608	0.4528	–	–

**Table 5**  
Evaluation of SimpleText Task 1 (boolean measures).

Run	Top.	Rel.	MRR	Precision			MAP
				5	10	20	
Elastic	72	1+	0.5315	0.3333	0.2139	0.1229	0.4100
Automatic	72	1+	0.5003	0.2917	0.1931	0.1333	0.3706
Manual	72	1+	0.5289	0.2750	0.1375	0.0687	0.2813
Manual	52	1+	0.7324	0.3808	0.1904	0.0952	0.3895
Elastic	72	2+	0.4673	0.2889	0.1792	0.1035	0.3619
Automatic	72	2+	0.4404	0.2417	0.1528	0.1028	0.3192
Manual	72	2+	0.4537	0.2417	0.1208	0.0604	0.2599
Manual	52	2+	0.6282	0.3346	0.1673	0.0837	0.3599
Elastic	72	4+	0.1889	0.0889	0.0542	0.0312	0.1447
Automatic	72	4+	0.2048	0.0778	0.0458	0.0333	0.1580
Manual	72	4+	0.2118	0.0889	0.0444	0.0222	0.1504
Manual	52	4+	0.2933	0.1231	0.0615	0.0308	0.2082

at least 1 result.

We see that the manual run, retrieving between 1 and 5 results per query, has superior early precision, higher than the Elastic Search baseline. We also see that our automatic run obtains very reasonable performance with an NDCG@10 of 38%. The performance in comparison to the original Elastic baseline (scoring an NDCG@10 of 43%) may look unimpressive, as in terms of relevance ranking we do not outperform the baseline. However, recall that our automatic run had a different aim, radically filtering the abstracts to a reading level agreeable with the intended layperson user. Hence we also include the Flesch-Kincaid Grade Level (FKGL) readability scores, and observe that the automatic run is able to return abstracts that are on average 2 years or school levels lower. That is, the reading level of the automatic run is around 12, corresponding to exit level compulsory education, or high school diploma, which would be accessible to the target audience of educated citizens. The baseline approach, in contrast, suggests a reading level requiring college or university education.

Assuming that users can select from a ranked list, it is of interest to analyze if, and how many, relevant and highly relevant abstracts are in the runs. Table 5 evaluates the runs with

**Table 6**  
Results for the SimpleText Task 2: Selecting rare terms.

Run	Total	Evaluated		Score_3		Score_5	
			+Limits		+Limits		+Limits
Selecting rare terms	263,022	1,315	1,175	105	69	60	49

Boolean quantization on various levels of relevance. In the top part of the table, we evaluate on all levels of relevance (with “1” meaning marginally relevant). Here we see very reasonable early precision scores, in particular for the manual run. The high average precision reflects suggests even good performance at higher recall levels. In the middle part of the table, we evaluate on relevance level 2 and higher (with “2” meaning relevant), confirming the superiority of the manual run to return those abstracts with higher levels of relevance. In the bottom part of the table, we evaluate on relevance level 4 and higher (with “4” meaning relevant to the article context), and observe that our automatic run not only returns abstracts that are easier to read, it also outperforms the baseline system in returning those abstracts of direct interest to the article context.

There are three conclusions based on our analysis. First, for every query and every level of relevance, there exist retrieved documents at a variety of readability levels. Second, a straightforward filter on readability level is able to retrieve abstracts that have a readability suitable for an educated citizen. Third, filtering on readability level leads to a small loss of retrieval effectiveness (as some relevant abstracts have high levels of text complexity), but still obtains a very reasonable performance in particular for retrieving highly relevant articles.

### 3.3. Task 2: Complexity Spotting

We continue with Task 2, asking to locate the most difficult concepts in a sentence extracted from a potentially relevant abstract, retrieved in response to a general query prompted by a popular science article. We submitted a single run, using an IDF based approach to find the least common term in the sentence, while boosting the most likely part of speech (i.e., nouns or noun phrases).

Table 6 shows the results of our official submission to Task 2, retrieving 263k terms for the 117k sentences in the test corpus. We see that a large fraction of highlighted terms (89%) has correct term limits, which is reassuring as we focused on selecting unigram tokens or words rather than complex concepts or phrases. Term difficulty is judged on a scale ranging from 0 (term needs no explanation) until 7 (impossible to understand). A fair fraction of terms selected (8%) has a high level of difficulty (3 or higher) and the majority of those (5%) a very high level of difficulty (5 or higher), with lower fractions exactly hitting the term limits (5% and 4% respectively).

At the time of writing, no ground truth is released for Task 2. Fortunately, the organizers released train data in an earlier stage of the track, consisting of 282 sentences with 453 manually extracted complex concepts. By treating every sentence as a “query” and every extracted concept as a “document identifier” (the string of characters after tokenizing and removing white-space),



**Table 7**  
Evaluation of SimpleText Task 2 (ranked list measures)

Run	MRR	Precision			NDCG			MAP
		5	10	20	5	10	20	
Longest terms	0.3577	0.1128	0.0720	0.0415	0.2994	0.3355	0.3542	0.2863
Rarest terms	0.3727	0.1340	0.0840	0.0440	0.3462	0.3847	0.3907	0.3088

**Table 8**  
Evaluation of SimpleText Task 2 (set measures).

Run	Precision			Recall			F1		
	1	2	3	1	2	3	1	2	3
Longest terms	0.2376	0.1950	0.1513	0.1690	0.2652	0.3052	0.1903	0.2152	0.1933
Rarest terms	0.2305	0.2092	0.1690	0.1727	0.2910	0.3424	0.1883	0.2307	0.2152

we can calculate standard ranked list measures as shown in Table 7. We observe very reasonable precision scores, for both selecting the longest and for selecting the least common terms. With the IDF approach being more effective than the simple word length. Qualitative inspection reveals that this is particularly caused by abbreviations who tend to be short character strings.

The gold standard also selects multi-term concepts, which are systematically missed by our single term or word-based approaches. Still, NDCG (35 to 40%) and MAP (around 30%) remain very reasonable. The NDCG scores reflect the graded concept difficulty level of the ground truth, normalized against the ideal ranking of the most difficult term first, and show impressive performance for these straightforward approaches, and a clearer advantage for the IDF approach over the length based approach. Qualitative inspection reveals that the concepts annotated as the most difficult, have often not the highest lexical complexity, but factor in the importance or centrality of the concept for understanding the sentence and abstract at hand.

As there are very few terms selected per sentence, leading to a necessarily low precision at 10 or even at 5, we also calculate set based precision, recall, and F1 for the first three results in Table 8. We see here that the set based F1 on the train data is highest after selecting 2 terms per sentence.

There are three conclusions based on our analysis. First, although spotting complex terms is a hard problem in general, straightforward approaches obtain very reasonable performance in the context of a single sentence. Second, using text statistics such as inverse word frequencies performs better than using only local features such as word length, and are particularly helpful to locate the most difficult terms. Third, lexical complexity as used in readability measures is not enough to locate the ground truth concepts, and we have to combine complexity with importance of the concept in the broader context of the information request.

**Table 9**

Results for SimpleText Task 3: zero-shot KiS (Keep it Simple) Model [6].

Run	Task	Sentences	FKGL			Compression	
			Mean	Median	Ratio	Mean	Median
No change	Train	648	15.46	15.4	0.00	1.00	1.00
KiS Model	Train	648	12.78	12.7	0.81	1.15	0.99
No change	Test	116,763	14.85	14.7	0.00	1.00	1.00
KiS Model	Test	116,763	12.06	11.9	0.79	1.33	1.01

### 3.4. Task 3: Text Simplification

We continue with Task 3, asking to perform text simplification proper, by rewriting a sentence extracted from a potentially relevant abstract, retrieved in response to a general query prompted by a popular science article. We only perform post-submission experiments, based on the zero-shot application of an existing neural text simplification model from [6], called the Keep it Simple (KiS) model. The model is based on GPT-medium, using a straightforward unsupervised training task with an explicit loss in terms of fluency, saliency, and simplicity. We are interested in this model as it is fully trained in an unsupervised way, and could be retrained or fine-tuned for the corpus or other academic texts without the need for huge human training data.

Table 9 shows the results of applying the KiS model zero-shot on the train and test data in terms of the generated output. To give an indication of whether the output is indeed simplified, we analyze the Flesch-Kincaid Grade Level (FKGL) of input and output sentences and the resulting compression in token length. We make the following three observations. First, we see a mean and median level of 15 in the scientific abstracts, which we lower by about 3 levels or years of education, with FKGL 12 corresponding to a high school diploma level. Second, we also look at the percentage of sentences where the FKGL is lowered, and see that this is the case in around 80% of sentences. Note that here, the dummy “no change” approach fails miserably, as not a single sentence is simplified. Third, in terms of sentence length, we see no significant compression, as the generated sentences are on par with the input sentences. This may be related to the corpus, as the input sentences from scientific abstracts tend to be not very long with a mean length of 25.8 (train) and 24.2 (test) tokens, and a median length of 24 and 23 respectively. This very significant reduction in text complexity is an encouraging result showing the promise to realize the general aims of the track.

Table 10 shows examples of the generated output for the first three sentences in the train corpus. On the one hand, we see no major issues in language aspects – every generated sentence is grammatical and coherent – and no issues with generating uncontrolled untrue information not contained in the input sentence. On the other hand, we see only very conservative changes, mostly very light editing in terms of deletions and substitutions. While the examples and earlier analysis is showing that we are moving in the right direction, the ground truth is a far more significant simplification. Hence, developing dedicated text simplification approaches for scientific text remains an important open problem.

For Task 3, we only performed post-submission experiments and our runs were not judged in terms of Lexical complexity, Syntax complexity, and Information loss. Hence we only show

**Table 10**

Examples simplifications: zero-shot KiS (Keep it Simple) Model [6] (sentences G11.1\_2892036907\_{1,2,3})

Input	Output	Reference
In the modern era of automation and robotics, <u>autonomous</u> vehicles are <u>currently</u> the focus of academic and industrial research	In the modern era of automation and robotics, vehicles are increasingly the focus of academic and industrial research.	Current academic and industrial research is interested in autonomous vehicles.
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 <u>too</u>	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.	Drones are increasingly used in the civilian and commercial domain and need to be autonomous.
Due to guidelines set by the governments regarding the operation ceiling of civil drones, road-tracking based navigation is <u>gaining interest</u> .	Due to guidelines set by the governments regarding the operation ceiling of civil drones, road-tracking based navigation is gaining traction.	Governments set guidelines on the operation ceiling of civil drones. So, road-tracking based navigation is attracting interest.

**Table 11**

Evaluation results for the SimpleText Task 3: zero-shot KiS (Keep it Simple) Model [6].

Model	Task	Evaluated	SARI	Bleu	Precision			
					1-gram	2-gram	3-gram	4-gram
No change	Train	648	0.5571	0.4204	0.6010	0.4531	0.3712	0.3089
KiS Model	Train	648	0.3984	0.2809	0.4881	0.3176	0.2319	0.1733

the automatic evaluation based on the human reference simplifications. Table 11 shows the automatic evaluation scores for Task 3, using standard SARI and Bleu scores. At the time of writing no test ground truth is available, so we only report scores on the train data. Note we apply a zero-shot model that is neither trained nor fine-tuned in any way on the CLEF SimpleText data, the evaluation on the train corpus is still an independent evaluation of the model’s quality. On the train corpus, with a single human simplified reference sentence, the KiS model obtains a Bleu score of 0.2809 and a SARI score of 0.3984. To put this number into perspective, the original paper reports scores in the range of 0.26 to 0.43 on a Wikipedia corpus [7]. Hence, a 40% SARI score is promising in terms of effectiveness.

We also include the dummy “no change” approach making no changes whatsoever, defaulting to returning the input sentence as is. Unlike in machine translation where this would result in very low, if any, token overlap, this proves a competitive approach in terms of the SARI and Bleu scores, as naturally the reference simplification will retain many tokens and n-grams of the original sentence. Recall from above that this no-change approach simplifies not a single sentence, resulting in a 0% of sentences scoring lower on the FKGL scores. This clearly indicates that we need to evaluate multiple aspects to capture the essence of text simplification.

There are three conclusions based on our analysis. First, an off-the-shelf text simplification

model based on GPT-2 is able to rewrite the sentences from academic abstracts with competitive SARI and Bleu scores against high quality human text simplifications. Second, although the model's revisions are conservative, there are no errors introduced and the output is fluent and without loss of information. Third, in terms of readability level, the simplification reduces the level from 15 (college level, undergraduate studies) to 12 (high school diploma, exit level compulsory education), suggesting a readability level suitable for a large fraction of educated citizens.

## 4. Discussion and Conclusions

This paper detailed the University of Amsterdam's participation in the CLEF 2022 SimpleText track.

We performed an extensive analysis of the corpus in terms of the text complexity, leading to the following findings. First, a negative result is that we can confirm that scientific abstracts on average are complex, validating and confirming the presumption of laypersons avoiding scientific information. Second, a positive result is that we also found that the variation of complexity is dramatic, and a large fraction of abstracts is within the readability level of an average educated adult citizen. Third, standard search engines are optimized for topic relevance and are completely ignoring secondary aspects such as the reading level of the text.

Next, we made two submissions to the first task retrieving academic abstracts in response to a query based on a popular science article, and found the following. First, for every query and every level of relevance, there exist retrieved documents at a variety of readability levels. Second, a straightforward filter on readability level is able to retrieve abstracts that have a readability suitable for educated citizens. Third, filtering on readability level leads to a small loss of retrieval effectiveness (as some relevant abstracts have high levels of text complexity), but still obtains a very reasonable performance in particular for retrieving highly relevant articles.

We also participated in the second task, asking to spot difficult terms in sentences from academic abstracts, and made the following observations. First, although spotting complex terms is a hard problem in general, straightforward approaches obtain very reasonable performance in the context of a single sentence. Second, using text statistics such as inverse word frequencies performs better than using only local features such as word length, and are particularly helpful to locate the most difficult terms. Third, lexical complexity as used in readability measures is not enough to locate the ground truth concepts, and we have to combine complexity with importance of the concept in the broader context of the information request.

Finally, we performed exploratory experiments with a large neural text simplification model based on GPT-2, and arrived at the following conclusions. First, an off-the-shelf text simplification model based on GPT-2 is able to rewrite the sentences from academic abstracts with competitive SARI and Bleu scores against high quality human text simplifications. Second, although the model's revisions are conservative, there are no errors introduced and the output is fluent and without loss of information. Third, in terms of readability level, the simplification reduces the level from 15 (college level, undergraduate studies) to 12 (high school, exit level compulsory education), suggesting a readability level suitable for a large fraction of educated citizens.

## Acknowledgments

This research was conducted as part of the final research projects of the Bachelor in Artificial Intelligence at the University of Amsterdam. We thank the coordinator Dr. Sander van Splunter for his support and flexibility to work around the CLEF deadlines. This research is funded in part by the Netherlands Organization for Scientific Research (NWO CI # CISC.CC.016), and the Innovation Exchange Amsterdam (POC grant). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

## References

- [1] L. Grossman, Time's person of the year: You, *TIME* 168 (2006).
- [2] 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.
- [3] E. SanJuan, S. Huet, J. Kamps, L. Ermakova, Overview of the CLEF 2022 SimpleText Task 1: Passage selection for a simplified summary, in: [8], 2022.
- [4] L. Ermakova, I. Ovchinnikova, J. Kamps, D. Nurbakova, S. Araújo, R. Hannachi, Overview of the CLEF 2022 SimpleText Task 2: Complexity spotting in scientific abstracts, in: [8], 2022.
- [5] L. Ermakova, I. Ovchinnikova, J. Kamps, D. Nurbakova, S. Araújo, R. Hannachi, Overview of the CLEF 2022 SimpleText Task 3: Query biased simplification of scientific texts, in: [8], 2022.
- [6] P. Laban, T. Schnabel, P. N. Bennett, M. A. Hearst, Keep it simple: Unsupervised simplification of multi-paragraph text, in: *ACL/IJCNLP'21: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Association for Computational Linguistics, 2021, pp. 6365–6378. URL: <https://doi.org/10.18653/v1/2021.acl-long.498>.
- [7] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, *Trans. Assoc. Comput. Linguistics* 4 (2016) 401–415. URL: [https://doi.org/10.1162/tacl\\_a\\_00107](https://doi.org/10.1162/tacl_a_00107). doi:10.1162/tacl\_a\_00107.
- [8] G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), *Proceedings of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, CEUR-WS.org, 2022.